

Algebra, Abstract

Ki Hang Kim
Fred W. Roush
Alabama State University

I. Sets and Relations
II. Semigroups
III. Groups
IV. Vector Spaces
V. Rings
VI. Fields
VII. Other Algebraic Structures

GLOSSARY

Binary relation Set of ordered pairs (a, b), where a, b belong to given sets A, B; as the set of ordered pairs of real numbers (x, y) such that $x > y$.

Equivalence relation Binary relation on a set S often written $x \sim y$ if (x, y) is in the relation such that (1) $x \sim x$; (2) if $x \sim y$ then $y \sim x$; and (3) if $x \sim y$ and $y \sim z$ then $x \sim z$ for all x, y, z in S.

Field Structure having operations of commutative, associative distributive addition and multiplication with additive and multiplicative identities and inverses of nonzero elements, as the real or rational numbers.

Function Binary relation from a set A to a set B that to every a in A associates a unique b in B such that (a, b) is in the (binary) relation, as a polynomial associates its value to a given x.

Group Structure having one associative operation with identity and inverses.

Homomorphism Function f from one structure S to another structure T such that, for every operation $*$ and all x, y in S, $f(x * y) = f(x) * f(y)$.

Ideal Subset $\Im$ of a ring containing the sum and difference of any two elements of $\Im$, and the product of any element of $\Im$ with any element of the ring.

Identity Element e such that, for all x for a given operation $*$, $e * x = x * e = x$.

Incline Structure having two associative and commutative operations $+$ and $\times$, satisfying the distributive law and (1) $x + x = x$ and (2) $x + (x \times y) = x$.

Isomorphism Homomorphism f from a set S to a set T such that the inverse mapping g from T to S where $g(y) = x$ if and only if $f(x) = y$ is also a homomorphism.

Partial order Binary relation R on a set S such that for all x, y, z in S (1) (x, x) is in R; (2) if (x, y) and (y, x) are both in R, then $x = y$; and (3) if (x, y) and (y, z) are in R, so is (x, z).

Quadratic form A function defined on a ring $\Re$ by $\Sigma a_{ij} x_i x_j$.

Quotient group For a normal subgroup N of a group G, the group formed by equivalence classes $\bar{x}$ under the relation that $x \sim y$ if xy^{-1} in N, with multiplication given by $\bar{x}\,\bar{y} = \overline{xy}$.

Ring Structure having one associative and commutative operation with identity and inverses, and a second operation associative and distributive with the first.

Semigroup Structure having an associative operation.

Simple Structure S such that, for every homomorphism f to another structure, either $f(x) = f(y)$ for all x, y in S or $f(x) \neq f(y)$ for all $x \neq y$ in S.

Subgroup Subset S of a group G containing the product and inverse of any of its elements. It is called normal if, for every g in G, s in S, the element gsg^{-1} is in S.

ABSTRACT ALGEBRA is the study of laws of operations and relations such as those valid for the real numbers and similar laws for new systems. It studies the class of all possible systems having given laws, for example, associativity. There is no standard object that it studies such as the real numbers in analysis. Two systems are isomorphic in abstract algebra if there is a one-to-one (1–1 for short) correspondence between their elements such that relations and operations agree in the two systems.

I. SETS AND RELATIONS

A. Sets

A set is any precisely defined collection of objects. The objects in a set are called elements (or members) of the set. The set $A = \{1, 2, x, y\}$ has elements $1, 2, x, y$ and we write $1 \in A$, $2 \in A$, $x \in A$, and $y \in A$. There is no set of all sets, but given any set we can obtain other sets by operations listed below, and for any set A and property $P(x)$ we have a set $\{x \in A : P(x)\}$ of all elements of A having the property. Things such as "all sets" are called classes rather than sets. There is a set of all real numbers and one of all isosceles triangles in three-dimensional space.

We say that A is a subset of B if all elements of A are in B. This is written $A \subset B$, or $B \supset A$. The set is called a proper subset if $A \neq B$; otherwise it is called an improper subset. For finite sets, a subset is proper if and only if it has strictly fewer elements. The empty set $\emptyset$ is the set with no elements.

If $A \subset B$ and $B \subset A$, then $A = B$. The union $\cup_{\mathfrak{F}} A$ of a family $\mathfrak{F}$ of sets is the set whose elements are all things that belong to at least one set $A \in \mathfrak{F}$. The intersection $\cap_{\mathfrak{F}} A$ is the set of all elements lying in every set A in $\mathfrak{F}$.

The power set $\mathfrak{P}(A)$ is the set of all subsets of A. The relative complement $B \backslash A$ (or $B - A$) is the set of all elements in B but not in A. For fixed set $B \supset A$, this is called the complement of A in B.

The Cartesian product $A_1 \times A_2 \times \cdots \times A_n$ of n sets $A_1, A_2, \ldots, A_n$ is the set of all n-tuples $(a_1, a_2, \ldots, a_n)$ such that $a_i \in A_i$ for all $i = 1, 2, \ldots, n$. Infinite Cartesian products can be similarly defined with i ranging over any index set $\mathbf{I}$.

B. Binary Relations

The truth or falsity of any mathematical statement about a relationship $x\,R\,y$, for instance, $x^2 + y^2 = 1$, can be determined from the set of ordered pairs (x, y) such that $x\,R\,y$. In fact, R is equivalent to the relationship that $(x, y) \in \{(x, y) \in A \times B : x\,R\,y\}$, where A and B are sets containing x and y.

The set of ordered pairs then represents the relationship and is called a binary relation. In general, a binary relation from A to B is a subset of $A \times B$, that is, a set of ordered pairs.

The union, intersection, and complement of binary relations from A to B are defined on them as subsets of $A \times B$. Another operation is composition. If R is a binary relation from A to B and S is a binary relation from B to C, then $R \circ S$ is $\{(x, z) \in A \times C : (x, y) \in R$ and $(y, z) \in S$ for some $y \in B\}$. Composition is distributive over union. For R, R_1, S, S_1, T, from A to B, A to B, B to C, B to C, C to D, we have $(R \cup R_1) \circ S = (R \circ S) \cup (R_1 \circ S)$ and $R \circ (S \cup S_1) = (R \circ S) \cup (R \circ S_1)$. It is associative; we have $(x, w) \in (R \circ S) \circ T$ if and only if for some $y \in B$ and $z \in C$, $(x, y) \in R$, $(y, z) \in S$, and $(z, w) \in T$. The same condition is obtained for $R \circ (S \circ T)$.

The identity relation 1_A is $\{(a, a) : a \in A\}$. This relation acts as an identity on either side for $R \subset A \times B$, $S \subset B \times C$; that is, $R \circ 1_B = R$, and $1_B \circ S = S$.

The transpose R^{T} of a binary relation R from A to B is $\{(b, a) \in B \times A : (a, b) \in R\}$. It is also called converse and inverse.

C. Functions

A partial function from a set A to a set B is a relation f from A to B such that if $(x, y) \in f$, $(x, z) \in f$ then $y = z$. If $(x, y) \in f$, we write $y = f(x)$ since this condition means y is unique. A partial function is a function defined on a subset of a set considered. Its domain is $\{a \in A : (a, b) \in f$ for some $b \in B\}$.

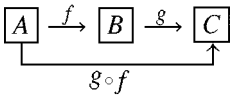
A function is a partial function such that for all $x \in A$ there exists $y \in B$ with $(x, y) \in f$.

A function is 1–1 if and only if whenever $(x, z) \in f$, $(y, z) \in f$ we have $x = y$. This is the transpose of the definition of a partial function. Functions and partial functions are thought of in several ways. A function may be

considered the output of a process in which x is an input, as $x^2 + x + 1$ adds a number to its square and 1. We may consider it as assigning an element of B to any element of A. Or we may consider it to be a map of A into the set B in which the point $x \in A$ is represented by $f(x) \in B$. Then $f(x)$ is called the image of x. Also for a subset $C \subset A$ the image $f(C)$ is $\{f(x) : x \in C\}$ the set of images of points of C. We may also consider it to be a transformation. Strictly, however, the term *transformation* is sometimes reserved for functions from a set to itself.

A 1–1 function sends distinct values of A to distinct points. For a 1–1 function on a finite set C, $f(C)$ will have exactly as many elements as C.

The composition of functions $f : A \to B$ and $g : B \to C$ is given by $g(f(x))$. It is written $g \circ f$. Composition is associative since it is a special case of composition of binary relations except that the order is reversed. The identity relation on a set is also called the identity function. A composition of partial (1–1) functions is respectively a partial (1–1) function.



A function $f : A \to B$ is *onto* if $f(A) = B$, that is, if every $y \in B$ is the image of some x. A composition of onto functions is onto. A 1–1 onto function is called a 1–1 correspondence. For a 1–1 onto function f, the inverse (converse) of f is defined by $f^{-1} = \{(y, x) : (x, y) \in f\}$ or $x = f^{-1}(y)$ if and only if $y = f(x)$. It is characterized by $f \circ f^{-1} = 1_B$, $f^{-1} \circ f = 1_A$. Moreover, for any binary relations R, S if $R \circ S = 1_B$, $S \circ R = 1_A$ both R and S must be 1–1 correspondences and S must be R^{T}. A 1–1 correspondence from a finite set to itself is called a permutation.

D. Order Relations

That a binary relation R from a set A to itself is (OR-1) reflexive means $(x, x) \in R$ for all $x \in A$; (OR-2) symmetric means if $(x, y) \in R$, then $(y, x) \in R$; and (OR-3) transitive means if $(x, y) \in R$ and $(y, z) \in R$, then $(x, z) \in R$. Frequently $x\ R\ y$ is written for $(x, y) \in R$, so that transitivity means if $x\ R\ y$ and $y\ R\ z$, then $x\ R\ z$. For instance, if $x \geq y$ and $y \geq z$, then $x \geq z$. So $x \geq y$ is transitive, but $x \neq y$ is not transitive.

The following relations are reflexive, symmetric, and transitive on the set of geometric figures: $x = y$ (same point set), $x \simeq y$ (congruence), $x \sim y$ (similarity), x has the same area as y. A reflexive, symmetric, transitive relation is called an equivalence relation. For any function f from A to B the relation $\{(x, y) \in A \times A : f(x) = f(y)\}$ is an equivalence relation.

The set $\bar{x} = \{y \in A : x\ R\ y\}$ is called the equivalence class of x (the set of all elements equivalent to the same element x). The set of equivalence classes $\bar{x}$ is called the quotient set A/R. The function $f(x) = \bar{x}$ is called the quotient map $A \to A/R$. The relation $\{(x, y) \in A \times A : f(x) = f(y)\}$ is precisely R.

Any two equivalence classes are disjoint or equal: If $x\ R\ z$ and $y\ R\ z$, by symmetry $z\ R\ y$ and by transitivity $x\ R\ y$. Any element belongs in the equivalence class of itself. A family $\mathfrak{F}$ of nonempty subsets of a set S is called a partition if (1) whenever $C \neq D$ in $\mathfrak{F}$, we have $C \cap D = \emptyset$; (2) $\cup_{\mathfrak{F}}\ C = S$. Therefore, the set of equivalence classes always forms a partition. Conversely, every partition $\mathfrak{F}$ arises from the equivalence relation $\{(x, y) \in A \times A$: for some $C \in \mathfrak{F}$, $x \in C$ and $y \in C\}$.

An important equivalence is congruence modulo m of integers. We say $x \equiv y (\mathrm{mod}\ m)$ for integers x, y, m if there exists an integer h such that $x - y = hm$, that is, if m divides $x - y$. If $x - y = hm$ and $y - z = km$, then $x - z = x - y + y - z = hm + km$. So if $x \equiv y (\mathrm{mod}\, m)$ and $y \equiv z (\mathrm{mod}\, m)$, then $x \equiv z (\mathrm{mod}\ m)$. This proves transitivity.

A relation that is reflexive and transitive is called a quasiorder. If it satisfies also (OR-4) antisymmetry if $(x, y) \in R$ and $(y, x) \in R$ then $x = y$, it is called a partial order. If a partial order satisfies (OR-5) completeness for all $x, y \in A$ either $(x, y) \in R$ or $(y, x) \in R$, it is called a total order, or a linear order.

For every total order on a finite set A, the elements of A can be labeled $a_1, a_2, \ldots, a_n$ in such a way that $a_i\ R\ a_j$ if $i < j$. An isomorphism between a binary relation R_1 on a set A_1 and R_2 on A_2 is a 1–1 correspondence $f : A_1 \to A_2$ such that $(x, y) \in R_1$ if and only if $(f(x), f(y)) \in R_2$. Therefore, every total order on a finite set is isomorphic to the standard order on $\{1, 2, \ldots, n\}$.

There are many nonisomorphic infinite total orders on the same set and many nonisomorphic partial orders on finite sets of n elements. The structure of quasiorders can be reduced to that of partial orders. For every quasiorder R on any set A, the relation $\{(x, y) : (x, y) \in R$ and $(y, x) \in R\}$ is an equivalence relation. The quasiorder gives a partial order on the set of equivalence classes of R and $(x, y) \in R$ if and only if $(\bar{x}, \bar{y}) \in R_1$.

The structure of partial orders on small sets can be described by diagrams known as Hasse diagrams. An element x of a partial order is called minimal (maximal) if for no $y \neq x$ does $(y, x) \in R$ ($(x, y) \in R$) where $(x, y) \in R$ is taken as $x \leq y$ in the order. Every partial order on a finite set has at least one minimal and at least one maximal element. Represent all minimal elements of the partial order as points at the same horizontal level of the bottom of the diagram. From then on, the ith level consists of elements z not in previous levels but such that for at least one y on

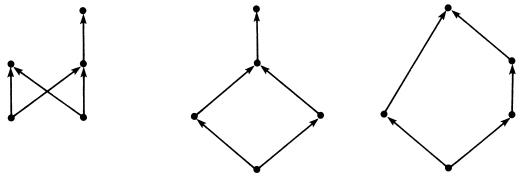


FIGURE 1 Hasse diagrams of partially ordered sets.

a previous level $y < z$ and for no x is $y < x < z$. That is, z is the very next element greater than y. For all such z, y draw a line segment from z to y. Figure 1 gives the Hasse diagrams of three partially ordered sets (posets for short).

A poset is called a lattice if every pair of elements in it have a least upper bound and a greatest lower bound. An upper (lower) bound on a set S is an element x such that $y \geq x$ ($y \leq x$) for all $y \in S$. Every linearly ordered set is a lattice, and any family of subsets of a set containing the union and intersection of any two of its members. These lattices are distributive: If $\wedge$, $\vee$ denote greatest lower bound, least upper bound, respectively, then $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ and $x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$. A lattice is called modular if these laws hold whenever two of x, y, z are comparable [two elements a, b are comparable for R if $(a, b) \in R$ or $(b, a) \in R$]. There exist nonmodular lattices, as the last diagram of Fig. 1.

A binary relation R is called a strict partial order if it is transitive and (OR-6) irreflexive, for no x does $(x, x) \in R$. There is a 1–1 correspondence between partial orders R (as $x \leq y$) on A and strict partial orders ($x < y$) obtained as $R \backslash \{(a, a) : a \in A\}$. An irreflexive binary relation R on A is called a semiorder if for all $x, y, z, w \in A$, (1) if $(x, y) \in R$ and $(z, w) \in R$, then either $(x, w) \in R$ or $(z, y) \in R$; (2) if $(x, y) \in R$ and $(y, z) \in R$, then either $(w, z) \in R$ or $(x, w) \in R$. Semiorders can be represented always as $\{(x, y) : f(x) > f(y) + \delta\}$ for some real valued function f and number δ, and are strict partial orders. Figure 2 shows the relationship of many of the relations taken up here.

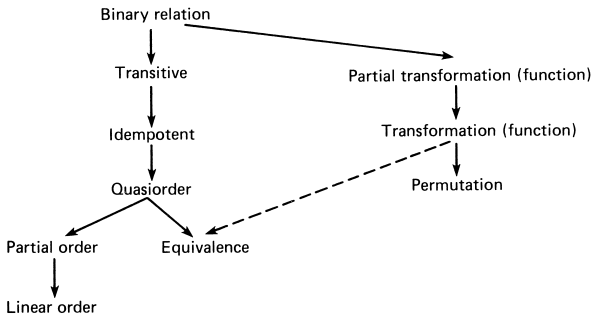


FIGURE 2 Classification of binary operations.

E. Boolean Matrices and Graphs

The Boolean algebra $\mathfrak{B}$ has elements $\{0, 1\}$ and operations $+, \cdot, ^c$ given by $0+0=0, 0+1=1+0=1+1=1, 0 \cdot 0 = 1 \cdot 0 = 0 \cdot 1 = 0, 1 \cdot 1 = 1, 0^c = 1, 1^c = 0$. There are many interpretations and uses of the Boolean algebra $\mathfrak{B}$: (1) propositional logic where 0 is false, 1 is true, $+$ is "or," $\cdot$ is "and," ^{c} is "not," (2) switching circuits where 0 means no current flows, 1 means current flows; and (3) $\mathfrak{B} \times \mathfrak{B} \times \cdots \times \mathfrak{B}$ can be taken as the set of subsets of an n-element set $\{y_i\}$, where $(x_1, x_2, \ldots, x_n)$ corresponds to the subset $\{y_i : x_i = 1\}$; (4) 0 means zero in $\mathbf{R}$, 1 means some positive number in $\mathbf{R}$, where $\mathbf{R}$ denotes the set of all real numbers. The algebra $\mathfrak{B}$ satisfies the same laws as the algebra of sets under $\cup, \cap, \sim$ where $\sim$ denotes the complementation since it is a Boolean algebra.

An $n \times m$ Boolean matrix $A = (a_{ij})$ is an $n \times m$ rectangle of elements of $\mathfrak{B}$. The entry in row (horizontal) i and column (vertical) j is denoted a_{ij}. To every binary relation R from a set $X = \{x_1, x_2, \ldots, x_n\}$ to set $Y = \{y_1, y_2, \ldots, y_m\}$ we can assign a Boolean matrix $M = (m_{ij})$, where m_{ij} is 0 or 1 according to whether (x_i, y_j) does or does not lie in R. This gives a 1–1 onto correspondence from binary relations from X to Y to $n \times m$ Boolean matrices. Many questions can be dealt with simply in the Boolean matrix form. Boolean matrices are multiplied and added by the same formulas as for ordinary matrices, except that operations are Boolean. Composition (union) of binary relations corresponds to product (sum) of Boolean matrices.

A directed graph (digraph for short) consists of a set V of elements called vertices (represented by points) and a set E or ordered pairs of V known as edges. Each ordered pair (x, y) is represented by a segment with an arrow going from point x to point y. Thus digraphs are essentially the same as binary relations from a set to itself. Any $n \times n$ (n-square) Boolean matrix or binary relation on a labeled set can be represented as a digraph, and conversely. Figure 3 shows the Boolean matrix and graph of a binary relation on $\{1, 2, 3, 4\}$.

F. Blockmodels

Often binary relations are empirically obtained. Such binary relations can frequently be simplified by blocking the Boolean matrices: dividing the set of indices into disjoint subsets, relabeling to get members of the same subset adjacent, and dividing the matrix into blocks. Each nonzero block is replaced by a single 1 entry and each zero block by a single 0 entry. Many techniques called clustering exist for dividing the total set into subsets. One (CONCOR) is to take iterated correlation matrices of the rows.

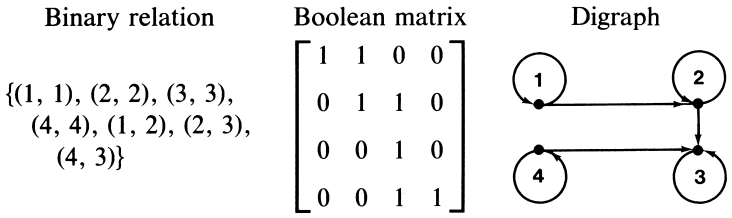


FIGURE 3 Matrix and graph of a relation.

Provided that every nonzero block has at least one 1 in each row the replacement of blocks by single entries will preserve all Boolean sums and products.

G. General Relational Structures

A general finite relational structure on a set S is an indexed family R_α of subsets of $S \cup (S \times S) \cup (S \times S \times S) \cup \cdots$. Such structures include order structures and operational structures such as multiplicative ones as subsets $\{(x, y, z) \in S \times S \times S : x * y = z\}$. A homomorphism of relational structures (index the same way) R_α on S_1 to T_α on S_2 consists of a function $f : S_1 \to S_2$ such that if g is the mapping $S_1 \cup (S_1 \times S_1) \cup (S_1 \times S_1 \times S_1) \cup \cdots$ to $S_2 \cup (S_2 \times S_2) \cup (S_2 \times S_2 \times S_2) \cup \cdots$ which is f on each coordinate then $g(R_\alpha) \subset T_\alpha$ for each α. An isomorphism of relational structures is a 1–1 onto homomorphism such that $g(R_\alpha) = T_\alpha$ for each α. The quotient structure of R_α on S associated with an equivalence relation E on S is the structure $T_\alpha = g(R_\alpha)$, for f the mapping $S \to S/E$ assigning to each element its equivalence class.

H. Arithmetic of Residue Classes

Let $\mathbf{Z}$ denote the set of all integers. Let E_m be the equivalence relation $\{(x, y) : \mathbf{Z} \times \mathbf{Z} : x - y = km$ for some $k \in \mathbf{Z}\}$. This is denoted $x \equiv y (\mathrm{mod}\, m)$. We have previously noted that it is an equivalence relation.

It divides the integers into exactly m equivalence classes, for $m \neq 0$. For $m = 3$, the classes are $\bar{0} = \{\ldots, -9, -6, -3, 0, 3, 6, 9, \ldots\}$, $\bar{1} = \{\ldots, -8, -5, -2, 1, 4, 7, 10, \ldots\}$, $\bar{2} = \{\ldots, -7, -4, -1, 2, 5, 8, 11, \ldots\}$. Any two members of the same class are equivalent (3 divides their difference).

This relation has the property that if $x \equiv y (\mathrm{mod}\, m)$ then, for any $z \in \mathbf{Z}$, $x + z \equiv y + z$ since m divides $x + z - (y + z) = x - y$ and $xz \equiv yz$. Such a relation in general is called a congruence. For any congruence, we can define operations on the classes by $\overline{x + y} = \bar{x} + \bar{y}$, and $\overline{xy} = \bar{x}\bar{y}$.

Let $\mathbf{Z_m}$ be the set of equivalence classes of $\mathbf{Z}$ under congruence module $\mathbf{m}$ and under $+, \times$ quotient operators.

Operations in $\mathbf{Z_5}$ are given in Fig. 4.

II. SEMIGROUPS

A. Generators and Relations

A binary operation is a function that given two entries from a set S produces some element of a set T. Therefore, it is a function from the set $S \times S$ of ordered pairs (a, b) to T. The value is frequently denoted multiplicatively as $a * b$, $a \circ b$, or ab. Addition, subtraction, multiplication, and division are binary operations.

The set S is said to be closed under the operation if the product always lies in S itself. The positive integers are not closed under subtraction or division.

The operation is called associative if we always have $(a \circ b) \circ c = a \circ (b \circ c)$. We have noted that this always holds for composition of functions or binary relations. Conversely, if closure and associatively hold, the set can always be represented by a set of functions under composition.

A set with a binary operation satisfying associativity and closure is called a semigroup. The positive integers form a semigroup under multiplication or addition. An element e is called a left (right) identity in a semigroup S if for all $x \in S$, $ex = x (xe = x)$. A semigroup with two-sided identity is called a monoid. To represent a semigroup as a set of functions under composition, first add a two-sided identity element e to obtain a monoid M. Then for each x in M define a function $f_x : M \to M$ by $f_x(y) = x \circ y$.

The set generated by a subset $G \subset S$ is the set of all finite products $\{x_1 x_2 \cdots x_n : n \in \mathbf{Z^+}, x_i \in G\}$, where $\mathbf{Z^+}$ denotes the set of all positive integers. The set satisfies

$+$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$
$\bar{0}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$
$\bar{1}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{0}$
$\bar{2}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{0}$	$\bar{1}$
$\bar{3}$	$\bar{3}$	$\bar{4}$	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{4}$	$\bar{4}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$

$\cdot$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$
$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$
$\bar{1}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$
$\bar{2}$	$\bar{0}$	$\bar{2}$	$\bar{4}$	$\bar{1}$	$\bar{3}$
$\bar{3}$	$\bar{0}$	$\bar{3}$	$\bar{1}$	$\bar{4}$	$\bar{2}$
$\bar{4}$	$\bar{0}$	$\bar{4}$	$\bar{3}$	$\bar{2}$	$\bar{1}$

FIGURE 4 Addition and multiplication of module 4 residue classes.

closure since $(x_1x_2\cdots x_n)(y_1y_2\cdots y_n)$ has agian the form required. A subset of a semigroup satisfying closure is itself a semigroup and is called a subsemigroup. The set G is called a set of generators for S if G generates the entire semigroup.

For a given set G of generators of S, a relation is an equation $x_1x_2\cdots x_n = y_1y_2\cdots y_n$, where $x_i, y_i \in G$.

A homomorphism of semigroups is a mapping $f: S_1 \to S_2$ such that $f(x \circ y) = f(x)f(y)$ for all $x, y \in S_1$. For example, e^z is a homomorphism from the additive semigroup of real numbers $(\mathbf{R}, +)$ to the multiplicative semigroup of complex numbers $(\mathbf{C}, \times)$, since $e^{x+y} = (e^x)(e^y)$. Here $\mathbf{C}$ denotes the set of all complex numbers. The trace and determinant of n-square matrices are homomorphisms, respectively, from the additive and multiplicative semigroups of n-square matrices to the additive and multiplicative semigroups of real numbers.

An isomorphism of semigroups is a 1–1 onto homomorphism. The inverse function will then also be an isomorphism. Isomorphic semigroups are structurally identical.

Two semigroups having the same generating set G and the same relations are isomorphic, since the mapping $f(x_1 \circ x_2 \circ \cdots \circ x_n) = x_1 * x_2 * \cdots * x_n$ will be a homomorphism where $\circ, *$ denote the two operations. The identity of the sets of relations guarantees that this is well defined, that is, if $x_1 \circ x_2 \circ \cdots \circ x_m = y_1 \circ y_2 \circ \cdots \circ y_n$, then $f(x_1 \circ x_2 \circ \cdots \circ x_m) = f(y_1 \circ y_2 \circ \cdots \circ y_n)$.

For any set of generators G and set R of relations, a semigroup S can be produced having these generators and satisfying the relations such that if $f(G) \to T$ is any function such that for every relation $x_1 \circ x_2 \circ \cdots \circ x_m = y_1 \circ y_2 \circ \cdots \circ y_n$ the relation $f(x_1) * f(x_2) * \cdots * f(x_m) = f(y_1) * f(y_2) * \cdots * f(y_n)$ holds then f extends to a homomorphism $S \to T$.

To produce S we take the set of all "words" $x_1x_2\cdots x_m$ in G, that is, sequences of elements in G. Two "words" are multiplied by writing one after the other $x_1x_2\cdots x_my_1y_2\cdots y_n$. We define an equivalence relation on words by $w_1 \sim w_2$ if w_2 can be obtained from w_1 by a series of replacments of a $x_1x_2\cdots x_nb$ by $ay_1y_2\cdots y_mb$, where a, b are words (or are empty) and $x_1x_2\cdots x_n = y_1y_2\cdots y_m$ is a relation in R. Then S is the set of equivalence classes of words.

The semigroup S is called the semigroup with generators S and defining relations R. The fact that multiplication is well defined in S follows from the fact that $a \sim b$ is a congruence. This means that if $a \sim b$ then for all words $x, ax \sim bx$ and $xa \sim xb$, and that $\sim$ is an equivalence relation.

For all semigroup homomorphisms $f: S \to T$ the relation $f(x) = f(y)$ is a congruence. Conversely any congruence gives rise to a homomorphism from S to the set of equivalent classes.

B. Green's Relations

For analyzing the structure of semigroups in many cases it is best to decompose it into a family of equivalence classes.

Let S be a semigroup and let M be S with an identity element if S lacks one. Let $\mathcal{R}$ be the relation for $x, y \in S$ that for some $a, b \in M$, $xa = y$, $yb = x$. Let $\mathcal{L}$ be the relation for some $a, b \in M$, $ax = y$, $by = x$. Let $\mathcal{J}$ be the relation that for some $a, b, c, d \in M, axc = y, byd = x$. What these equations express is a kind of divisibility, that either x, y can be a factor of the other. For example, in $\mathbf{Z}^+$ under multiplication $(2, 5) \notin \mathcal{J}$ because 2 is not a factor of 5.

There are two other important relations: $\mathcal{H} = \mathcal{L} \cap \mathcal{R} = \{(x, y) : x \mathcal{L} y \text{ and } x \mathcal{R} y\}$ and $\mathcal{D} = \mathcal{L} \circ \mathcal{R} = \mathcal{R} \circ \mathcal{L}$. These are both equivalence relations also and are known as Green's relations.

For the semigroup of transformations on a finite set T, $f \mathcal{L} g$ if and only if the partitions given by the equivalence relations $\{(x, y) \in T \times T : f(x) = f(y)\}$, $\{(x, y) \in T \times T : g(x) = g(y)\}$ coincide, $f \mathcal{D} g$ if and only if $f \mathcal{J} g$ if and only if f, g have the same number of image elements, and $f \mathcal{R} g$ if and only if their images are the same.

In a finite semigroup, $\mathcal{J} = \mathcal{D}$. The entire semigroup is always broken into $\mathcal{J}$-classes, which are partially ordered by divisibility. The $\mathcal{D}$-classes, which are in these, are broken into $\mathcal{H}$-classes $H_{ij} = R_i \cap L_j$, where R_i, L_i are the $\mathcal{R}, \mathcal{L}$-classes in a given $\mathcal{D}$-class.

The $\mathcal{H}$-classes can be laid out in a matrix (H_{ij}) called the eggbox picture of a semigroup. There exists a 1–1 correspondence between any two $\mathcal{H}$-classes.

A left ideal in a semigroup S is a set $\mathcal{I} \subset S$ such that for all $x \in S$, $y \in \mathcal{I}$ the element $xy \in \mathcal{I}$. Right and two-sided ideals are similarly defined by $yx \in \mathcal{I}$, $yxz \in \mathcal{I}$ for all $y, z \in S$, respectively. An element x generates the principal left, right two-sided ideals $Sx = \{yx : y \in S\}$, $xS = \{xy : y \in S\}$, $SxS = \{yxz : y, z \in S\}$. Two elements are $\mathcal{L}$-, $\mathcal{R}$-, $\mathcal{J}$-equivalent (assuming S has an identity, otherwise add one) if and only if they generate the same left, right, or two-sided ideals, respectively.

C. Binary Relations and Boolean Matrices

The set of binary relations on an n-element set under composition forms a semigroup that is isomorphic to B_n, the semigroup of n-square Boolean matrices under Boolean matrix multiplication. A $1 \times n (n \times 1)$ Boolean matrix is called a row (column) vector. Two vectors are added or multiplied by a constant (0 or 1) as matrices: $(1, 0, 1) + (0, 1, 0) = (1, 1, 1)$, $0v = 0$, and $1v = v$. A set of Boolean vectors is called a subspace if it is closed under sums and contains 0. The span of a set W of Boolean vectors is the set of all finite sums of elements of W, including 0.

The row (column) space of a Boolean matrix is the space spanned by its row (column) vectors. Two Boolean matrices are $\mathcal{J}$- (also $\mathcal{D}$-) equivalent if and only if their row spaces are isomorphic. The basis for a space of Boolean vectors spanned by S is $\{x \in S : x \neq 0$ and x is not in the span of $S\backslash\{x\}\}$. Two subspaces are identical if and only if their bases are the same. Bases for the row, column spaces of a Boolean matrix are called row, column bases. Two Boolean matrices are $\mathcal{L}$- ($\mathcal{R}$-) equivalent if and only if their row (column) bases (and so spaces) coincide.

A semigroup of the form $\{x^n\}$ is called cyclic. Every cyclic semigroup is determined by the index $k = \inf\{k \in \mathbf{Z}^+ : x^{k+m} = x^k$ for some $m \in \mathbf{Z}^+\}$, and the period $d = \inf\{d \in \mathbf{Z}^+ : x^{k+d} = x^k\}$. The set of powers $k, k+1, \ldots$ repeat with period d. If an n-square Boolean matrix A is reflexive ($A \geq I$), then $A = AI \leq A^2 \leq A^3 \leq \cdots \leq A^{n-1} = A^n$, an idempotent. Here I denotes the identity matrix. Also if A is fully indecomposable, meaning that there exists no nonempty $K, L \subset \{1, 2, \ldots, n\}$ with $|K| + |L| = n$, where $|S|$ denotes the cardinality of a set S, and $a_{ij} = 0$ for $i \in K$, $j \in L$, then $A^{n-1} = J$ and so A has period 1 and index at most $n-1$. Here J is the Boolean matrix all entries of which are 1.

In general, an n-square Boolean matrix has period equal to the period of some permutation on $\{1, 2, \ldots, n\}$ and index at most $(n-1)^2 + 1$.

D. Regularity and Inverses

An element x of a semigroup S is said to be a group inverse of y if S has a two-sided identity ϵ and $xy = yx = \epsilon$. A semigroup in which every element has a group inverse is called a group. In most semigroups, few elements are invertible in this sense. A function is invertible if and only if it is 1–1 onto. A Boolean matrix in invertible if and only if it is the matrix of a permutation (permutation matrix).

There are many weaker ideas of inverse. The two most important are regularity $yxy = y$ and Thierrin–Vagner inverse $yxy = y$ and $xyx = x$. An element y has a Thierrin–Vagner inverse if and only if it is regular: If $yxy = y$ then xyx is a Thierrin–Vagner inverse. A semigroup in which all elements are regular is called a regular semigroup. The semigroups of partial transformation, transformations, and n-square matrices over a field are regular. In the semigroup of n-square Boolean matrices, an element is regular if and only if its row space forms a distributive lattice as a poset.

An idempotent is an element x such that $xx = x$. An element x is regular if and only if its $\mathcal{L}$-equivalence class contains an idempotent: if $xyx = x$ then xy is idempotent. The same holds for $\mathcal{R}$-equivalence classes. Therefore, if two elements are $\mathcal{L}$ or $\mathcal{R}$-equivalent (therfore $\mathcal{D}$-equivalent), one is regular if and only if the other is.

If an $\mathcal{H}$-class contains an idempotent, then it will be closed under multiplication, and multiplication by any of its elements gives a 1–1 onto mapping from it to itself. Therefore, it forms a group. Conversely any group in a semigroup lies in a single $\mathcal{H}$-class since under multiplication any two elements are $\mathcal{L}$-equivalent and $\mathcal{R}$-equivalent.

E. Finite State Machines

The theory of automata deals with different classes of theoretical machines representing robots, calculators, and similar devices. The simplest are the finite state machines. There are two essentially equivalent varieties: Mealy machines and Moore machines.

A Mealy machine is a 5-tuple (S, X, Z, ν, μ), where S, X, Z are sets, ν a function $S \times X$ to S, and μ a function $S \times X$ to Z. The same definition holds for Moore machines except that μ is a function S to Z.

Here S is the set of internal states of the machine. For a computer this could include all posibilities as to which circuit elements are on or off. The set X is the set of inputs, which could include a program and data. The set Z is the set of outputs, that is, the desired response. The function μ gives the particular output from a given internal state and input. The function ν gives the next internal state resulting from a given state and input. For a computer this is determined by the circuitry. For example, a flip-flop will change its internal state if it receives an input of 1; otherwise the internal state will remain unchanged.

A Mealy machine to add two n-digit binary numbers, a, b can be constructed as follows. Let the ith digits of a_i, b_i be inputs, so $X = \{0, 1\} \times \{0, 1\}$. Let the carry from the $i-1$ digit be c_i. It is an internal state, so $S = \{0, 1\}$. The output is the ith digit of the answer, so $Z = \{0, 1\}$. The function ν gives the next carry. It is 1 if and only if $a_i + b_i + c_i > 1$. The function μ gives the output. It is 1 if and only if $a_i + b_i + c_i$ is odd. Figure 5 gives the values of ν and μ.

With a finite state machine is associated a semigroup of transformations, the transformations $f_x(s) = \nu(s, x)$ of

X		S	ν	μ
0	0	0	0	0
0	1	0	0	1
1	0	0	0	1
1	1	0	1	0
0	0	1	0	1
0	1	1	1	0
1	0	1	1	0
1	1	1	1	1

FIGURE 5 Machine to add two binary numbers.

the state space, and all compositions of them $f_{x_1x_2\cdots x_n}(s) = f_{x_1}(f_{x_2}\cdots f_{x_n}(s))$. This is called the semigroup of the machine.

Two machines are said to have the same behavior if there exists a binary relation R from the initial states of one machine to the intial state of the other such that if $(s, t) \in S$ then for any sequence $x_1, x_2, \ldots, x_k$ of inputs machine 1 in state s gives the same sequence of outputs as machine 2. A machine M_1 equivalent to a given machine M having a minimal number of states can be constructed as follows. Call two states of M equivalent if, for any sequence of inputs, the same sequence of outputs is obtained. This gives an equivalence relation on states. Then let the states of M_1 be the equivalence classes of states of M.

No finite state machine M can multiply, as the above machine adds binary number of arbitrary length. Suppose such a machine has n states. Then suppose we multiply the number 2^k by itself in binary notation (adding zeros in front until the correct number of digits in the answer is achieved). Let f_x be the transformation of the state space given by inputs of 0, 0 and let a be the state after the inputs 1, 1. The inputs 0, 0 applied to state $f_x^i(a)$ give output 0 for $i = 0, 1, 2, \ldots, k-2$ but output 1 for $i = k-1$. Yet the transformation f on a set of n elements will have index at most n. Therefore, if $k > n$ then $f_x^{k-1}(a)$ will coincide with $f_x^j(a)$ for some $j < k-1$. This is a contradiction since one yields 0 output, the other 1 output. It follows that no such machine can exist.

F. Mathematical Linguistics

We start with a set W of basic units considered words. Mathematical linguistics is concerned with the formal theory of sentences, that is, sequences of words that are grammatically allowed and the grammatical structure of sentences (or longer units). This is syntax. Meaning (semantics) is usually not dealt with.

For a set X, let X^* be the set of finite sequences from X including the empty sequence e. For instance, if X is $\{0, 1\}$, then X^* is $\{e, 0, 1, 00, 01, 10, 11, 000\cdots\}$. For a more important example, we can consider the family of all sequences of logical variables p, q, r, and $\vee$ (or), $\wedge$ (and), (,) (parentheses), $\rightarrow$ (if then), $\sim$ (not). The set of logical formulas will be a subset of this.

A phrase structure grammar is a quadruple (N, W, ρ, ψ), where W (set of words) is nonempty and finite, N (nonterminals) is a finite set disjoint from W, $\psi \in N$, and ρ is a finite subset of $((N \cup W)^* \setminus W^*) \times (N \cup W)^*$. The set N is a set of valid grammatical forms involving abstract concepts such as ψ (sentence) or subject, predicate, object. The set ρ (productions) is a set of ways we can substitute into a valid grammatical form to obtain another, more specific one. The element ψ is called the starting symbol. If $(x, y) \in \rho$, we are allowed to change any occurrence of x with y. Members of W are called terminals.

Consider logical formulas involving operations $\vee, \sim$ and variables p, q, r. Let $W = \{p, q, r, \vee, (,), \sim\}$, $N = \{\psi\}$. We derive formulas by successive substitution, as ψ, $(\psi \vee \psi)$, $((\psi \vee \psi) \vee \psi)$, $((\psi \vee \psi) \vee \sim \psi)$, $((p \vee \psi) \vee \sim \psi)$, $((p \vee q) \vee \sim \psi)$, $((p \vee q) \vee \sim r)$.

An element $y \in (N \cup W)^*$ is said to be directly derived from $x \in (N \cup W)^*$ if $x = azb$, $y = awb$ for some $(z, w) \in \rho$, $a, b \in (N \cup W)^*$. An indirect derivation is a sequence of direct derivations. Here $\rho = \{(\psi, p), (\psi, q), (\psi, r), (\psi, \sim\psi), (\psi, (\psi \vee \psi))\}$.

The language determined by a phrase structure grammar is the set of all $a \in W^*$ that can be derived from ψ.

A grammar is called context free if and only if for all $(a, b) \in \rho$, $a \in N$, $b \neq e_0$. This means that what items can be substituted for a given grammatical element do not depend on other grammatical elements. The grammar above is context free.

A grammar is called regular if for all $(a, b) \in \rho$ we have $a \in N$, $b = tn$, where $t \in W$, $n \in N$, or $n = e_0$. This means at each derivation we go from $t_1t_2\cdots t_rn$ to $t_1t_2\cdots t_rt_{r+1}m$, where t_i are terminals, n, m nonterminals, $(n, t_{r+1}m) \in \rho$. So we fill in one terminal at each step, going from left to right. The grammar mentioned above is not regular.

To recognize a grammar is to be able to tell whether or not a sequence from W^* is in the language. A grammar is regular if and only if some finite state machine recognizes it. The elements of W are input 1 at a time and outputs are "yes, no," meaning all symbols up to the present are or are not in the language. Let the internal states of the machine be in 1–1 correspondence with all subsets of N, and let the initial state be ψ. For a set S_1 of nonterminals and input x let the next state be the set S_2 of all nonterminals z such that for some $u \in S_1$, (u, xz) is a production. Then at any time the state consists of all nonterminals that could occur after the given seqence of inputs. Let the output be "yes" if and only if for some $u \in N$, $(u, x) \in \rho$. This is if and only if the previous inputs together with the current input form a word in the language.

For the converse, if a finite state machine can recognize a language, let W be as in the language, N be the set of internal states, ψ the initial state, the productions the set of pairs (n_1, xn_2) such that if the machine is in state n_1 and x is input, state n_2 is the next state, and the set of pairs (n_1, x) such that in state n_1 after input x the machine answers "yes."

A further characterization of regular language is the Myhill–Nerode theorem. Let W^* be considered a semigroup of words. Then a language $L \subset W^*$ is regular if and only if the congruence $\{(x, y) \in W^* \times W^* : axb \in L$ if and only if $ayb \in L$ for all $a, b \in L\}$ has finitely many equivalence classes. This is if and only if there exists a

finite semigroup H, a homomorphism $h: W^* \to H$ is onto, and a subset $S \subset H$ such that $L = h^{-1}(S)$.

III. GROUPS

A. Examples of Groups

A group is a set G together with a binary operation here denoted $\circ$ such that (1) $\circ$ is a function $G \times G \to G$ (closure); (2) $\circ$ is associative, for example, $(x \circ y) \circ z = x \circ (y \circ z)$; (3) there exists a two-sided identity e such that $e \circ x = x \circ e = x$ for all $x \in G$; (4) for all $x \in G$ there exists $y \in G$ with $x \circ y = y \circ x = e$. Here y is known as an inverse of x. From now on we suppress $\circ$.

For any semigroup S with a two-sided identity e, let $S^* = \{x \in S : xy = yx = e \text{ for some } y \in S\}$, the set of invertible elements. The element $y = x^{-1}$ is unique since if $y_1 x = e$, $xy_2 = e$ then $y_1 = y_1 x y_2 = y_2$. Then S^* satisfies closure since $(ab)^{-1} = b^{-1}a^{-1}$ and is a group where a^{-1} is an inverse of a.

A group with a finite number of elements is called a finite group; otherwise, it is called an infinite group. If a group G is finite, and G contains n elements, we say that the order of G is n and we write $|G| = n$. If G is infinite, we write $|G| = \infty$. In general, $|G|$ denotes the cardinality of a set G.

The group of all permutations of $\{1, 2, \ldots, n\}$ is called $\mathcal{S}_n$, the symmetric group of degree n. The degree is the number of elements in the domain (and range) of a permutation. Therefore, $\mathcal{S}_n$ has degree n, order $n!$. A subgroup of $\mathcal{S}_n$ is formed by the set of all transformation of the form $1 \to k+1, 2 \to k+2, \ldots, n-k \to n, n-k+1 \to 1, \ldots, n \to k$.

For any set P of subsets of n-dimensional Euclidean space $\mathbf{E^n}$, let T be the union of the subsets of P. A symmetry of P is a mapping $f: T \to T$ such that (S–1) for $x, y \in T$, $d(x, y) = d(f(x), f(y))$, where $d(x, y)$ denotes the distance from x to y; and (S–2) for $A \subset T$, $A \in P$ if and only if $f(A) \in P$. That is, f preserves distances and sends the subsets C of P, for example, points and lines, so other subsets of P. The inverse off is its inverse function, which also satisfies (S–1) and (S–2).

The sets $\mathbf{R}$, $\mathbf{Z}$, $\mathbf{Z_m}$ under addition are all groups. The group $\mathbf{Z_m}$, the group of permutations $\{1 \to k+1, 2 \to k+2, \ldots, n \to k\}$, and the group of rotational symmetries of an n-sided regular polygon are isomorphic. That is, there exists a 1–1 correspondence f between any two such that $f(xy) = f(x)f(y)$ for all x, y in the domain. A group isomorphic to any of these is called a cyclic group of order m.

The group of all symmetries of a regular n-sided polygon has order $2n$ and is called the dihedral group of order n. It is isomorphic to the group of functions $\mathbf{Z_n} \to \mathbf{Z_n}$ of the form $f(x) = \pm x + k$. Its multiplication table is given in Fig. 6, where x_i is $x + i - 1$ if $i > n+1$, $-x + i - n - 1$ if $i > n$.

	x_1x_2	$\cdots$	x_n	$x_{n+1}x_{n+2}$	$\cdots$	x_{2n}
x_1	x_1x_2	$\cdots$	x_n	$x_{n+1}x_{n+2}$	$\cdots$	x_{2n}
x_2	x_2x_3	$\cdots$	x_1	$x_{n+2}x_{n+3}$	$\cdots$	x_{n+1}
$\vdots$		$\cdots$				
x_n	x_nx_1	$\cdots$	x_{n-1}	$x_{2n}x_{n+1}$	$\cdots$	x_{2n-1}
x_{n+1}	$x_{n+1}x_{2n}$	$\cdots$	x_{n+2}	x_1x_n	$\cdots$	x_2
$\vdots$		$\cdots$			$\cdots$	
x_{2n}	$x_{2n}x_{2n-1}$	$\cdots$	x_{n+1}	x_nx_{n-1}	$\cdots$	x_1

FIGURE 6 Multiplication table of the dihedral group.

B. Fundamental Homomorphism Theorems

A group homomorphism is a function $f: G \to H$ satisfying $f(xy) = f(x)f(y)$ for all x, y in G. If f is 1–1 and onto it is called an isomorphism. For any group G of order n there is a 1–1 homomorphism $G \to \mathcal{S}_n$; label the elements of G as $x_1, x_2, \ldots, x_n$. For any x in G, the mapping $a \to xa$ gives a 1–1 onto function $G \to G$, so there exists a permutation π with $xx_i = x_{\pi(i)}$. This is a homomorphism since if x, y are sent to π, ϕ, we have $yxx_i = yx_{\pi(i)} = x_{\phi(\pi(i))}$. There is an isomorphism from $\mathcal{S}_n$ to the set of n-square permutation matrices (either over $\mathbf{R}$ or a Boolean algebra), and it follows that any finite group can be represented as a group of permutations or of invetible matrices.

Groups can be simplified by homomorphisms in many cases. The determinant represents the group of nonsingular n-square matrices in the multiplicative group of nonzero real numbers (it is not 1–1). For a bounded function g, $\int gf\,dx$ is a homomorphism from the additive group of integrable functions f to the additive $\mathbf{R}$.

The kernel of a homomorphism $f: G \to H$ is $\{x : f(x) = e\}$. Every homomorphism sends the identity to the identity and inverses to inverses. Existence of inverses means we can cancel on either side in groups: If $ax = ay$ then $a^{-1}ax = a^{-1}ay$, $ex = ey$, $x = y$. Therefore, the identity is the unique elements satisfying $ee = e$. Under a homomorphism $f(e) = f(ee) = f(e)f(e)$ so $f(e)$ is an identity. So e is in the kernel. A homomorphism is 1–1 if and only if e is its only element: If $f(x) = f(y)$, then $f(xy^{-1}) = f(x)f(y^{-1}) = f(x)f(y)^{-1} = f(x)f(x)^{-1} = e$.

If x, y are in the kernel, so is xy since $f(xy) = f(x)f(y) = ee = e$. If x is in the kernel, so is x^{-1}. Any

subset of group that is closed under products and inverses is a subgroup. Both the kernel and image of a homomorphism are subgroups.

The mapping $c_g : G \to G$ defined by $c_g(x) = gxg^{-1}$ is called a conjugation. An isomorphism from a group (or other structure) to itself is called an automorphism. Since $c_g(xy) = gxyg^{-1} = gxg^{-1}gyg^{-1}$, c_g is a homomorphism. If $h = g^{-1}$, then c_g is the inverse of c_h. Therefore, the mappings c_g are automorphisms of a group. They are called the inner automorphisms. This gives a homomorphism of any group to its automorphisms, which also form a group, the automorphism group. A subgroup N is normal if $c_g(x) \in N$ for all $x \in N$ and $g \in G$. The kernel is always a normal subgroup.

If $f : G \to H$ has kernel K and is onto, the group G is said to be an extension of K by H. Whenever there exists a homomorphism f from K to Aut(H) for any groups K, H, a particular extension exists called the semidirect product. Here Aut(H) denotes the automorphism group of H. This has as its set $K \times H$ and products are defined by $(k_1, h_1)(k_2, h_2) = [k_1k_2, f(k_2)(h_1)(h_2)]$, $k_1, k_2 \in K$ and $h_1, h_2 \in H$. The dihedral group is a semidirect product of $\mathbf{Z_2}$ and $\mathbf{Z_m}$.

The groups K, H will generally be simpler than G. A theory exists classifying extensions. If a group has no normal subgroups except itself and $\{e\}$, it is called simple. The alternating group, the group of permutations in $\mathcal{S}_n$ whose matrices have positive determinant, is simple for $n > 4$. Its order is $n!/2$. For n odd, the group of n-square real-valued invertible matrices of determinant 1 is simple. For n even a homomorphic image of $\mathcal{S}_n$ with kernel $\{I, -I\}$ is simple, where I is an identity matrix. All finite and finite-dimensional differentiable, connected simple groups are known and most are constructed from groups of matrices.

For any subgroup H of a group G there exist equivalence relations defined by $x \sim y$ if $xy^{-1} \in H (y^{-1}x \in H)$. These equivalence classes are called left (right) cosets. The left (right) coset of a is $Ha = \{ha : h \in H\}$ ($aH = \{ah : h \in H\}$). There exists a 1–1 correspondence $H \to Ha$ given by $x \to xa$ ($x \to ax$ for right cosets). Therefore, all cosets have the same cardinality as H. If $[G : H]$ denotes the number of right (or left) cosets, then $|G| = |H|[G : H]$, which is known as Lagrange's theorem. So the order of a subgroup of a finite group must divide the order of the group. If G is a group of prime order p, let $x \in G$, $x \neq e$. Then $\{x^n\}$ forms a subgroup whose order divides and must therefore equal p. So $G = \{x^n\}$.

If N is a normal subgroup and $x \sim y$, for all $g \in G$ then $gx \sim gy$ and $xg \sim yg$ since $gxy^{-1}g^{-1}$ and xy^{-1} are in N if xy^{-1} in N. It follows that if $a \sim b$ and $c \sim d$, then $ac \sim bd$. Therefore, the equivalence classes form a group G/N called the quotient group. Its order is $|G|/|N|$.

For any homomorphism $f : G \to H$ with kernel K, image M every equivalence class of K is sent to a single element. That is, f gives a mapping $G/K \to M$. This mapping is an isomorphism.

If A, B are normal in a group G, and B is a subgroup of A, then

$$\frac{G/B}{A/B} = G/A$$

If A is a normal subgroup and B is any subgroup of G, and $AB = \{ab : a \in A, b \in B\}$ then,

$$\frac{AB}{A} = \frac{B}{A \cap B}$$

C. Cyclic Groups

Let G be a group. If there exists $a \in G$ such that $G = \{a^k : k \in \mathbf{Z}\}$, then G is called a cyclic group generated by a, and a is called a generator of G. For addition, the definition becomes:
There exists $a \in G$ such that for each $g \in G$ there exists $k \in \mathbf{Z}$ such that $g = ka$.

A group G is said to be commutative when $xy = yx$ for all x, y in G. Commutative groups are also called Abelian. In a commutative group, $gxg^{-1} = x$ for all $g \in G$. Therefore, every subgroup of a commutative group is normal. Let $\mathbf{R^n}$ denote the set of all n-tuples of real numbers. Then ($\mathbf{R^n}$, +) is a commutative group. For any groups $G_1, G_2, \ldots, G_n$, the set $G_1 \times G_2 \times \cdots \times G_n$ under componentwise multiplication $(x_1, \ldots, x_n) \times (y_1, \ldots, y_n) = (x_1y_1, \ldots, x_ny_n)$ is a group called the Cartesian or direct product of $G_1, G_2, \ldots, G_n$. If all G_i are commutative, so is the direct product. For any indexed family G_α of groups, coordinatewise multiplication makes the Cartesian product a group $\times\alpha G_\alpha$. The subset of elements (x_α) such that $\{\alpha : x_\alpha = e\}$ is finite is also a group sometimes called the direct sum.

A set $G_0 \subset G$ is said to generate the set $\{x_1x_2 \cdots x_n : x_i \in G_0$ or $x_i^{-1} \in G_0$ for all i and $n \in \mathbf{Z}\}$. Every finitely generated Abelian group is isomorphic to a direct product of cyclic (1 generator) groups.

A group of words on any set X is defined as the set of finite sequences (with e) $x_1^{a(1)}x_2^{a(2)} \cdots x_n^{a(n)}$, $x_i \in X$, $a(n) \in \mathbf{Z}$. We can reduce such words until $a(i) \neq 0$, $x_i \neq x_{i+1}$ by adding exponents if and only if they have the same reduced form. Equivalence classes of words form a group F.

Relations among generators in a group can be written in the form $x_1^{a(1)}x_2^{a(2)} \cdots x_n^{a(n)} = e$. For any set G_0 and set of relations R in the elements of G_0, there exists a group G defined by these relations such that any mapping $f : G_0 \to H$ extends to a unique homomorphism $g : G \to H$ if the relations hold in H. The group G is

defined to be F/N, where N is the subgroup generated by $\{yx_1^{a(1)}x_2^{a(2)}\cdots x_n^{a(n)}y^{-1} : x_1^{a(1)}x_2^{a(2)}\cdots x_n^{a(n)} = e$ is a relation of R, $y \in F\}$.

The dihedral group has generators, x, y and defining relations $x^m = e$, $y^2 = e$, $yx = x^{-1}y$.

The word problem for a group G with given sets of generators and defining relations is to find an algorithm which for any product $x_1^{a(1)}x_2^{a(2)}\ldots x_n^{a(n)}$ will determine (in a finite number of steps) whether this product is e. For general groups with finite G_0, R, the word problem is unsolvable; that is, no algorithm exists.

If a cyclic group $\{x^m\}$ has two powers equal, then some power is e. Let d be the least positive power equal to e. Then $x^0, x^1, x^2, \ldots, x^{d-1}$ are distinct and for any m if $m = qd + r$ then $x^m = (x^d)^q x^r = e^q x^r = x^r$. It follows that the group is isomorphic to $\mathbf{Z_m}$ under addition. If no two powers are equal, then $f(x^m) = m$ defines an isomorphism to $\mathbf{Z}$ under addition.

The multiplicative groups $\mathbf{Z_m^*} = \mathbf{Z_m}\backslash\{\bar{0}\}$ must be distinguished from $\mathbf{Z_m}$. For m prime we will later prove $\mathbf{Z_m^*}$ is cyclic (of order $m-1$). For $m = 8$, $S = \{\bar{1}, \bar{3}, \bar{5}, \bar{7}\} \subset \mathbf{Z_8^*}$ with multiplication $\bar{1} = e$, $\bar{3}^2 = \bar{5}^2 = \bar{7}^2 = \bar{1}$, $\bar{3}\bar{5} = \bar{7}$. It is isomorphic to $\mathbf{Z_2} \times \mathbf{Z_2}$, is not cyclic, and is the noncyclic group of smallest order.

D. Permutation Groups

In the symmetric group $\mathcal{S}_n$ of all permutations of $\{1, 2, \ldots, n\}$ elements are written in two standard ways. The notation

$$\begin{pmatrix} 1 & 2 & 3 & \ldots & n \\ a_1 & a_2 & a_3 & \ldots & a_n \end{pmatrix}$$

denotes the function $f(1) = a_1$, $f(2) = a_2$, $f(3) = a_3, \ldots, f(n) = a_n$. It is convenient to compose permutations written in this notation. To find

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 3 & 2 & 1 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 4 & 2 & 1 \end{pmatrix}$$

as functions, on the right go through 1 to 5 on top. For each number find its image below; then locate that image on the upper left. Below it is the final image.

$$\begin{matrix} 1{-}5{-}1 \\ 2{-}4{-}2 \\ 3{-}3{-}4 \\ 4{-}2{-}5 \\ 5{-}1{-}3 \end{matrix} \longrightarrow \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 4 & 5 & 3 \end{pmatrix}$$

The inverse of an element is obtained by looking up i on the bottom and taking the element above it, $i = 1$ to 5.

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 5 & 1 & 2 & 3 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 1 & 2 \end{pmatrix}$$

In computer work, permutations can be stored as arrays $F(N)$, $G(N)$ with composition $F(G(N))$ [or $u = G(x)$: $y = F(u)$].

The second standard notation is cyclic notation. A k-cycle $(x_1x_2\cdots x_k)$ is the permutation f such that $f(x_1) = x_2$, $f(x_2) = x_3, \ldots, f(x_k) = x_1$, and $f(y) = y$ for $y \neq x_i$. If for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, k$, $x_i \neq y_j$, then $(x_1x_2\cdots x_m)$ and $(y_1y_2\ldots y_k)$ commute.

Any permutation ρ can be written uniquely as a product of disjoint cycles of length greater than 1. In such a representation x, y will lie in the same cycle if and only if $\rho(x) \neq x$, $\rho(y) \neq y$ and $\rho^k(x) = y$ for some $k \in \mathbf{Z}$. This defines an equivalence relation, so take the sets of cycles to be the equivalence classes. In each equivalence class, choose an element x, and the cycle if it has size d must be $(x\rho(x)\rho^2(x)\cdots\rho^{d-1}(x))$.

Two permutations are conjugate if and only if they have the same number of cycles of each length. The order of a group element x is the order of the cyclic subgroup it generates, that is, the least m such that $x^m = e$. If $x = z_1z_2\cdots z_k$ in cycle form, then since $z_iz_j = z_jz_i$, $x^m = z_1^m z_2^m \cdots z_k^m$ and $x^m = e$ if and only if all $z_i^m = e$. Therefore, the order of x is the least common multiple of the lengths of its cycles.

A permutation ρ acts on the expression $\Pi_{i<j}(x_1 - x_j)$ by taking it to $\Pi_{i<j}(x_{\rho(i)} - x_{\rho(j)})$. Since each pair $\{i, j\}$ is sent to $\{\rho(i), \rho(j)\}$ in a 1–1 fashion, each factor goes to plus or minus another factor in the same expression. The sign of a permutation is defined to be $+1$ or -1 according as the whole expression goes to $+$ or $-$ itself. This is a group homomorphism. Permutations of sign $+1$ (-1) are called even (odd). Those of sign $+1$ form the kernel of the homomorphism, the alternating group $\mathcal{A}_n$. Any 2-cycle (transposition) is odd. A k-cycle $(x_1x_2\cdots x_k)$ is a product of $(k-1)$ transpositions $(x_kx_{k-1})(x_{k-1}x_{k-2})\cdots(x_2x_1)$ so it has sign $(-1)^{k-1}$. The symmetric group is generated by any transposition $t = (ab)$ and any n-cycle of the form $z = (ab\cdots)$. By conjugation $z^n t z^{-n}$ we obtain a family of transpositions t_i. Products of these give a new family of elements w_i. Conjugation of t_i by w_j give all transpositions. Products of these give all cycles. Products of cycles give all permutations.

E. Orbits

A group G of permutations of a set S is also said to be a group acting on a set. The function evaluation $g(s)$ defines a function $G \times S \to S$ such that $g(h(s)) = (g \circ h)(s)$. The relation on S; $\{(s, t) : g(s) = t$ for some $g \in G\}$ is an equivalence relation. If $g(s) = t$ and $h(t) = u$, then $h(g(s)) = u$,

so it is transitive. Its equivalence classes are called orbits. For cyclic groups, the orbits are the underlying sets of the cycles.

A permutation group is said to be transitive if S has a single orbit. This will be true if and only if G does not lie in a conjugate of some $S(n_1) \times S(n_2) \times \cdots \times S(n_k) \subset S(n)$, where $n_i > 0, k > 1, \Sigma n_i = n$, and $S(n_i)$ acts on $n_1 + n_2 + \cdots + n_{i-1} + 1, n_1 + n_2 + \cdots + n_{i-1} + 2, \ldots, n_1 + n_2 + \cdots + n_{i-1} + n_i$. Otherwise, it is called intransitive. For any element x, the isotropy subgroup of x is $H = \{g \in G : gx = x\}$. Then for any a, the coset aH maps x to a single element ax. Different cosets give different elements. Therefore, there is a 1–1 correspondence between cosets and elements of the orbit of x. So $[G : H]$ is the cardinality of the orbits of x.

Any group acts on itself by conjugation. The orbits are the equivalence classes under the relation for some g in G, $x = gyg^{-1}$, and are called conjugacy classes. Their size, since they are orbits, divides $|G|$.

Let G be a group of order p^n, p prime, $n \in \mathbf{Z}^+$. Suppose the identity e was the only conjugacy class containing only one element. Then all other conjugacy classes have order $p^i, i > 0$. So G has $kp + 1$ elements for some k. This is false. If every element of G has an order power of p, then G is called a p-group. Therefore, for any p-group G the set $C = \{c : cg = gc \text{ for all } g \in G\}$ has size larger than 1. For any group C so defined is a normal subgroup and is commutative, called the center.

F. Symmetry

An algebraic structure on a set S is a family of subsets $\mathfrak{F}_\alpha$ indexed on $\alpha \in \mathbf{I}$ of $S_1 = S \cup (S \times S) \cup (S \times S \times S) \cup \cdots$. An automorphism of this structure is a 1–1 onto mapping $f : S \to S$ such that for all $\alpha \in \mathbf{I}$, $T \in \mathfrak{F}_\alpha$ if and only if $f(T) \in \mathfrak{F}_\alpha$. For operational structures, this is any 1–1 mapping that is a homomorphism: $f(xy) = f(x)f(y)$. The complex number system has the automorphism $a + bi \to a - bi$. The additive real numbers has the automorphism $x \to -x$.

Especially in a geometric setting, automorphisms are called symmetries. A metric symmetry of a geometric figure is an isometry (distance-preserving map) on its points, which also gives a 1–1 correspondence on its lines and other distinguished subsets. Finite groups of isometries in three-dimensional space include the isometry groups of a regular n-gon (dihedral), cube and octahedron (extension of $\mathbf{Z_2}$ by $\mathcal{S}_4$), icosahedron and dodecahedron (extension of $\mathbf{Z_2}$ by $\mathcal{A}_5$).

The possible isomorphism types of n-dimensional symmetric groups, finite or continuous, are limited. One of the principal applications of group theory is to make use of exact or approximate symmetries of a system in studying it.

G. Enumeration

The equation $|O_X| = |G|/|H|$, where O_X is the orbit of X, G a group, and H the isotropy subgroup of X, simplifies a number of enumeration problems.

Another problem is to count the number of orbits. A lemma of Burnside states that for a group G acting on a set X the number of orbits is $|G|^{-1} \Sigma_{g \in G} f_g$, where f_g is the number of elements fixed by g, that is, $|\{x : gx = x\}|$. This is proved by counting the set $S = \{(x, g) : g(x) = x\}$ first over each g (yielding Σf_g) and then over each x (yielding $\Sigma |G|/|O_x|$).

Polya's theory of counting was motivated partly by a desire to count the number of distinct structures of a chemical compound with known numbers of elements and known graph but unknown locations of elements within the diagram. Suppose we have a hexagonal compound containing four carbon and two silicon atoms (Fig. 7; the reality of such a compound is not considered here). In how many distinct ways can the atoms be arranged? The symmetry group of a hexagon is the dihedral group of order 12.

In the Redfield–Polya theory, we first obtain a polynomial

$$P_G(x_1, x_2, \ldots, x_k) = \frac{1}{|G|} \sum_{g \in G} x_1^{c_1(g)} x_2^{c_2(g)} \cdots x_k^{c_k(g)}$$

where $c_i(g)$ is the number of cycles of g of length i and k is the maximum length of any cycle. The dihedral group breaks down as shown in Table I. Then

$$P_G(x_1, x_2, \ldots, x_n) = \tfrac{1}{12}\left(x_1^6 + 2x_6^1 + 2x_3^2 + x_2^3 + 3x_2^2 x_1^2 + 3x_2^3\right)$$

For instance, e having six 1-cycles yields term x_1^6. Redfield and Polya computed the number of labelings of X by labels $\theta_1, \theta_2, \ldots, \theta_m$ such that θ_i occurs exactly n_i times. They proved that this is the coefficient of $\theta_1^{n_1} \theta_2^{n_2} \cdots \theta_m^{n_m}$ in

$$P_G\left(\theta_1 + \theta_2 + \cdots + \theta_m, \theta_1^2 + \theta_2^2 + \cdots + \theta_m^2, \ldots, \theta_1^k + \theta_2^k + \cdots + \theta_m^k\right)$$

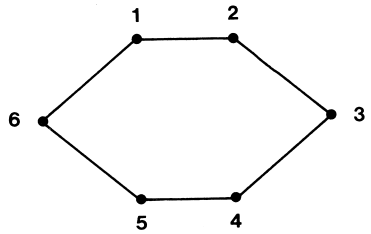


FIGURE 7 Hexagonal molecule.

TABLE I Cycle Structure of Hexagonal Symmetries

Rotations			Reflections		
e	6 1-cycles	$c_1 = 6$	—	—	—
(123456) (654321)	1 6-cycle	$c_6 = 1$	About 2 vertices (26) (35) (13) (46) (15) (24)	2 2-cycles	$c_1 = c_2 = 2$
(135) (246) (153) (246)	2 3-cycles	$c_3 = 2$	—	—	—
(14) (25) (36)	3 2-cycles	$c_2 = 3$	About an edge center (12) (36) (45) (23) (14) (65) (16) (25) (34)	3 2-cycles	$c_2 = 3$

In the problem above this yields the coefficient of $\theta_1^2\theta_2^4$ in

$$\tfrac{1}{12}\Big[(\theta_1+\theta_2)^6 + 2\big(\theta_1^6+\theta_2^6\big) + \big(\theta_1^3+\theta_2^3\big)^2 + 4\big(\theta_1^2+\theta_2^2\big)^3 + 3\big(\theta_1^2+\theta_2^2\big)^2(\theta_1+\theta_2)^2\Big]$$

That coefficient is

$$\frac{1}{12}\left(\binom{6}{2} + 4\binom{3}{1} + 3(3)\right) = \frac{1}{12}(15+12+9) = 3$$

IV. VECTOR SPACES

A. Vector Space Axioms

A vector space is a set having a commutative group addition, and a multiplication by another set of quantities (magnitudes) called a field. A field is a set $\mathbf{F}$ such as $\mathbf{R}$ or $\mathbf{C}$ having addition and multiplication $\mathbf{F}\times\mathbf{F}\to\mathbf{F}$ such that the axioms in Table II hold for all x, y, z and some 0, 1 in $\mathbf{F}$.

The standard example of a vector space is the set $\mathbf{v_n}$ of n-tuples of members of $\mathbf{F}$. Let $(x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n) \in \mathbf{v_n}$. The $(x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n) = (x_1+y_1, x_2+y_2, \ldots, x_n+y_n)$ and multiplication by c is given by $c(x_1, x_2, \ldots, x_n) = (cx_1, cx_2, \ldots, cx_n)$ where $c \in \mathbf{F}$. For $\mathbf{F} = \mathbf{R}$, $\mathbf{v_3}$ is the set of vectors consideered in physics. These can be regarded as positions (displacement), velocities, or forces: Each adds in the prescribed way.

TABLE II Field Axioms

Property	Operation: Addition	Operation: Multiplication
Commutative	$x+y=y+x$	$xy=yx$
Associative	$(x+y)+z=x+(y+z)$	$(xy)z=x(yz)$
Identity	For all $x\in\mathbf{F}$, $x+0=0$	For all $x\in\mathbf{F}$, $x1=x$
Inverse	For all $x\in\mathbf{F}$, there exists $-x$ such that $x+(-x)=0$	For all $x\in\mathbf{F}$ $(x\neq 0)$, there exists x^{-1} such that $xx^{-1}=1$
Distributive	$x(y+z)=xy+xz$	

A general vector space is a set $\mathbf{v}$ having an addition $\mathbf{v}\times\mathbf{v}\to\mathbf{v}$ and a scalar multiplication $\mathbf{F}\times\mathbf{v}\to\mathbf{v}$ such that $(\mathbf{v}, +)$ is a commutative group and for all $a, b\in\mathbf{F}$, v, $w\in\mathbf{v}$; $a(v+w)=av+aw$, $(a+b)v=av+bv$, $(ab)v=a(bv)$, and $cv=v$, where $c=1\in\mathbf{F}$. Incidentally, $\mathbf{F}$ is called the field of scalars, and multiplication by its elements scalar multiplication.

B. Basis and Dimension

The span $\langle S\rangle$ of a subset $S\subset\mathbf{v}$ is $\{a_1s_1+a_2s_2+\cdots+a_ns_n : n\in\mathbf{Z^+}, s_i\in S, a_i\in\mathbf{F}\}\cup\{0\}$. The elements $a_1s_1+a_2s_2+\cdots+a_ns_n$ are called linear combinations of $s_1, s_2, \ldots, s_n$. A sum of two linear combinations is a linear combination and a multiple by $c\in\mathbf{F}$ of a linear combination is a linear combination. A subset of a vector space closed under addition and scalar multiplication is called a subspace. Therefore, the span of any set is a subspace.

A subspace is a vector space in itself. Its span is itself. It follows that $\langle S\rangle$ lies in any subspace containing S.

An indexed family t_α, $\alpha\in\mathbf{I}$ of vectors is said to be linearly independent if t_γ is not a linear combination of t_α, $\alpha\neq\gamma$ and no $t_\gamma=0$ for all γ. Otherwise, it is called linearly dependent. A set of vectors W is likewise said to be linearly independent if for all $x\in W$, $x\notin\langle W\backslash\{x\}\rangle$.

A basis for a vector space is a linearly independent spanning set for the vector space. A chain in a poset is a subset that is linearly ordered by the partial order. Zorn's lemma states that in a poset X in which every chain has a maximal element there is a maximal element of X. Take X to be the family of linearly independent sets in $\mathbf{v}$. Every chain has its union as an upper bound. A maximal linearly independent set therefore exists. It is a basis. Therefore, every vector space has a basis.

A linearly ordered set is well ordered if every nonempty subset has a least element. Using Zorn's lemma it can be shown that every set can be well ordered. From every set $\{x_\alpha\}$ of vectors where the index set is well ordered we can pick a subset $\{x_\gamma\}$ that is linearly independent and spans the same set. Let $B = \{x_\gamma : x_\gamma \notin \langle\{x_\alpha : \alpha < \gamma\}\rangle\}$. Then no element of B is linearly dependent on previous vectors. But if some element of B were linearly dependent, let $y = \Sigma a_i x_i$, $a_i \neq 0$. Let x_n be the last of the x_i in the ordering. Then $x_n = (1/a_n)(y - \Sigma_{i \neq n} a_i x_i)$ is linearly dependent on previous vectors. This is a contradiction.

Two bases for a vector space have the same cardinal number. For finite bases $u_1, u_2, \ldots, u_n, v_1, v_2, \ldots, v_m$ let $m < n$. Expressions $u_i = \Sigma a_{ij} v_j$ exist since $\langle\{v_i\}\rangle = \langle\{u_i\}\rangle$. Since $m < n$, we can find a nonzero solution of the m linear equations $\Sigma_i w_i a_{ij} = 0$, $j = 1$ to m in n variables w_i. Then $\Sigma w_i u_i = 0$, not all $w_i = 0$. This contradicts linear independence of the u_i.

The cardinality of a basis for a vector space is called the dimension of the vector space.

For vector spaces $\mathbf{v}$, $\mathbf{w}$ a homomorphism (isomorphism) is a (1–1 and onto) function $f : \mathbf{v} \to \mathbf{w}$ such that $f(x + y) = f(x) + f(y)$ and $f(cx) = cf(x)$ for all $x, y \in \mathbf{v}$, $c \in \mathbf{F}$. If two vector spaces over $\mathbf{F}$ have the same dimension, choose bases u_α, v_α for them. Every element of $\mathbf{v}$ has a unique expression $\Sigma a_i u_i$, $u_i \in \{u_\alpha\}$, and every element of $\mathbf{w}$ has a unique expression $\Sigma a_i v_i$, $v_i \in \{v_\alpha\}$. Therefore, $f(\Sigma a_i u_i) = \Sigma a_i v_i$ defines an isomorphism $\mathbf{v}$ to $\mathbf{w}$. Conversely if two vector spaces are isomorphic, their dimensions must be equal since if $\{u_\alpha\}$ is a linearly independent (spanning) set for $\mathbf{v}$, $\{f(u_\alpha)\}$ will be a linearly independent (spanning) set for $\mathbf{w}$.

C. Linear Transformations

A linear transformation is another phrase for homomorphism of vector spaces. To every subspace $\mathbf{w}$ of a vector space $\mathbf{v}$ is associated the quotient space $\mathbf{v}/\mathbf{w}$. This is the set of equivalence classes of the equivalence relation $x - y \in \mathbf{w}$. Transitivity follows from $(x - y) + (y - z) = x - z \in \mathbf{w}$ if $x - y$, $y - z$ do. If $x - y \in \mathbf{w}$, $u - v \in \mathbf{w}$, then $(x + u) - (y + v) \in \mathbf{w}$. So addition of equivalence classes is well defined; so is scalar multiplication since $a(x - y) \in \mathbf{w}$ if $x - y$ does.

The quotient space, additively, is the same as the quotient group. Let $\{s_\sigma\}$ be a basis for $\mathbf{w}$. We can find a basis for $\mathbf{v}$ of the form $\{s_\sigma\} \cup \{t_\tau\}$, where $\{s_\sigma\} \cap \{t_\tau\} = \emptyset$. Let q_γ be a well-ordered basis for $\mathbf{v}$. Delete each q_α that is linearly dependent on prior q_γ together with the s_σ. The remaining set with s_σ is a spanning set and is linearly independent. This is a basis for $\mathbf{v}$, which contains a basis for $\mathbf{w}$.

Let $\{s_\sigma\} \cup \{t_\tau\}$ denote the resulting basis. Since every element of $\mathbf{w}$ has the form $\Sigma a_i s_i + \Sigma b_i t_i$ it will be equivalent to $\Sigma b_i t_i$. No two distinct sums $\Sigma b_i t_i$ can have difference a nonzero sum $\Sigma a_i s_i$ so the classes of t_i are linearly independent in $\mathbf{v}/\mathbf{w}$. Therefore $\bar{t}_\tau$ gives a basis for $\mathbf{v}/\mathbf{w}$.

The external direct sum of two vector spaces $\mathbf{v}$, $\mathbf{w}$ is $\mathbf{v} \times \mathbf{w}$ with operations $(v_1, w_1) + (v_2, w_2) = (v + v_2, w_1 + w_2)$, and $c(v, w) = (cv, cw)$. It has as basis the disjoint union of any bases for $\mathbf{v}$, $\mathbf{w}$. It is denoted $\mathbf{v} \oplus \mathbf{w}$. Therefore, if $\mathbf{w} \subset \mathbf{v}$, $\mathbf{v}$ is isomorphic to $\mathbf{w} \otimes \mathbf{v}/\mathbf{w}$. Moreover, $\dim(\mathbf{w}) + \dim(\mathbf{v}/\mathbf{w}) = \dim(\mathbf{v})$, where $\dim(\mathbf{v})$ denotes the dimension of $\mathbf{v}$.

A complement to a subspace $\mathbf{w} \subset \mathbf{v}$ is a subspace $\mathbf{u} \subset \mathbf{v}$ such that $\mathbf{u} \cap \mathbf{w} = \{0\}$, $\mathbf{u} + \mathbf{w} = \mathbf{v}$, where $\mathbf{u} + \mathbf{w} = \{u + w : u \in \mathbf{u}, w \in \mathbf{w}\}$. Every subspace has a complement by the construction above with $\mathbf{u} = \langle\{t_\alpha\}\rangle$. The mapping $\Sigma a_i t_i \to \Sigma a_i \bar{t}_i$ gives an isomorphism $\mathbf{u} \to \mathbf{v}/\mathbf{w}$.

For a linear transformation $f : \mathbf{v} \to \mathbf{w}$, the image set $\mathrm{Im}(f)$ and the null space (kernel) $\mathrm{Nu}(f) = \{v \in \mathbf{v} : f(v) = 0\}$ are both subspaces. Results on group homomorphism imply that f gives an isomorphism $\bar{f} : \mathbf{v}/\mathrm{Nu}(f) \to \mathrm{Im}(f)$.

A linear transformation is 1–1 (onto) if and only if $\mathrm{Nu}(f) = 0$ ($\mathrm{Im}(f) = \mathbf{w}$). The dimension of $\mathrm{Im}(f)$ is called the rank of f.

Choose bases x_i, y_i, z_i for vectors spaces $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$. Let f be a homomorphism from $\mathbf{u}$ to $\mathbf{v}$ and g a homomorphism $\mathbf{v}$ to $\mathbf{w}$. There exist unique coefficients a_{ij}, b_{ij} such that $f(x_i) = \Sigma a_{ij} y_i$ and $g(y_i) = \Sigma b_{ij} z_j$. Then $g(f(x_i)) = g(\Sigma a_{ij} y_j) = \Sigma a_{ij} g(y_j) = \Sigma a_{ij} \Sigma b_{jk} z_k$. Therefore, the composition $g \circ f$ sends x_i to $\Sigma_{j,k} a_{ij} b_{jk} z_k$. This rule defines a product on nm-tuples a_{ij}, which must be associative since composition of functions is. It is distributive since $f(g(x) + h(x)) = f(g(x)) + f(h(x))$ and $(g + h)(f(x)) = g(f(x)) + h(f(x))$.

D. Matrices and Determinants

An $n \times m$ matrix A is an nm-tuple (a_{ij}) indexed on $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$. Matrices are represented as rectangles of elements where a_{ij} is placed in horizontal line i (row i) and vertical line j (column j).

Addition is entrywise. The entries of both matrices in location (i, j) are added to form the (i, j)-entry of the sum. The (i, j)-entry of the product is formed by multiplying row i of the first factor by column j of the second factor entry by entry and adding all these products. For example,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

This is the same as the operation defined at the end of the last section. To multiply two n-square matrices by the

formula requires n^3 multiplications and $n^2(n-1)$ additions. There does exist a slightly improved computer method known as fast matrix multiplication.

Matrices obey the following laws: (1) $A+B=B+A$, (2) $(A+B)+C=A+(B+C)$, (3) $(AB)C=A(BC)$, (4) $A(B+C)=AB+AC$, and (5) $0+A=A$. Here 0 denotes the zero matrix all the entries of which are 0. It is an additive identity.

There is a 1–1 correspondence from the set of linear transformations from an n-dimensional vector space to an m-dimensional vector space to the set of $n \times m$ matrices, given bases in these. As in the last subsection (a_{ij}) corresponds to f such that $f(x_i)=\Sigma a_{ij} y_j$.

Consider linear transformations from a vector space with basis $\{x_i\}$ to itself. Let f be represented as (a_{ij}). Let $\{w_j\}$ be another basis, where $x_i=\Sigma b_{ij} w_j$, $w_j=\Sigma c_{ji} x_i$. Then $f(w_j)=\Sigma c_{ji} f(x_i)=\Sigma c_{ji} a_{ik} x_k=\Sigma c_{ji} a_{ik} b_{km} w_m$. From $w_j=\Sigma c_{ji} x_i=\Sigma c_{ji} b_{ik} w_k$ it follows that $CB=I$, where I is the matrix

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ & & \cdots\cdots & & \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

is known as a (multiplicative) identity. It acts as a two-sided identity $IA=AI=A$. We also have $x_i=\Sigma b_{ij} w_j=\Sigma b_{ij} c_{jk} x_k$ so $BC=I$. Therefore, $B=C^{-1}$. Here C^{-1} denotes the (multiplicative) inverse of C. Then, expressed in terms of w_j, f is $CAB=CAC^{-1}$. Two matrices X, YXY^{-1} are said to be similar.

A linear transformation $\mathfrak{T}$ represented as (a_{ij}) sends $\Sigma c_i x_i$ to $\Sigma c_i a_{ij} y_j$. The matrix product $(c_i)(a_{ij})$ is $(\Sigma c_i a_{ij})$. A row (column) vector is an $1 \times n (n \times 1)$ matrix. So the linear transformation $\mathfrak{T}$ is equivalent to that given by matrix multiplication on row vectors. Matrices also act as linear transformation on column vectors.

The rank of a matrix is its rank as a linear transformation on row vectors. Taking column vectors gives the same number.

The image space is spanned by the rows of the matrix since $(\Sigma_i c_i a_{ij})$ is the sum of c_i times the ith row of A. Therefore, the row rank is the maximum size of a set of linearly independent rows. The ith row (column) of A is denoted $A_{i*}(A_{*i})$.

In general, an n-square matrix X is invertible if there exists Y with $XY=I$ or $YX=I$. Either equation implies the other and is equivalent to the rank of the matrix being n.

The determinant of a matrix A is defined to be

$$\det(A)=\sum_{\pi \in \mathcal{S}_n} \operatorname{sgn}(\pi) a_{1\pi(1)} a_{2\pi(2)} \cdots a_{n\pi(n)}$$

where the summation is over all permutations π, and $\operatorname{sgn}(\pi)$ denotes the sign of π and so $\operatorname{sgn}(\pi)$ is ± 1 according to whether π is even or odd. As before, a permutation is even or odd according to whether it is the product of an odd or even number of transpositions. It has the following properties, which can be proved in turn. Let $(a_{ij})^{\mathrm{T}}$ denote (a_{ji}), called the transpose of A. This operation changes rows into columns and columns into rows.

(D-1) Since $\operatorname{sgn}(\sigma)=\operatorname{sgn}(\sigma^{-1})$, $\det(A)=\det(A^{\mathrm{T}})$, $\sigma \in \mathcal{S}_n$.
(D-2) If $A_{i*}=B_{i*}=C_{i*}$ for $i \neq k$ and $C_{k*}=A_{k*}+B_{k*}$, then $\det(C)=\det(A)+\det(B)$.
(D-3) If the rows of A are permuted by a permutation π, the determinant of A is multiplied by $\operatorname{sgn}(\pi)$.
(D-4) If two rows (columns) of A are equal, then $\det(A)=0$.
(D-5) If any row is multiplied by k, then the determinant is multiplied by k.
(D-6) If A_{i*} is replaced by $A_{i*}-kA_{j*}$, $i \neq j$, then the determinant is unchanged.
(D-7) If $a_{ij}=0$ for $i>j$, then $\det(A)=a_{11}a_{22}\cdots a_{nn}$.
(D-8) $\det(AB)=\det(A)\det(B)$.
(D-9) $\det(A) \neq 0$ if and only if A has an inverse.
(D-10) Let $A[i|j]$ the submatrix of A obtained by deleting row i and column j. The (i, j)th-cofactor of a_{ij} is $(-1)^{i+j}$ $\det A[i|j]$ and it is denoted as $C[i|j]$:

$$\det(A)=\sum_{j=1}^{n} a_{rj} C[r|j]=\sum_{i=1}^{n} a_{is} C[i|s]$$

Property (D-1) ensures that for properties of determinants stated in terms of their rows, equivalent properties can be stated in terms of columns. These sometimes will not be explicitly mentioned.

If follows from (D-4), (D-9), and (D-10) that if A has an inverse

$$A^{-1}=\frac{1}{\det(A)}(C[i|j]^{\mathrm{T}})$$

From (D-7) and (D-8), $\det(I)=1$, $\det(A^{-1})=1/\det(A)$.

E. Boolean Vectors and Matrices

Most of the theory of the last section holds in the case of Boolean matrices. Boolean row (column) vectors are $1 \times n (n \times 1)$ Boolean matrices. The set V_n is the set of all Boolean n-tuples, and additive homomorphisms preserving 0 from V_n to V_m are in 1–1 correspondence with Boolean matrices.

Matrices over a Boolean algebra or field can be multiplied or added in blocks using the same formulas as

for individual entries $(A_{ij})+(B_{ij})=(A_{ij}+B_{ij})$, and $(A_{ij})(B_{ij})=(\Sigma A_{ik}B_{kj})$. Here the set of all indices has been partitioned into r subsets $S_1, S_2, \ldots, S_r$ and A_{ij} is the submatrix consisting of all a_{km} such that $(k, m) \in S_i \times S_j$. Usually the numbers in each S_t are assumed adjacent. For example,

$$\left[\begin{array}{cc|cc} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \\ \hline 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{array}\right]\left[\begin{array}{cc|cc} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \\ \hline 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{array}\right]$$

$$= \left[\begin{array}{cc} \begin{bmatrix}1&2\\3&4\end{bmatrix}\begin{bmatrix}1&2\\3&4\end{bmatrix}+\begin{bmatrix}5&6\\7&8\end{bmatrix}\begin{bmatrix}0&1\\1&0\end{bmatrix} & \begin{bmatrix}1&2\\3&4\end{bmatrix}\begin{bmatrix}5&6\\7&8\end{bmatrix}+\begin{bmatrix}5&6\\7&8\end{bmatrix}\begin{bmatrix}1&1\\1&1\end{bmatrix} \\ \begin{bmatrix}0&1\\1&0\end{bmatrix}\begin{bmatrix}1&2\\3&4\end{bmatrix}+\begin{bmatrix}1&1\\1&1\end{bmatrix}\begin{bmatrix}0&1\\1&0\end{bmatrix} & \begin{bmatrix}0&1\\1&0\end{bmatrix}\begin{bmatrix}5&6\\7&8\end{bmatrix}+\begin{bmatrix}1&1\\1&1\end{bmatrix}\begin{bmatrix}1&1\\1&1\end{bmatrix} \end{array}\right]$$

A block lower triangular form is one such that $A_{ij}=0$ for $i<j$. Every Boolean matrix can be put uniquely into a block triangular form such that main diagonal blocks A_{ii} are indecomposable or zero. This means $\Sigma_{n=0}^{\infty}A_{ii}^{n}=J$, the matrix consisting entirely of 1 for each i.

If the Boolean matrix is idempotent, every block in this block triangular form consists identically of 0 or of 1.

A Boolean matrix is invertible if and only if it is the matrix of a permutation. The row (column) rank of a Boolean matrix is the number of vectors in row (column) basis. An n-square Boolean matrix is nonsingular if it has row and column rank n and it is regular. Incidentally, the row and column rank of a Boolean matrix are not necessarily equal when $n \geq 4$. For a nonsingular Boolean matrix, there exists an analog of the formula for inverses. The permanent of a Boolean matrix A is

$$\operatorname{per}(A)=\sum_{\pi\in\mathcal{S}_n} a_{1\pi(1)}a_{2\pi(2)}\cdots a_{n\pi(n)}$$

A nonsingular Boolean matrix A has a unique Thierrin–Vagner inverse given by $\operatorname{per}(A[i|j])^{\mathrm{T}}$. The theories of Boolean matrices and semirings found more and more applications related to computers in the 1990s such as communication complexity and Schein rank, image processing, and the study of parallel computation.

F. Characteristic Polynomials

The linear transformations on a vector space are isomorphic (as binary relations) by an isomorphism of the vector space if and only if their matrices are similar. Similarity is an equivalence relation. To describe it completely we need a set of similarity invariants that distinguish any two nonsimilar matrices.

The most important similarity invariant is the characteristic polynomial defined by $p(\lambda)=\det(\lambda I-A)$. If $B=XAX^{-1}$, then

$$\begin{aligned}\det(\lambda I-XAX^{-1}) &= \det(X(\lambda I-A)X^{-1}) \\ &= \det(X)p(\lambda)(\det(X))^{-1} \\ &= p(\lambda)\end{aligned}$$

Therefore, it is a similarity invariant. Its roots are called eigenvalues (characteristic values, latent values).

The matrix A itself satisfies $p(A)=0$. The coefficient of λ^{n-1} is minus Σa_{ii}. This quantity is called the trace of A and it is denoted by $\operatorname{Tr}(A)$. We have $\operatorname{Tr}(AB)=\operatorname{Tr}(BA)$ and the trace is an additive homomorphism. The last coefficient is $(-1)^n\det(A)$.

An eigenvector (characteristic vector, latent vector) v of an n-square matrix is a vector such that $vA=\lambda v$ for some λ. There exists an eigenvector for λ if and only if $v(A-\lambda I)=0$ if and only if $A-\lambda I$ has rank less than n if and only if $p(\lambda)=0$. That is, λ must be an eigenvalue.

V. RINGS

A. Ring of Integers

The ring of integers satisfies the following algebraic axioms: Let $x, y, z \in \mathbf{Z}$. (R-1) $x+y=y+x$; (R-2) $(x+y)+z=x+(y+z)$; (R-3) $x(y+z)=xy+xz$; (R-4) $(y+z)x=yx+zx$; (R-5) $(xy)z=x(yz)$ (R-6) there exists 0 such that $0+x=x$ for all x; (R-7) for all x there exists $-x$ such that $x+(-x)=0$; (R-8) $xy=yx$; (R-9) if $xz=yz$, then $x=y$ or $z=0$; (R-10) there exists 1 such that $1\cdot x=x$ for all x.

Any structure of a set with two binary operations satisfying (R-1) to (R-7) is called a ring. If (R-8) holds, it is called a commutative ring. The sets of n-square matrices over $\mathbf{Z}$, $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{C}$ satisfy (R-1) to (R-7) and (R-10) but not (R-8) and (R-9). Here $\mathbf{Q}$ denotes the set of all rational numbers. A ring satisfying (R-10) is called a ring with unit. A ring satisfying (R-1) to (R-10) is called an integral domain.

In addition the integers satisfy the properties that there exists a set $\mathbf{P} \subset \mathbf{Z}$ called the positive integers such that (R-11) $\mathbf{P}$ is closed under addition, (R-12) $\mathbf{P}$ is closed under multiplication, and (R-13) for all integers exactly one of these holds: $x=0$, $x \in \mathbf{P}$, $-x \in \mathbf{P}$. (Obviously, $\mathbf{P}=\mathbf{Z}^{+}$.) A ring satisfying these is called an ordered ring. $\mathbf{R}$ but not $\mathbf{C}$ is an ordered ring.

The final axiom is the induction property (R-14): If a nonempty set S is contained in $\mathbf{P}$ and $1 \in S$ and $1 + x \in S$ whenever $x \in S$, then $S = \mathbf{P}$.

The ordinary rules of algebra follow from these axioms, as $(1 + (-1) + 1) \cdot (-1) = (0 + 1) \cdot (-1) = 1 \cdot (-1) = -1$. Therefore, $-1 + (-1)(-1) + (-1) = -1$, $1 + (-1) + (-1)(-1) + (-1) + 1 = 1 + (-1) + 1$, and $(-1)(-1) = 1$.

B. Euclidean Domains

Many properties of $\mathbf{Z}$ are shared by the ring of polynomials over any field, and the rings $\{a + bi : a, b \in \mathbf{Z}\}$ and $\{a + b[(1 + \sqrt{3})/2]i : a, b \in \mathbf{Z}\}$. To deal with these we define Euclidean domains. A Euclidean domain is a ring $\mathfrak{E}$ satisfying (R-1) to (R-10) provided with a function $\omega(x)$ defined from nonzero elements of $\mathfrak{E}$ into $\mathbf{Z}$ such that (R-15) $\omega(xy) \geq \omega(x)$; (R-16) for all $x, y \in \mathfrak{E}$ there exist $q, r \in \mathfrak{E}$ such that $x = qy + r$, and either $\omega(r) < \omega(y)$ or $r = 0$; and (R-17) $\omega(x) \geq 0$. For polynomials, let ω be the degree. For other cases, let it be $|x|$.

In any ring with unit $\mathfrak{R}$, the relation $x|y$ can be defined by there exists $z \in \mathfrak{R}$ with $zx = y$. This relation is transitive. If $x|y$, $x|z$, then $x|y + z$ since $y = ax$, $z = bx$, $y + z = (a + b)x$ for some a, b. If $x|y$ then $xa|ya$ for any a. Since $1x = x$, it is a quasiorder. It is called (right) divisibility.

In an integral domain, suppose $y|x$ and $x|y$. Then $x = ry$, $y = sx$ for some r, s. If either is zero then both are. Otherwise $rsx = ry = x = 1x$ and $rs = 1$ by (R-9). So r, s are invertible.

A greatest common divisor (g.c.d.) of a set $S \subset \mathfrak{E}$ is an element $g \in \mathfrak{E}$ such that $g|x$ for all $x \in S$ and if $y|x$ for all $x \in S$ then $y|g$. Any two g.c.d.'s of the same set must be multiples of one another, and so multiples by an invertible element. Let g be a g.c.d. of S and h be a g.c.d. of a set T and m be a g.c.d. of $\{g, h\}$. Then m is a g.c.d. of $S \cup T$. The g.c.d. is a greatest lower bound in the partial order of the set of equivalence classes of elements of $\mathfrak{E}$ under the relation $x \sim ay$ if a is invertible.

In a Euclidean domain, the Euclidean algorithm is an efficient way to calculate the g.c.d. of two elements x, y. Assume $\omega(x) \geq \omega(y)$. Let $d_0 = x$, $d_1 = y$. Choose a sequence d_i, q_i for $i > 0$ by $d_i = q_{i+1}d_{i+1} + d_{i+2}$, where $\omega(d_{i+2}) < \omega(d_{i+1})$. The process terminates since $\omega(x)$ is always a nonnegative integer, and in the last stage we have a remainder $d_{k+2} = 0$. Then $d_{k+1}|d_k$. By induction using the equations $d_i = q_{i+1}d_{i+1} + d_{i+2}$ it will divide the right side and therefore the left side. So it divides all d_i. Also $d_{i+2} = d_i - q_{i+1}d_{i+1}$ is a linear combination of d_i, d_{i+1} so by induction d_{k+1} is a linear combination of a, b. Therefore, if x divides a, b, it divides d_{k+2}. Therefore, the element $d_{k+1} = ra + sb$ is a g.c.d. of a, b.

When an argument is made by induction, it is understood that a statement $P(n)$ holds for $n = 1$ and that there is an argument showing that if $P(1)$, $P(2)$, ..., $P(n-1)$ hold so does $P(n)$. Let $S = \{n \in P : P(n) \text{ is true}\}$. Then by the induction axiom (R-14), S is all of P. Arguments by induction are especially common in the theory of $\mathbf{Z}$.

The following is an example of the Euclidean algorithm for polynomials:

$$x^3 + 1 = x(x^2 + 1) + (-x + 1)$$
$$x^2 + 1 = (-x + 1)(-x - 1) + 2$$
$$-x - 1 = 2\left(\frac{-x - 1}{2}\right)$$

Therefore, 2 (equivalently 1) is a g.c.d. of $x^2 + 1$, $x^3 + 1$. We can now express it in terms of them:

$$2 = (x^2 + 1) - (-x + 1)(-x - 1)$$
$$= (x^2 + 1) - (-x + 1)(x + 1)$$
$$2 = (x^2 + 1) + ((x^3 + 1) - x(x^2 + 1))(x + 1)$$
$$2 = (x^2 + 1) + (x + 1)(x^3 + 1) - x(x + 1)(x^2 + 1)$$
$$2 = (1 - x - x^2)(x^2 + 1) + (x + 1)(x^3 + 1)$$

In the second step, we substituted $x^3 + 1 - x(x^2 + 1)$ for $-x + 1$.

By induction there is a g.c.d. of $x_1, x_2, \ldots, x_k$ that is linear combination $s_1x_1 + s_2x_2 + \cdots + s_kx_k$ of them. If we multiply by invertible elements, we obtain such an expression for any g.c.d.

If a is invertible, then $\omega(ab) \leq \omega(b)$ and $\omega(a^{-1}ab) \leq \omega(ab)$, so the two are equal. Conversely, if $\omega(ab) = \omega(b)$, $b \neq 0$, $b \in \mathfrak{E}$, then a is a unit [divide b by ab; if $\omega(r) < \omega(ab) = \omega(b)$ we have a contradiction].

From this we can establish the basic facts of prime factorizations. An element $p \in \mathfrak{E}$ is called prime if it is not invertible but whenever $p = xy$ one of x, y is invertible. It follows that for any a if $p|a$ then the g.c.d. of p, a is invertible.

Suppose 1 is a g.c.d. of c, a. Let $1 = ar + cs$. If $c|ab$, then $c|abr + cbs = b$. Therefore, if p, a prime, divides xy, it divides x or y. To factor an element y of $\mathfrak{E}$ into primes choose a divisor x that has minimal $\omega(x)$ among nonivertible elements. If $x = ut$ and u is not invertible then $\omega(u) = \omega(x)$ and so t is invertible and so x is prime. Since x is not invertible, $\omega(y/x) < \omega(y)$. So after a finite number of prime factors the process ends.

By induction we can show that if p is a prime and divides $y_1y_2 \cdots y_k$, then p divides some y_i. From this it can be proved that if $a = p_1p_2 \cdots p_m = q_1q_2 \cdots q_n$, where p_i, q_i are primes, then $n = m$ and the q_i can be renumbered so that $p_i = u_iq_i$, where u_i is invertible. Since $p_i|q_1q_2 \cdots q_n$, p_1 divides some q_i. Label it q_1. Since both are primes,

$p_1 = q_1$. This reduces n, m by 1 and the process can be repeated.

C. Ideals and Congruences

A ring homomorphism is a function $f : \mathfrak{R} \to \mathfrak{S}$ satisfying $f(x+y) = f(x) + f(y)$ and $f(xy) = f(x)f(y)$. That is, it is a semigroup homomorphism for multiplication and a group homomorphism for addition.

The following are examples of ring homomorphisms. (1) The mapping from n-square matrices to m-square matrices for $m > n$, which adds to a matrix $m-n$ rows and columns of zero. (2) The mapping $f : \mathbf{Z} \to \mathbf{Z_m}$, $f(x) = \bar{x}$. For any two rings $\mathfrak{R}$, $\mathfrak{S}$ a product ring $\mathfrak{R} \times \mathfrak{S}$ is defined by operations $(a, b) + (c, d) = (a + c, b + d)$, $(a, b)(c, d) = (ac, bd)$. (3) The mapping $\mathfrak{R} \times \mathfrak{S} \to \mathfrak{R}$ sending (a, b) to a (or b) is a ring homomorphism. (4) For any complex number c, the evaluation $f(c)$ is a homomorphism from the ring of all functions $f : \mathbf{C} \to \mathbf{C}$ into $\mathbf{C}$ itself.

A subring of a ring is a subset closed under addition, subtraction, and multiplication. An ideal is a subset S of a ring $\mathfrak{R}$ closed under addition and subtraction (if $x, y \in S$, then $x - y \in S$) and such that if $z \in \mathfrak{R}$, $x \in S$, then $xz, zx \in S$. The set of even integers is an ideal in the set of intergers.

A congruence on a ring $\mathfrak{R}$ is an equivalence relation $x \sim y$ such that for all $a \in \mathfrak{R}$ if $x \sim y$ then $x + a \sim y + a$, $xa \sim ya$, and $ax \sim ay$. For any congruence $\mathfrak{R}$ if $x \sim y$, $z \sim w$ then $x + z \sim y + z \sim y + w$ and $xz \sim yw$. Then addition and multiplication of equivalence classes $\bar{x} + \bar{y} = \overline{x+y}$, $\bar{x}\bar{y} = \overline{xy}$ are well defined and give a ring called the quotient ring. The function $x \to \bar{x}$ is a homomorphism.

There is a 1–1 correspondence between ideals and congruences. To a congruence associate the set $x \sim 0$, which is an ideal $\mathfrak{I}$. Then the congruence is defined by $x \sim y$ if and only if $x + (-y) \sim y + (-y) = 0$. Therefore, $x \sim y$ if and only if $x - y \in \mathfrak{I}$. Every ideal in this way determines a congruence.

The equivalence classes have the form $x + \mathfrak{I} = \{x + m : m \in \mathfrak{I}\}$. The quotient ring is denoted $\mathfrak{R}/\mathfrak{I}$.

All ideals in the integers—in fact, all subgroups—have the form $m\mathbf{Z} = \{mx : x \in \mathbf{Z}\}$. The congruence associated with these is $a \equiv b (\mathrm{mod}\ m)$ if and only if $a - b = km$ for some k. By the general theory, such congruences can be added, subtracted, multiplied, or raised to a power. Congruences can sometimes decide whether or not equations are solvable in whole numbers. For example, $x^2 + y^2 + z^2 = w$ is not possible if $w \equiv 7 (\mathrm{mod}\ 8)$. To prove this it can first be noted that $x^2 \equiv K (\mathrm{mod}\ 8)$ where $k = 0, 1, 4$, and no three elements of {0, 1, 4} add to 7(mod 8). The ring of all polynomials in a variable x having coefficients in a field $\mathbf{K}$ is denoted $\mathbf{K}[x]$. As a set it is a subset of a Cartesian product of countably many copies of $\mathbf{K}$, sequences $(a_0, a_1, \ldots)$ such that only finitely many a_i are nonzero. The a_i are the coefficients of the polynomial and multiplication can be defined by $(a_n)(b_n) = (\Sigma_r a_r b_{n-r})$.

In it or any Euclidean domain, all ideals have the form $\{f(x)y : y \in \mathbf{K}[x]\}$ for some polynomial $f(x)$. Given an ideal $\mathfrak{I}$, take a nonzero element $f(x)$ in it such that $\omega(f(x))$, the degree, is minimal.

If $f(x)$ is another element of the ideal, choose $q(x)$, $r(x)$ such that $r(x) = 0$ or $\omega(r(x)) < \omega(f(x))$ for $g(x) = f(x)q(x) + r(x)$. Then $r(x) = g(x) - f(x)q(x) \in \mathfrak{I}$ since $g(x)$, $f(x)$ do. So $\omega(r(x)) < \omega(f(x))$ contradicts minimality of $\omega(f(x))$. So $r(x) = 0$. So every element in $\mathfrak{I}$ is a multiple of $f(x)$. Conversely multiples of $f(x)$ belong in $\mathfrak{I}$ since it is an ideal.

Ideals of the form $a\mathfrak{R} = \{ax : x \in \mathfrak{R}\}$ are called principal, and a is said to be a generator. Therefore, Euclidean domains are principal ideal domains, that is, integral domains in which all ideals are principal.

In all principal ideal domains, all elements can be uniquely factored in terms of primes and invertible elements. Let $\mathfrak{R}_m$ where $m \in \mathbf{Z}$ is not divisible by the square of a prime denote $\{a + [b(1 + \sqrt{m})/2] : a, b \in \mathbf{Z}\}$ or $\{a + b\sqrt{m} : a, b \in \mathbf{Z}\}$ according to whether $m \equiv 1 (\mathrm{mod}\ 4)$ or not. If $0 < m < 25$, then $\mathfrak{R}_m$ is a principal ideal domain if and only if $m \neq 10, 15$. If $-130 < m < 0$, then $\mathfrak{R}_m$ is a principal ideal domain if $m = -1, -2, -3, -7, -11, -19, -43, -67$. (This is a result proved by techniques of algebraic number theory.) Unique factorization does not hold in $\mathfrak{R}_m$ if it is not a principal ideal domain.

For noncommutative rings, there are separate concepts of right and left ideals. A subset S of a ring $\mathfrak{R}$ closed under addition and subtraction is a left (right) ideal if and only if $ax \in S$ ($xa \in S$) for all $x \in S$, $a \in \mathfrak{R}$. The ring of n-square matrices has no (two-sided) ideals except itself and zero but for any subspace $\mathbf{W}$, $\mathfrak{I} = \{M \in M_n(\mathbf{F}) : vM \in \mathbf{W}\}$ is a right ideal. A proper ideal is an ideal that is a proper subset. The trivial ideal is {0}.

Suppose a ring $\mathfrak{R}$ has not all products zero and has no proper nontrivial right ideals. If $ab \neq 0$ for some b, then $\{x : ax = 0\}$ is a proper right ideal and is therefore zero. The set $\{ax\}$ is a nonzero right ideal and is therefore $\mathfrak{R}$. So multiplication on the left by a is 1–1 and onto.

Let $\mathfrak{I} = \{a : ab = 0 \text{ for all } b \in \mathfrak{R}\}$. Then $\mathfrak{I}$ is a proper right ideal and is therefore 0. So for all $a \neq 0$ left multiplication by a is 1–1 and onto. So if $ab = 0$, then $a = 0$ or $b = 0$.

For any $a \neq 0$ since $\{ax\} = \mathfrak{R}$, there exist $\boldsymbol{e}$, a^{-1} such that $a\boldsymbol{e} = a$, $aa^{-1} = \boldsymbol{e}$. Then for any $x \in \mathfrak{R}$, $a\boldsymbol{e}x = ax$. So $\boldsymbol{e}x = x$ since left multiplication is 1–1. From $ab \neq 0$ for $a \neq 0$, $b \neq 0$, it follows that right multiplication is also 1–1. Since $x\boldsymbol{e}x = xx$ and $x\boldsymbol{e} = x$, $\boldsymbol{e}$ is an identity. From $aa^{-1}a = \boldsymbol{e}a = a = a\boldsymbol{e}$, it follows that a^{-1} is an inverse of a. Therefore, in $\mathfrak{R}$, all nonzero elements have inverses.

A ring with unit in which all nonzero elements have inverses is called a division algebra. The best known division algebra that is not a field is the ring of quaternions. As a set, it is $\mathbf{R}\times\mathbf{R}\times\mathbf{R}\times\mathbf{R}$. Elements are written as $a+b\,\mathfrak{i}+c\,\mathfrak{j}+d\,\mathfrak{k}$. Expressions are multiplied by expanding and using the rules $\mathfrak{i}^2=\mathfrak{j}^2=\mathfrak{k}^2=-1$, $\mathfrak{i}\mathfrak{j}=\mathfrak{k}, \mathfrak{j}\mathfrak{k}=\mathfrak{i}, \mathfrak{k}\mathfrak{i}=\mathfrak{j}, \mathfrak{j}\mathfrak{i}=-\mathfrak{k}, \mathfrak{k}\mathfrak{j}=-\mathfrak{i}, \mathfrak{i}\mathfrak{k}=-\mathfrak{j}$. Any multiplication defined from a basis by expanding terms is distributive. The multiplication on basis elements and their negative $\pm1, \pm\mathfrak{i}, \pm\mathfrak{j}, \pm\mathfrak{k}$ is a finite group of order 8. This implies associativity.

D. Structure of $\mathbf{Z_n}$

The ring $\mathbf{Z_n}$ is the ring of congruence classes modulo n. It is the quotient ring $\mathbf{Z}/\Im_n$, where $\Im_n=\{nx : x\in\mathbf{Z}\}$. For any quotient ring $\mathfrak{R}/\Im$, ideals of the quotient ring are in 1–1 correspondance with ideals of $\mathfrak{R}$ containing $\Im$. An ideal $\Im_m\supset\Im_n$ if and only if $m|n$. So the ideals of $\mathbf{Z_n}$ are in 1–1 correspondence with positive integers m dividing n. The ring $\mathbf{Z_n}$ has exactly n elements, since each integer x can be expressed as $nk+r$ for a unique $r\in\{0,1,\ldots,n-1\}$, and then $x\equiv r(\mathrm{mod}\ n)$. It has a unit and is commutative.

Two elements of an integral domain are called relatively prime if 1 is a g.c.d. of them. If x, n are relatively prime, then $rx+sn=1$ for some r and s. Thus, $\bar r\bar x=1$ so $\bar x$ is invertible. Conversely, if $\bar r\bar x=1$, then $rx-1=sn$ for some s, and 1 is a g.c.d. of x and n. Thus, the number of invertible elements equals $\Phi(n)$, the number of positive integers $m<n$ that are relatively prime to n.

The Chinese remainder theorem asserts that if all pairs of $n_1, n_2, \ldots, n_m$ are relatively prime and $a_i\in\mathbf{Z}$, $i=1$ to m, then there exists $x\in\mathbf{Z}$ and $x\equiv a_i(\mathrm{mod}\ n_i)$ for $i=1$ to m. Any two such x are congruent modulo $n_1n_2\cdots n_m$.

The set of invertible elements is closed under products since $(xy)^{-1}=y^{-1}x^{-1}$ and contains inverses and identity, so it forms a group for any ring. Here it is a finite group. For any invertible element x, the elements $1, x, \ldots, x^{k-1}$ form a subgroup if k is the order of x. By the Lagrange's theorem, the order of any subgroup of a group divides the order of the group. Therefore, $k|\Phi(n)$. From $x^k=1$, follows $x^{\Phi(n)}=1$. This proves a theorem of Euler and Fermat. For any $x\in\mathbf{Z_n}$, if x is invertible, then $x^{\Phi(n)}=1$.

If p is prime, then $1, 2, \ldots, p-1$ are all relatively prime to p, so $\Phi(p)=p-1$, and $x^{p-1}\equiv1(\mathrm{mod}\ p)$ if $x\not\equiv0$). Then $x^p\equiv x(\mathrm{mod}\ p)$ for all $x\in\mathbf{Z_p}$. Assume p is prime for the rest of this section.

The multiplicative group $\mathbf{Z_p^*}=\{\bar x\not\equiv\bar0 : \bar x\in\mathbf{Z_p}\}$ is a commutative group of order $p-1$. The ring $\mathbf{Z_p}$ is a field since $\mathbf{Z_p^*}$ is a group. Polynomials over $\mathbf{Z_p}$ can be uniquely factored into primes.

Over any field $\mathbf{K}$, if a polynomial $p(x)$ satisfies $p(k)=0$, where $k\in\mathbf{K}$, then let $p(x)=(x-k)q(x)+r$. By substitution $x=k$ we find $r=0$. So if $p(k)=0$, then $(x-k)|p(x)$. Thus, a polynomial of order n cannot have more than n distinct roots since each root gives a degree 1 factor $(x-k)$.

The polynomial $x^{p-1}-1$ has exactly $p-1$ roots in $\mathbf{Z_p}$, that is, $1, 2, \ldots, p-1$. Thus, it factors as $c(x-1)\cdots(x-(p-1))$. Since the, coefficient of x^{p-1} is 1, we have $c=1$. The polynomial x^r-1 divides $x^{p-1}-1$ for $r|p-1$, so by unique factorization it factors into linear factors and has exactly r roots. For any prime power r^t dividing $p-1$ take a root y_r of $x^{r^t}=1$, which is not a root of $x^{r^{t-1}}=1$. Then y_r has order precisely r^t. Take t maximal. The product u of all y_r will have order the least common multiple of r^t, which is $p-1$. Therefore, $u, u^2, \ldots, u^{p-1}$ are distinct and are all of $\mathbf{Z_p^*}$ since it has exactly $p-1$ elements. This proves $\mathbf{Z_p^*}$ is cyclic. Therefore, it is isomorphic to $(\mathbf{Z_{p-1}}, +)$.

In $\mathbf{Z_n}$ for $m|n$, an element y is a multiple of m if and only if $(n/m)y\equiv0$. Therefore, in $\mathbf{Z_p^*}$, an element y is an mth power for $m|p-1$ if and only if $y^{(p-1)/m}\equiv1(\mathrm{mod}\ p)$. The mth powers form a multiplicative subgroup of order $(p-1)/m$.

E. Simple and Semisimple Rings

A ring is called simple if it has no proper nontrivial (two-sided) ideals. A commutative simple ring is a field. It is simplest to treat the case of finite dimensional algebras. An algebra over a field $\mathbf{F}$ is a ring $\mathfrak{R}$ provided with a multiplication $\mathbf{F}\times\mathfrak{R}\to\mathfrak{R}$ such that (1) $(ax)y=a(xy)=x(ay)$ for all $a\in\mathbf{F}$, $x, y\in\mathfrak{R}$; and (2) $\mathfrak{R}$ is a vector space over $\mathbf{F}$. The quaternions, the complex numbers, and the ring of all functions $\mathbf{R}$ to $\mathbf{R}$ are all algebras over $\mathbf{R}$.

A division algebra is an algebra that is a division ring. Division algebras can be classified in terms of fields. A field $\mathbf{F}$ is called algebraically closed if every nonzero polynomial $p(x)=a_0x^n+a_1x^{n-1}+\cdots+a_nx^0$, $a_i\in\mathbf{F}$, $a_0\neq0$, $n\neq0$ has a root $r\in\mathbf{F}$. Suppose we have a division algebra $\mathfrak{D}$ over an algebraically closed field $\mathbf{F}$ of finite dimension n. Let $a\in\mathfrak{D}$. Then $1, a, \ldots, a^n$ are $n+1$ elements in an n-dimensional vector space so they cannot be linearly independent (they could be extended to a basis). So some combination $p(a)=c_0+c_1a+\cdots+c_na^n=0$. Choose the degree of p minimal. Let $\deg(p)$ denote the degree of p. If $\mathrm{Deg}(p)>1$, then p can be factored over $\mathbf{F}$. Each factor is nonzero evaluated at a, but their product is zero. This is impossible in a division algebra where nonzero elements are invertible. Therefore, $\mathrm{Deg}(p)=1$, so $c_0+c_1a=0$, as $a\in\mathbf{F}$. So $\mathfrak{D}=\mathbf{F}$. Any such division algebra coincides with $\mathbf{F}$. This applies to $\mathbf{F}=\mathbf{C}$.

A finite dimensional $\mathbf{F}$-algebra is simple if and only if it is isomorphic to the ring $M_n(\mathfrak{D})$ of n-square matrices over $\mathfrak{D}$. More generally, every ring having no infinite strictly

decreasing sequence of distinct left ideals (Artinian ring) that is simple is an $M_n(\mathfrak{D})$.

A finite dimensional **F**-algebra $\mathfrak{A}$ is called semisimple if as a vector space it is an internal direct sum of minimal right ideals $\mathfrak{R}_i$. This means every $x \in \mathfrak{A}$ can be uniquely expressed as Σx_i, $x_i \in \mathfrak{R}_i$. Left ideals can equivalently be used.

This is equivalent to any of the following conditions. (1) The multiplicative semigroup of $\mathfrak{R}$ is regular; (2) every left ideal of $\mathfrak{A}$ is generated by an idempotent; and (3) $\mathfrak{A}$ has no two-sided ideals $\mathfrak{I}$ such that $\mathfrak{I}^m = \{x_1 x_2 \cdots x_m : x_i \in \mathfrak{I}\}$ is zero for some m. Ideals $\mathfrak{I}$ such that $\mathfrak{I}^m = 0$ are called nilpotent.

A finite dimensional semisipmle **F**-algebra is isomorphic to a direct sum of rings $M_n(\mathfrak{D})$.

If $\mathfrak{I}$ and $\mathfrak{J}$ are nilpotent two-sided ideals, so is $\mathfrak{I} + \mathfrak{J}$ since $(\mathfrak{I} + \mathfrak{J})^{2^n} \subset \mathfrak{I}^n + \mathfrak{J}^n$. This implies that every finite dimensional algebra has a maximal two-sided nilpotent ideal, the Jacobson radical, and its quotient ring by this ideal is semisimple.

All finite division rings are fields.

F. Group Rings and Group Representations

A group ring of a group G over a ring $\mathfrak{R}$ is formally the set of all functions $f : G \to \mathfrak{R}$ such that $|\{g : f(g) \neq 0\}|$ is finite. Multiplication is given by $(fh)(x) = \Sigma_{yz=x} f(y)h(z)$ and addition is the usual sum of functions. Distributivity follows from the linear form of this expression. Associativity follows from $\{(u, v, w) : uv = y \text{ and } yw = x \text{ for some } y\} = \{(u, v, w) : vw = z, uz = x \text{ for some } z\}$. In fact, semigroups also have semigroup rings.

The group ring can be thought of as the set of all sums $r_1 g_1 + r_2 g_2 + \cdots + r_n g_n$ of group elements with coefficients in $\mathfrak{R}$. For coefficients in **Z**, we have, for instance, $(2g + e)(3e - 2g) = 3e - 2g + 6g - 4g^2 = 3e + 4g - 4g^2$.

A representation of a group G is a homomorphism h from G into the ring $M_n(\mathbf{F})$ of n-square matrices over **F**. Two representations f, h are called equivalent if there exists an invertible matrix $M \in M_n(\mathbf{F})$ with $f(g) = Mh(g)M^{-1}$ for all $g \in G$. This is an equivalence relation.

Every representation f of a group G defines a ring homomorphism $h : F(G) \to M_n(\mathbf{F})$ by $h(\Sigma r_i g_i) = \Sigma r_i h(g_i)$ such that $h(1) = I$ where I is an n-square identity matrix. This is a 1–1 correspondence since if h is a ring homomorphism $h(g)$ gives a group representation.

For $\mathbf{F} = \mathbf{C}$, every group representation is equivalent to a unitary representation, that is, one in which every matrix $f(g)$ satisfies $f(g)f(g)^{*\mathrm{T}} = I$. Here * is complex conjugation. Define a modified inner product on row vectors by $b(x, y) = \Sigma_g x f(g)(yf(g))^{*\mathrm{T}}$. Then $b(x, y) = b(xf(h), yf(h))$ for $h \in G$ since multiplication by h permutes the group elements g and so permutes the terms in the sum. We have $b(x, x) \geq 0$ and $b(x, x) = 0$ only if $x = 0$ since each term will be a vector $vv^{*\mathrm{T}} = \Sigma |v_i^2|$. We can inductively choose a basis u_i for row vectors such that $b(u_i, u_i) = 1$, $b(u_i, u_j) = 0$ if $i \neq j$. In this basis, let $u_i f(g) = \Sigma u_k a_{ki}$. Then $b(u_i, u_j) = b(\Sigma u_k a_{ki}, \Sigma u_k a_{kj}) = \Sigma a_{kj}^* a_{ki} b(u_k, u_k)$. So $\Sigma a_{kj}^* a_{ki} = 1, 0$ according to whether $i = j$ or $i \neq j$. This proves the matrix is now unitary.

For $\mathbf{F} = \mathbf{R}$, essentially the same can be done. A real unitary matrix is called orthogonal. Orthogonal matrices are geometrically products of rotations and reflections.

G. Modules

A module differs from a vector space only in that the ring involved may not be a field. A left (right) $\mathfrak{R}$-module is a set $\mathfrak{M}$ provided with a binary operation denoted as addition and a multiplication $\mathfrak{R} \times \mathfrak{M} \to \mathfrak{M}$ ($\mathfrak{M} \times \mathfrak{R} \to \mathfrak{M}$), where $\mathfrak{R}$ is a ring such that the axioms in Table III hold for all $x, y \in \mathfrak{M}$, $r, s \in \mathfrak{R}$, and some $0 \in \mathfrak{M}$.

For group representations, we consider only unitary modules, those in which $1x = x$ for all $x \in \mathfrak{M}$. Henceforth, we consider left modules.

A homomorphism of modules is a mapping $f : \mathfrak{M} \to \mathfrak{R}$ such that $f(x + y) = f(x) + f(y)$ and $f(rx) = rf(x)$ for all $x, y \in \mathfrak{M}$, and $r \in \mathfrak{R}$. A submodule of $\mathfrak{M}$ is a subset closed under addition, subtraction, and multiplication by elements of $\mathfrak{R}$. For any submodule $\mathfrak{N} \subset \mathfrak{M}$, the equivalence relation $x \sim y$ if and only if $x - y \in \mathfrak{N}$ is a module congruence, that is, $x + z \sim y + z$, $rx \sim ry$ for all $z \in \mathfrak{M}$, $r \in \mathfrak{R}$. Then the equivalence classes form a new module called the quotient module.

Modules are a type of algebraic action, that is, a mapping $G \times S \to S$ for a structure G and set S. Figure 8 classifies some of these.

Direct sums of modules are defined as for vector spaces. Unlike the vector space case not every finite dimensional module over a general algebra is isomorphic to a direct sum $\mathfrak{R} \oplus \mathfrak{R} \oplus \cdots \oplus \mathfrak{R}$ of copies of $\mathfrak{R}$ (with operations from $\mathfrak{R}$). If a ring is an **F**-algebra, all modules over it are vector spaces and this determines their dimension.

TABLE III Module Axioms

Left module	Right module
$(x + y) + z = x + (y + z)$	$(x + y) + z = x + (y + z)$
$x + 0 = x$	$0 + x = x$
$0x = 0$	$x0 = 0$
There exists $-x$ such that $x + (-x) = 0$ for all x	There exists $-x$ such that $x + (-x) = 0$ for all x
$r(x + y) = rx + ry$	$(x + y)r = xr + yr$
$(r + s)x = rx + sx$	$x(r + s) = xr + xs$
$r(sx) = (rs)x$	$(xs)r = x(sr)$

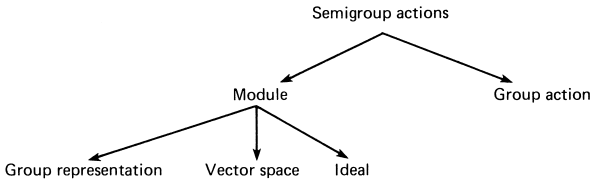


FIGURE 8 Algebraic actions.

There is a 1–1 correspondence between equivalence classes of representations in $M_n(\mathbf{F})$ and isomorphism classes of modules over $\mathbf{F}(G)$ of dimension n, where $\mathbf{F}(g)$ is the group ring of G over $\mathbf{F}$. For any group representation, the set of column vectors is a left module over $\mathbf{F}(G)$ with action given by $(\Sigma c_i g_i)v = \Sigma c_i h(g_i)v$. For any module of dimension n, the group action gives a set of linear transformations of a vector space. If we choose a basis, the matrices of these linear transformations give a group representation.

H. Irreducible Representations

A module is called irreducible if it has no proper nonzero submodules. Suppose $\mathfrak{N} \subset \mathfrak{M}$, where $\mathfrak{M}$, $\mathfrak{N}$ are finite dimensional $\mathbf{F}(G)$-modules. Then there exists a vector space complement $\mathbf{V}$ of $\mathfrak{M}$ in $\mathfrak{N}$. Then there is a natural isomorphism $f : \mathbf{V} \to \mathfrak{M}/\mathfrak{N}$ of vector spaces. Consider $f^{-1} : \mathfrak{M}/\mathfrak{N} \to \mathbf{V} \subset \mathfrak{M}$. Assume $|G| \neq 0$ in $\mathbf{F}$ (if $\mathbf{F}$ is $\mathbf{R}$ or $\mathbf{C}$, for instance). Let

$$h(x) = \frac{1}{|G|} \sum g^{-1} f^{-1}(gx)$$

Then $h(gx) = gh(x)$, so h is a module homomorphism. If we let h_1 be $h : \mathfrak{M}/\mathfrak{N} \to \mathfrak{M}$ composed with the quotient mapping $\mathfrak{M} \to \mathfrak{M}/\mathfrak{N}$, h_1 is the identity since the summand $g^{-1} f^{-1}(gx) = f^{-1}(x)$ because we have a module isomorphism. So h is 1–1. Its image has dimension $\mathrm{Dim}(\mathbf{V})$ and intersection $\{0\}$ with $\mathfrak{N}$ since h_1 is 1–1. Thus the image of h gives a module isomorphic to $\mathfrak{M}/\mathfrak{N}$ whose intersection with $\mathfrak{N}$ is $\{0\}$. Then $\mathfrak{M} \simeq \mathfrak{N} \oplus \mathrm{Im}(h)$, where $\mathrm{Im}(h)$ is the image of h.

Therefore, all representations are direct sums of irreducible modules since we can repeatedly express a module as a direct sum.

If we regard right ideals as modules, the group ring is then a direct sum of irreducible left ideals. These ideals are minimal. Their number is at most the dimension of $\mathfrak{R}(G)$. Therefore, $\mathfrak{R}(G)$ is semisimple and is isomorphic to a direct sum of matrix rings $M_n(\mathfrak{D})$. If $\mathbf{F} = \mathbf{C}$, then $\mathfrak{D} = \mathbf{C}$ so it is isomorphic to a direct sum of rings $M_n(\mathbf{C})$.

For any irreducible module $\mathfrak{N}$, let $0 \neq x \in \mathfrak{N}$. The mapping $\Sigma c_i g_i \to \Sigma c_i g_i$ gives a mapping $\mathfrak{R}(G)$ to $\mathfrak{N}$. This must be nonzero on some left ideal $\mathfrak{M}$ in a given direct sum decomposition $\mathfrak{R}(G) = \oplus \mathfrak{M}_i$. The mapping $\mathfrak{M} \to \mathfrak{N}$ has kernel a proper submodule and image a nontrivial submodule. So by irreducibility it has kernel zero and image $\mathfrak{N}$. Therefore, every irreducible module is isomorphic to a member of the finite set $\{\mathfrak{M}_i\}$.

Further use of these facts gives the basic facts about group representations such as orthogonality for group characters. The character of a representation is the function assigning to each group element g the trace of the matrix $h(g)$.

VI. FIELDS

A. Integral Domains

An integral domain is a commutative ring with unit having cancellation. Any subring of a field must be an integral domain.

Conversely, let $\mathfrak{R}$ be an integral domain. We construct a field, the field of fractions containing $\mathfrak{R}$. Let $S = \mathfrak{R} \times (\mathfrak{R} \setminus \{0\})$. We consider $(a, b) \in S$ as a/b. Therefore, since $a/b = c/d$ if $ad = bc$, we define an equivalence relation by $(a, b) \sim (c, d)$ if and only if $ad = bc$. Transitivity follows for $(a, b) \sim (c, d) \sim (e, f)$ since we have $ad = bc$, $cf = de$; therefore, $adf = bcf = bde$. By cancellation of $d \neq 0$, $af = be$. The relation is a multiplicative congruence if we define products by $(a, b)(c, d) = (ac, bd)$. This shows that equivalence classes can be multiplied under this definition.

Addition is defined by $(a, b) + (c, d) = (ad + bc, bd)$ by analog with ordinary fractions. A computation shows that the equivalence relation is also a congruence with respect to addition and that associativity and distributivity hold. The elements (0, 1) and (1, 1) are additive and multiplicative identities. A nonzero element (a, b) has the additive inverse $(-a, b)$ and the multiplicative invese (b, a). Therefore, the set of equivalence classes is a field.

If $\mathfrak{R}$ is the ring of polynomials over a field $\mathbf{K}$ in one variable, then the field is the field of rational functions.

B. Fields and Extensions

Let $(\mathbf{F_1}, *, \square)$ be a field. A subset $\mathbf{F_2}$ of $\mathbf{F_1}$ is a called a subfield $\mathbf{F_1}$ if the elements of $\mathbf{F_2}$ form a field under $*$ and $\square$.

The rings $(\mathbf{Z_p}, +, \cdot)$, where $\mathbf{p}$ is a prime number and $(\mathbf{Q}, +, \cdot)$ are fields. Let $\mathbf{F}$ be any field. If no sum of $n1 = (1 + 1 + \cdots + 1)$ is zero, then $\mathbf{Q}$ is isomorphic to the subfield $\{n1/m1 : m \neq 0\}$ of $\mathbf{F}$. If $n1 = 0$ where n is minimal, n must be a prime p; else we would have zero divisors $(p1)(q1) = 0$. Then $\mathbf{Z_p} = \{n1\}$ is a subfield of $\mathbf{F}$.

Let $\mathbf{E}$, $\mathbf{F}$ be fields. If $\mathbf{F} \subset \mathbf{E}$, then $\mathbf{E}$ is called an extension of $\mathbf{F}$. All fields are extensions of $\mathbf{Q}$ or $\mathbf{Z_p}$. The field $\mathbf{E}$ is a

vector space over $\mathbf{F}$. The dimension of $\mathbf{E}$ over $\mathbf{F}$ is denoted $[\mathbf{E} : \mathbf{F}]$ and is called the degree of the extension.

Suppose $\mathbf{F} \subset \mathbf{E} \subset \mathbf{D}$. Let x_i be a basis for $\mathbf{E}$ over $\mathbf{F}$ and y_i a basis for $\mathbf{D}$ over $\mathbf{E}$. Then every element of $\mathbf{D}$ can be written uniquely as $\Sigma c_i y_i$, $c_i \in \mathbf{E}$, and in turn this can be written uniquely as $\Sigma(f_{ij}x_j)y_i$, $f_{ij} \in \mathbf{F}$. So $x_j y_i$ are a basis and $[\mathbf{E}:\mathbf{F}][\mathbf{D}:\mathbf{E}] = [\mathbf{D}:\mathbf{F}]$.

Let $\mathbf{E}$ be an extension of $\mathbf{F}$ and let $t \in \mathbf{E}$. Then $\mathbf{F}(t)$ denotes the subfield of $\mathbf{E}$ generated by $\mathbf{F}$, that is, all elements of $\mathbf{E}$ that can be obtained from elements of $\mathbf{F}$, t by adding, multiplying, subtracting, and dividing.

Suppose $\mathbf{E}$ is a finite dimensional extension of degree n. Let x denote an indeterminate. Then the elements $1, t, t^2, \ldots, t^n$ are $n+1$ elements and so are linearly dependent. Then some sum $\Sigma_{i=0}^{n} a_i t^i = 0$. We can choose n minimal and assume by dividing by a_n that $a_n = 1$. If $a_n = 1$, we say that a polynomial is monic. We then get a polynomial $p(t) = 0$ of minimal degree called the minimal polynomial of t. If $f(t) = 0$ for any other polynomial $f(x)$, divide $p(x)$ into $f(x)$. Then $r(t) = f(t) - q(t)p(t) = 0$. If $r(x) \neq 0$ it would have lower degree than $p(x)$ by minimality. Therefore, $p(x)$ divides every other polynomial $f(x)$ such that $f(t) = 0$. Also by minimality $1, t, t^2, \ldots, t^{n-1}$ are linearly independent. The polynomial $p(x)$ must also be prime; else if $p(x) = r(x)s(x)$ then one of $r(t), s(t) = 0$. Prime polynomials are called irreducible.

We next show that $1, t, t^2, \ldots, t^{n-1}$ are a basis for $\mathbf{F}(t)$. They are linearly independent. Since $t^n = \Sigma_{i=0}^{n-1} a_i t^n$, $t^{n+j} = \Sigma_{i=0}^{n-1} a_i t^{i+j}$, every power of t can be expressed as a linear combination of $1, t, t^2, \ldots, t^{n-1}$. Therefore, their span is a subring. Suppose $f(t) \neq 0$. Then $p(x)$ does not divide $f(x)$. Since $p(x)$ is prime, 1 is a g.c.d. of $p(x)$, $f(x)$. For some $r(x)$, $s(x)$, $r(x)p(x) + s(x)f(x) = 1$. So $s(t)f(t) = 1$ and $s(t)$ is an inverse of $f(t)$. Therefore, the span of $1, t, t^2, \ldots, t^{n-1}$ is closed under division and is a field. So it must be the field $\mathbf{F}(t)$ generated by t.

There exists a homomorphism h from the ring of polynomials $\mathbf{F}[x]$ to $\mathbf{F}(t)$ defined by $h(f(x)) = f(t)$. From the usual laws of algebra it is a ring homomorphism that is also onto.

The kernel is the set of polynomials $f(x)$ such that $f(t) = 0$. We have already shown that this is the set of polynomials divisible by $p(x)$, or the ideal $p(x)\mathbf{F}(x)$ generated by $p(x)$.

Therefore, in summary of these results, let $p(x)$ be the minimum polynomial, of degree n, of t. Then $[\mathbf{F}(t):\mathbf{F}] = n$ with basis $1, t, t^2, \ldots, t^{n-1}$. The field $\mathbf{F}(t)$ is isomorphic to the quotient ring $\mathbf{F}[x]/p(x)\mathbf{F}[x]$, where $\mathbf{F}[x]$ is the ring of all polynomials in an indeterminate x.

Conversely, let $p(x)$ be a monic irreducible polynomial. Then we will show that $\mathbf{F}[x]/p(x)\mathbf{F}[x]$ is an extension field of $\mathbf{F}$ in which $t = \bar{x}$ has minimum polynomial $p(x)$. For any $f(x)$ such that $f(\bar{x}) \neq 0$, there will exist $r(x)$, $s(x)$ such that $r(x)f(x) + s(x)p(x) = 1$ since f, p have g.c.d. 1. So $f(t)$ has an inverse $r(t)$ and the quotient ring is a field.

C. Applications to Solvability and Constructibility

The problem of solving quadratic equations goes at least back to the Babylonians. In the ninth century, the Muslim mathematician Al-Khwarismi gave a version of the modern quadratic formula. In the mid-sixteenth century, Italian mathematicians reduced the solution of a cubic equation to the form

$$x^3 + mx = n$$

by dividing it by the coefficient of x^3 and making a substitution replacing x by some $x - c$. They then solved it as

$$x = a - b$$

$$a = \sqrt[3]{(n/2) + \sqrt{(n/2)^2 + (m/3)^3}}$$

$$b = \sqrt[3]{-(n/2) + \sqrt{(n/2)^2 + (m/3)^3}}.$$

They proceeded to solve quartic equations by reducing them to cubics. But no one was able to solve the general quintic using nth roots and in 1824 N.H. Abel proved this is impossible. E. Galois in his brief life proved this also, as part of a general theory that applies to all polynomials. This is based on field theory, and we describe it next.

In the extensions $\mathbf{F}(t)$, one root of a polynomial $p(t)$ has been added, or adjoined, to $\mathbf{F}$. Extensions obtained by adding all roots of a polynomial are called normal extensions. The roots can be added one at a time in any order.

Finite dimensional normal extensions can be studied by finite groups called Galois groups. The Galois group of a normal extension $\mathbf{F} \subset \mathbf{E}$ is the group of all field automorphisms of $\mathbf{E}$ that are the identity on $\mathbf{F}$. It will, in effect, permute the roots of a polynomial whose roots generate the extension. For example, let $\mathbf{F} = \mathbf{Q}(\xi)$ and let $\mathbf{E} = \mathbf{F}(\sqrt[3]{2})$, where $\xi = (-1 + i\sqrt{3})/2$. Then an automorphism of $\mathbf{E}$ exists taking $\sqrt[3]{2} \to \xi\sqrt[3]{2}$, $\xi\sqrt[3]{2} \to \xi^2(\sqrt[3]{2})$, $\xi^2(\sqrt[3]{2}) \to \sqrt[3]{2}$. The Galois group is cyclic of order 3, generated by this automorphism. Since the ratio ξ of two roots goes to itself, it is the identity on $\mathbf{Q}(\xi)$.

The order of the Galois group equals the degree of a normal extension. Moreover, there is a 1–1 correspondence between subfields $\mathbf{F} \subset \mathbf{K} \subset \mathbf{E}$ and subgroups of $H \subset G$, the Galois group of $\mathbf{E}$ over $\mathbf{F}$. To a subgroup H is associated the field $\mathbf{K} = \{x \in \mathbf{E} : f(x) = x \text{ for all } f \in \mathbf{K}\}$.

A splitting field of a polynomial p over a field $\mathbf{F}$ is a minimal extension of $\mathbf{F}$ over which p factors into factors

of degree 1. It is a normal extension and any two splitting fields are isomorphic.

Suppose a polynomial p is solvable by radicals over $\mathbf{Q}$. Let $\mathbf{E}$ be the splitting field of p over $\mathbf{Q}$. Each time we extract a radical the roots of the radical generate a normal extension $\mathbf{F_1}$ of a previous field $\mathbf{F_2}$. Let $\mathbf{E_i} = \mathbf{F_i} \cap \mathbf{E}$. Then $\mathbf{F_2}$ over $\mathbf{F_1}$ has cyclic Galois group, so $\mathbf{E_2}$ over $\mathbf{E_1}$ does also.

It follows that there exist a series of extensions $\mathbf{Q} = \mathbf{D_0} \subset \mathbf{D_1} \subset \cdots \subset \mathbf{D_n} = \mathbf{E}$ each normal over the preceding such that the Galois group of each over the other is cyclic. It follows that the Galois group G has a series of subgroups $G_n = \{e\} \subset G_{n-1} \subset \cdots \subset G_0 = G$ such that g_i is a normal subroup of G_{i-1} with cyclic quotient group. Such a group is called solvable.

The symmetric group of degree 5 has as its only nontrivial proper normal subgroup the alternating group which is simple. Therefore, it is not solvable. If $\mathbf{F}(x)$ is a degree 5 irreducible polynomial over $\mathbf{Q}$ with exactly two nonreal roots, there exists an order 5 element in its Galois group just because 5 divides the degree of the splitting field. Complex conjugation gives a transposition. Therefore, the Galois group is $\mathcal{S}_5$. So polynomials of degree 5 cannot in general be solved by radicals.

Conversely, it is true that every normal extension $\mathbf{E} \subset \mathbf{F}$ with cyclic Galois group can be generated by radicals. It can be shown that there is a single element θ such that $\mathbf{E} = \mathbf{F}(\theta)$ (consider all linear combinations θ of a basis for $\mathbf{E}$ over $\mathbf{F}$, and there being a finite number of intermediate fields).

Let the extension be cyclic of order n and let τ be such that $\tau^n = 1$ but no lower power. Let the automorphism g generate the Galois group. Let $t = \theta + \tau g(\theta) + \cdots + \tau^{n-1} g^{n-1}(\theta)$. Then t has n distinct conjugates (assuming $\tau \in \mathbf{F}$) $g^i(\theta) + \tau g^{i+1}(\theta) + \cdots + \tau^{n-1} g^{n-1+i}(\theta)$ and so its minimum polynomial has degree n. Since $g(t) = \tau^{-1}(t)$, the element $t^n = a$ is invariant under the Galois group and lies in $\mathbf{F}$. So $\theta, g(\theta), \ldots, g^{n-1}(\theta)$ lie in the splitting field of $x^n = a$, which must be $\mathbf{E}$.

Geometric constructions provide an application of field theory. Suppose we are given a unit line segment. What figures can be constructed from it by ruler and compass? Let the segment be taken as a unit length or the x axis. Wherever we construct a new point from existing ones by ruler and compass it is an intersection of a line or circle with a line or circle. Such intersections lead to quadratic equations. Therefore, if a point P is constructible, each coordinate must be obtained from rational numbers by adding, subtracting, multiplying, dividing, or taking square roots. Such quantities lie in an extension field of $\mathbf{E} \subset \mathbf{Q}$ such that there exist fields $\mathbf{E_0} = \mathbf{Q} \subset \mathbf{E_1} \subset \cdots \subset \mathbf{E_k} = \mathbf{E}$ and $\mathbf{E_n} = \mathbf{E_{n-1}}(\sqrt{a})$ for $a \in \mathbf{E_{n-1}}$. The degree of $[\mathbf{E} : \mathbf{Q}] = [\mathbf{E_n} : \mathbf{E_{n-1}}] \cdots [\mathbf{E_1} : \mathbf{E_0}]$ is a power of 2.

Therefore, if x is a coordinate of a constructible point, x lies in an extension of degree 2^n, in fact a normal extension of degree 2^n. But if $[\mathbf{Q}(x) : \mathbf{Q}]$ has degree not a power of 2, this is impossible since $[\mathbf{E} : \mathbf{Q}] = [\mathbf{E} : \mathbf{Q}(x)][\mathbf{Q}(x) : \mathbf{Q}]$.

In particular, duplicating a cube (providing a cube of volume precisely 2) and trisecting an angle of 60° lead to roots of irreducible cubics $x^3 - 2 = 0$ and $4\cos^3\theta - 3\cos\theta - \cos 60^\circ = 0$ and cannot be performed. Since π does not satisfy any monic polynomial with coefficients in $\mathbf{Q}$, the circle cannot be squared.

D. Finite Fields

The fields $\mathbf{Z_p}$ where p is a prime number are fields with a finite number of elements. All finite fiels have $n1 = 0$ for some n and are therefore extension fields of some $\mathbf{Z_p}$. If $\mathbf{Z_p} \subset \mathbf{E}$, then $\mathbf{E}$ is a vector space over $\mathbf{Z_p}$. Let $x_1, x_2, \ldots, x_n$ be a basis. Then all elements of $\mathbf{E}$ can be uniquely expressed as linear combinations $\Sigma c_i x_i$, where $c_i \in \mathbf{Z_p}$. This set has cardinality $|\{(c_1, c_2, \ldots, c_n)\}| = |\mathbf{Z_p^n}| = p^n$. So $|\mathbf{E}| = p^n$ if $[\mathbf{E} : \mathbf{Z_p}] = n$.

Let $\mathbf{E}$ have order p^n. Then the multiplicative group $\mathbf{E}^*$ has order $p^n - 1$. So every element has order dividing $p^n - 1$. So if $r \neq 0$, $r^{p^n - 1} = 1$. So for all $r \in \mathbf{E}$, $r^{p^n} = r$. Then for all $r \in \mathbf{E}$, $x - r$ divides the polynomial $x^{p^n} - x$. Therefore $x^{p^n} - x$ factors as precisely the product of $x - r$ for all $r \in \mathbf{E}$. Therefore, if m is a divisor of $p^n - 1$, the equation $x^m - 1$ divides $x^{p^n - 1} - 1$ and splits into linear factors. As with $\mathbf{Z_p}$ this means we can get an element u_i of order a maximum power of each prime dividing $p^n - 1$ and their product will have order $p^n - 1$ and generate the group. So $\mathbf{E}^*$ is cyclic.

Since $p | p!$ but not $r!$ for $r < p$, p prime we have $p | \binom{p}{r}$. In $\mathbf{E}$ since $p1 = 0$, every element satisfies $x + x + \cdots + x = px = 0$. So $(x + y)^n = \Sigma \binom{n}{r} x^r y^{n-r}$ by the binomial theorem, which holds in all commutative rings. In $\mathbf{E}^*$, $(x+y)^p = x^p + y^p$ since other terms are divisible by p. This implies by induction, $(x + y)^{p^r} = x^{p^r} + y^{p^r}$. Therefore, x^{p^r} is an automorphism of $\mathbf{E}$.

This gives a cyclic group of order n of automorphisms of $\mathbf{E}$ since if y generates the cyclic group then $y^{p^i} \neq y$ for $i < n$. This is the complete automorphism group of $\mathbf{E}$.

If an element z lies in a proper subfield $\mathbf{F}$ of $\mathbf{E}$, then $\mathbf{F}$ has order p^k and $k | n$ and $z^{p^k} = z$. Conversely, the set of $\{z : z^{p^k} = z\}$ is closed under sums and products and multiplicative inverses so it is a proper subfield. So if $z^{p^k} = z$ then z lies in a proper subfield.

For any irreducible polynomial $p(x)$ of degree k over $\mathbf{Z_p}$, there exists a field $\mathbf{H} = \mathbf{Z_p}[x]/p(x)\mathbf{Z_p}[x]$. It has degree k and so order p^k. If $t = \bar{x}$, then it is $\mathbf{Z_p}(t)$ where $p(t) = 0$ is the minimum polynomial of t. Since $t^{p^k} - t = 0$, we have $p(x) | x^{p^k} - x$, and if $k | n$, then $p(x) | x^{p^k} - x | x^{p^n} - x$.

Suppose $p(x)|x^{p^n}-x$ and is irreducible of degree k, $k \not| n$. Then in $\mathbf{Z_p}(t)$ of degree k we would have $t^{p^n}=t$ and so for all r in the field $r^{p^n}=r$. This is false. So an irreducible polynomial divides $x^{p^n}-x$ if and only if its degree divides n.

The derivative can be defined for polynomials in $\mathbf{Z_p}[x]$ and satisfies the usual sum, product, power rules. If $x^{p^n}-x=q(x)(p(x))^2$, then $(d/dx)(x^{p^n}-x)=-1=q'(x)(p(x))^2+2q(x)p(x)p'(x)$. So $p(x)$ divides -1. For p of degree greater than zero, this is false. Therefore, $x^{p^n}-x$ has each irreducible polynomial of degree k dividing n exactly once as a factor.

Suppose $x^{p^n}-x$ has no irreducible factors of degree n. Then all its factors divide $x^{p^k}-x, k<n$. So $x^{p^n}-x|(x^{p^{n-1}}-x)\cdots(x^p-x)$. Since $x^{p^n}-x$ has higher degree this is false. So $x^{p^n}-x$ has an irreducible factor $g(x)$ of degree n. Therefore, a field $\mathbf{H}$ exists of order exactly p^n.

Any field $\mathbf{F}$ of order p^n has an element r such that $x-r|g(x)|x^{p^n}-x$ since $x^{p^n}-x$ factors into linear factors. Then r has minimum polynomial $g(x)$. It follows that $\mathbf{Z_p}(r)$ has degree n and equals $\mathbf{F}$. And $\mathbf{F}$ is isomorphic to $\mathbf{H}$. So any two fields of order p^n are isomorphic.

E. Codes

Coding theory is concerned with the following problem. Consider information in the form of sequences $a_1a_2\cdots a_m$ over a q-element set assumed to be a finite field $\mathbf{F}$ of order q. We wish to find a function f encoding $a_1a_2\cdots a_m$ as another sequence $b_1b_2\cdots b_n$ such that, if an error of specified type occurs in the sequence (b_i), the sequence (a_i) can still be recovered. There should also be a readily computable function g giving $a_1a_2\cdots a_m$ from $b_1b_2\cdots b_n$ with possible errors. We assume any t or fewer errors could occur in (b_i).

We consider $a_1a_2\cdots a_m$ as an m-dimensional vector $(a_1, a_2, \ldots, a_m)$ over $\mathbf{F}$, that is, $a\in\mathbf{V_m}$ and (b_i) as a vector $b\in\mathbf{V_n}$. Therefore, a coding function is a function $\mathrm{f}:\mathbf{V_m}\to\mathbf{V_n}$. The resulting code is its image of C where $C\subset\mathbf{V_n}$.

Suppose two vectors v, w in a code differ in at most $2t$ places. Then let z agree with v in half these, w the other half, and agree with v and w where they agree. Then z could have been either v or w. Error correction is impossible.

Conversely, if two vectors v, w could give rise to the same vector z by at most t errors each, then they differ in at most $2t$ places.

Therefore, a code can correct t errors if and only if every pair of vectors differ in at least $2t+1$ places.

For such a code, $q^n\geq q^m(\binom{n}{0}+(q-1)\binom{n}{1}+\cdots+(q-1)^t\binom{\mathrm{n}}{\mathrm{t}}$ since the kth term on the right is the number of vectors that differ from a code vector on exactly r places, and for $r\leq t$ all these are distinct.

A perfect code is one such that equality holds. A linear code is one that is a subspace of $\mathbf{V_n}$. A cyclic code is a linear code in which each cyclic rearrangement $b_1b_2\cdots b_n\to b_nb_1\cdots b_{n-1}$ of a code vector is a code vector.

Let $\mathbf{V_n}$ be considered as the set of polynomials of degree less than n in x. Then cyclic rearrangement is multiplication by x if we set $x^n=1$. Therefore, cyclic codes are subspaces closed by multiplication by x in

$$\frac{\mathbf{F}[x]}{(x^n-1)\mathbf{F}[x]}$$

and are therefore ideals.

Such an ideal is generated by a unique monic irreducible polynomial $g(x)$. Let γ be an element of an extension field of $\mathbf{F}$ such that $\gamma^n=1$ but no lower power is 1. Let $g(x)$ be the least common multiple of the minimum polynomials of $\gamma^1, \gamma^2, \ldots, \gamma^{d-1}$, where $d\leq n$, d is relatively prime to n. Then $g(x)|x^n-1$ since all powers of γ are roots of x^n-1. By a computation involving determinants no element of the ideal of $g(x)$ can have fewer than d nonzero coefficients. So $t=(d-1)/2$ errors can be corrected. The number m is n minus the degree of $g(x)$. These are called the BCH (Bose–Chaudhuri–Hoquenghem) codes and if $n=q-1$ Reed–Solomon codes.

F. Applications to Latin Squares

An $n\times n$ Latin square is a matrix whose entries are elements of an n-element set such that each number occurs exactly once in every row and column. The following cyclic Latin square occurs for all n:

$$\begin{bmatrix} 1 & 2 & \cdots & n \\ 2 & 3 & \cdots & 1 \\ & \cdots\cdots & & \\ n & 1 & \cdots & n-1 \end{bmatrix}$$

Two Latin squares (a_{ij}), (b_{ij}) are orthogonal if for all i, j, the ordered pairs (a_{ij}, b_{ij}) are distinct.

Orthogonal Latin squares are used in experiments in which several factors are tested simultaneously. In particular, suppose we want to test 5 fertilizers, 5 soil types, 5 amounts of heat, and 5 plant varieties. Choose 5×5 orthogonal Latin squares. Take an experiment with variety i, heat amount j, soil type a_{ij}, fertilizer b_{ij} for all i, j, and n^2 experiments. Then for any two factors, each variation on one factor occurs in combination with each for the other factor exactly once. If we have k mutually orthogonal Latin squares we can test $k+2$ factors.

Suppose $n=p_1^{n_1}p_2^{n_2}\cdots p_r^{n_r}$, where p_i is a prime number. Let R be the direct product of fields of order $p_i^{n_i}$. Let $k=\inf\{p_i^{n_i}-1\}$. Choose k invertible elements from each field, $x_{ij}: i=1, 2, \ldots, k;\ j=1, 2, \ldots, r$. Then the

elements $z_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ are invertible and the difference of any two is invertible.

Define r, $n \times n$ Latin squares $M\langle s\rangle$ by $m\langle s\rangle_{ij} = z_s y_i + y_j$, where y_i, y_j run through all elements of R. Then this function is 1–1 in y_j. Since z_r is invertible, it is 1–1 in y_i. From the fact that $z_s - z_t$ is invertible, it follows that $(m\langle s\rangle_{ij}, m\langle t\rangle_{ij}) \neq (m\langle s\rangle_{hk}, m\langle t\rangle_{hk})$ unless $i = h$ and $j = k$. So all the square are orthogonal. If n is a prime power, then $n-1$ is the maximum number of orthogonal $n \times n$ Latin squares. However, there exists a pair of orthogonal $n \times n$ Latin squares for all $n \neq 2, 6$ constructed by R. C. Bose, S. S. Shrikhande, and E. T. Parker in 1958. The existence of 10×10 Latin squares disproved a long-standing conjecture of Euler.

A $k \times m$ orthogonal array is a $k \times m$ matrix (a_{ij}) such that for any $i \neq j$ the ordered pairs a_{is}, a_{js} are distinct for $s = 1$ to m.

An $(m+2) \times n^2$ orthogonal array with entries from an n-element set yields m mutually orthogonal Latin squares. The first two rows run through all pairs (i, j) so let $m\langle r\rangle_{ij}$ be the entry in row $r+2$ such that rows 1, 2 have entries (i, j) in that place.

For n a prime power congruent to 3 modulo 4, an orthogonal $4 \times n^2$ array can be constructed. Let $\mathbf{F}$ be a field of order n and let g generate $\mathbf{F}^*$ its multiplicative group. Because of the congruence condition -1 will not be a square in $\mathbf{F}$. Let $y_1, y_2, \ldots, y_k, k = \frac{1}{2}(n-1)$, be distinct symbols not in $\mathbf{F}$. Take as columns of the array all columns of the types shown in Table IV, where x ranges through $\mathbf{F}$, and i independently varies from $1, 2, \ldots, \frac{1}{2}(n-1)$, together with n columns

$$\begin{bmatrix} x \\ x \\ x \\ x \end{bmatrix}, \qquad x \in \mathbf{F}$$

and $(n-1)^2/4$ columns corresponding to a pair of orthogonal Latin squares of size $(n-1)/2 \times (n-1)/2$ with entries $y_1, y_2, \ldots, y_k$. This gives an orthogonal array yielding, for example, two 10×10 orthogonal Latin squares.

G. Applications to Projective Planes

A projective plane is a mathematical system with a binary relation called incidence (lying on) from a set P called the set of points to a set L called the set of lines, satisfying three axioms: (1) Exactly one line is determined by two distinct points; (2) two lines intersect in exactly one point; and (3) there exist four points no three of which are collinear. The last axiom is only to guarantee that the system does not reduce to various special cases.

TABLE IV Columns Used to Construct Two Orthogonal Latin Squares

y_i	$g^{2i}(g+1)+x$	$g^{2i}+x$	x
x	y_i	$g^{2i}(g+1)+x$	$g^{2i}+x$
$g^{2i}+x$	x	y_i	$g^{2i}(g+1)+x$
$g^{2i}(g+1)+x$	$g^{2i}+x$	x	y_i

Let $\mathfrak{D}$ be any division ring. Left modules over $\mathfrak{D}$ have many of the properties of vector spaces. Let $\mathfrak{S} = \mathfrak{D} \times \mathfrak{D} \times \mathfrak{D}$ as a left module. Let $\mathfrak{P}$ be the set of submodules of the form $\mathfrak{D}x$, $x \in \mathfrak{S}$. Let $\mathfrak{L}$ be the set of submodules of the form $\mathfrak{D}x + \mathfrak{D}y$, $x, y \in \mathfrak{S}$, where $x \neq dy$ for any $d \in \mathfrak{D}$. Then $\mathfrak{P}$ and $\mathfrak{L}$ are essentially one- and two-dimensional subspaces of $\mathfrak{S}$. Let incidence be the relation $U \subset V$. Then we have a projective plane. A choice of $\mathfrak{P} = \mathbf{R}$ gives the standard geometric projective plane. This type of projective plane is characterized by the validity of the theorem of Desargues.

Many more general systems also give rise to projective planes. A ternary ring (or groupoid) is a set $\mathbf{R}$ with an operation $\mathbf{R} \times \mathbf{R} \times \mathbf{R}$ to $\mathbf{R}$ denoted $x \circ m * b$. Every projective plane can be realized with $\mathfrak{P}$ the set $\mathbf{R} \times \mathbf{R} \cup \mathbf{R} \cup \{\infty\}$, where $\mathbf{R} \times \mathbf{R}$ is considered as a plane for the ternary ring, $\mathbf{R} \cup \{\infty\}$ the points on a line at infinity. Incidence is interpreted set theoretically by membership, and lines consist of sets $y = x \circ m * b$ together with a line at infinity. Necessary and sufficient conditions that the ternary ring give a projective plane are that for some $0, 1 \in \mathbf{R}$, for all $x, m, b, y, k \in \mathbf{R}$: (TR-1) $0 \circ m * b = x \circ 0 * b = b$; (TR-2) $1 \circ m * 0 = m \circ 1 * 0 = m$; (TR-3) there exists a unique z such that $a \circ m * z = b$; (TR-4) there exists a unique z such that $z \circ m * b = z \circ k * a$ if $k \neq m$; and (TR-5) there exist unique z, w such that $a \circ z * w = x$ and $b \circ z * w = y$ if $a \neq b$.

Finite fields give ternary rings $xm + b$ of all prime power orders. It is unknown whether ternary rings (TR-1) to (TR-5) exist not of prime power order. If $|\mathbf{R}| = m$, then $|\mathfrak{P}| = |\mathfrak{L}| = m^2 + m + 1$.

Projective planes are essentially a special case of block designs. A balanced incomplete block design of type (b, v, r, k, λ) consists of a family B_i, $i = 1$ to b of subsets of a set V having v elements such that (1) $|B_i| = k > \lambda$ for all i; (2) $|\{i : x \in B_i\}| = r$ for all $x \in V$; and (3) $|\{i : x \in B_i$ and $y \in B_i\}| = \lambda$ for all $x \in V$, $x \neq y$. These are also used in design of experiments where B_i is the ith set of experiments and V the set of varieties tested.

Let $A = (a_{ij})$ be the matrix such that $a_{ij} = 1$ if and only if the ith element of V occurs in B_i. Then $AA^{\mathrm{T}} = \lambda J + (r-\lambda)I$ and all column sums of A are k, where J is a matrix all of whose entries are 1. Moreover, these properties characterize balanced incomplete block designs. For $k = 3$, $\lambda = 1$, designs are called Steiner triple systems.

A permutation group G acting on a set S is called doubly transitive if for all $x \neq y, u \neq v$ in S there exists g in G with

$gx = u$ and $gy = v$. If $T \subset S$, then the family of subsets $\{B_i\} = \{g(T)\}$ forms a balanced incomplete block design.

VII. OTHER ALGEBRAIC STRUCTURES

A. Groupoids

As mentioned before, a groupoid is a set having one binary operation satisfying only closure. For instance, in a finite group, the operation $aba^{-1}b^{-1}$ called the commutator gives a groupoid. Congruences and homomorphism on groupoids are defined in the same way as for semigroups, and for every congruence there is a quotient groupoid. Sometimes as in topology, groupoids are used to construct quotient groupoids which are groups.

Groupoids are also used in combinatorics. For example, any $n \times n$ Latin square, its entries labeled $1, 2, \ldots, n$ is equivalent to a groupoid satisfying (q) any two of a, b, c determines the third uniquely in $ab = c$. A groupoid satisfying (q) is called a quasigroup. A loop is a quasigroup with two-sided identity. Figure 9 classifies some one-operation structures.

B. Semirings

A semiring is a system in which addition and multiplication are semigroups with distributivity on both sides. Any Boolean algebra and any subset of a ring closed under addition and multiplication as the nonnegative elements in **Z** or **R** are semirings. The set of matrices over a semiring with 0, 1 comprises a semiring with additive and multiplicative identity.

A homomorphism of semirings is a function f from $\mathfrak{R}$ to $\mathfrak{S}$ such that $f(x + y) = f(x) + f(y)$, and $f(xy) = f(x)f(y)$ for all $x, y \in \mathfrak{R}$. The equivalence relation $\{(x, y) : f(x) = f(y)\}$ is a congruence. This means that, if $x \sim y$, then $x + z \sim y + z$, $xz \sim yz$, and $zx \sim yz$ for all $z \in \mathfrak{R}$. Conversely, for every congruence there is a quotient semiring and a homomorphism to the quotient semiring $x \to \bar{x}$.

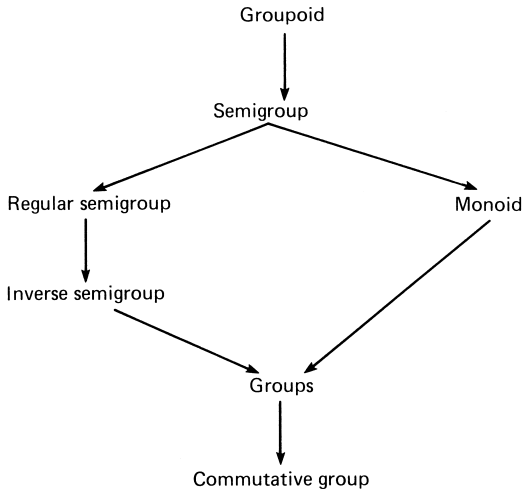


FIGURE 9 Classification of one-operation structures.

If the semiring is partially ordered, so are the semirings of n-square matrices over it.

Just as Boolean matrices represent binary relations, so matrices over semirings on [0, 1] can represent relations in which there is a concept of degree of relationship. One widely used semiring is the fuzzy algebra in which the operations are $\sup\{x, y\}$, $\inf\{x, y\}$, giving a distributive lattice. Fuzzy matrices have applications in clustering theory, where objects are partitioned into a hierarchy of subsets on the basis of a matrix $C = (c_{ij})$ giving the similarity between objects i and j. Applications of fuzzy matrices are also found in cybernetics.

Inclines are a more general class of semirings. An incline is a semiring satisfying $x + x = x$, $xy \leq x$, and $xy \leq y$ for all x, y. The set of two-sided ideals in any ring or semigroup forms an incline. The additive operation in an incline makes it a semilattice and, under weak restrictions such as compactness of intervals, a lattice. In a finitely generated incline for every sequence (a_i), $i \in \mathbf{Z}^+$, there exist $i < j$ with $a_i \geq a_j$. The set of n-square matrices over an incline is not an incline under matrix multiplication but is one under the elementwise product given by $(a_{ij}) \odot (b_{ij}) = (a_{ij}b_{ij})$. Inclines can be used to study optimization problems related to dynamic programming.

C. Nonassociative Algebras and Higher Order Algebras

A nonassociative ring is a system having two binary operations satisfying all the axioms for a ring except associativity of multiplication. A nonassociative algebra $\mathfrak{A}$ over a field **F** in addition is a vector space over the field satisfying the algebra property $a(bc) = b(ac) = (ab)c$ for all $a \in \mathbf{F}$, $b, c \in \mathfrak{A}$. There exists a nonassociative eight-dimensional division algebra over the real numbers called the Cayley numbers. It is an alternative ring, that is, $(yy)x = y(yx)$, and $(yx)x = y(xx)$ for all x, y. It has applications in topology and to projective planes.

A Lie algebra is an algebra in which for all a, b, c the product denoted $[a, b]$ satisfies (L-1) $[a, a] = 0$; (L-2) $[a, b] + [b, a] = 0$; and (L-3) $[[a, b], c] + [[b, c], a] + [[c, a], b] = 0$. In any associative algebra, the commutators $[a, b] = ab - ba$ define a Lie algebra. Conversely, for any Lie algebra an associative algebra called the universal enveloping algebra can be defined such that the Lie algebra is a subalgebra of its algebra of commutators. In many topological groups, for every element a there exists

a parametrized family $b(t), t \in \mathbf{R}$ of elements such that $b(1) = a$, $b(t+s) = b(t)b(s) = b(s)b(t)$. The operation $b(t)c(t)$ is commutative up to first order in t as $t \to 0$ and defines a real vector space. The operation $b(t)c(t)b^{-1}(t)c^{-1}(t)$ taken to second order in t then defines a Lie algebra. All finite dimensional connected, locally connected continuous topological groups can be classified using Lie algebras.

For any group G, let $\Gamma_i(G)$ denote the subgroup generated by all products of i elements under the operation of group commutator $xyx^{-1}y^{-1}$. Then $\Gamma_i(G)$ is a normal subgroup and $\Gamma_i(G)/\Gamma_{i+1}(G)$ is an Abelian group. Commutators give a product from $\Gamma_i(G)/\Gamma_{i+1}(G) \times \Gamma_j(G)/\Gamma_{j+1}(G)$ to $\Gamma_{i+j}(G)/\Gamma_{i+j+1}(G)$ satisfying (L-1) to (L-3). Such a structure is called a Lie ring. The sequence of normal subgroups $\Gamma_i(G)$ is called the lower central series.

A (noncommutative) Jordan algebra is a nonassociative algebra such that $(x^2y)x = x^2(yx)$, and $(xy)x = x(yx)$. Commutative Jordan algebras are used in constructing exceptional Lie algebras. For any associative algebra, the product $xy + yx$ defines a commutative Jordan algebra. They also are used in physics.

A median algebra is a set S with ternary product xyz such that (1) xyz is unchanged under any permutation of x, y, z; (2) $(x(xyz)w) = (xy(xzw))$; and (3) $xyy = y$ for all x, y, z, w. The product in any two factors is then a semilattice. The set of n-diemensional Boolean vectors is a median algebra under the componentwise product that xyz equals whichever of x, y, z is in the majority. The set of linear orders on n elements is a median algebra under the same operation.

D. Varieties

As mentioned earlier an algebraic operation on a set S is a function $S^n = S \times S \times \cdots \times S$ to S for some positive integer n. An operational structure is any labeled collection of operations on a set S. A substructure is a subset closed under all operations. A congruence is an equivalence relation such that for each operation if $x_i = y_i$ for $i \neq j$ and $x_j \sim y_j$, then $f(x_1, x_2, \ldots, x_n) \sim f(y_1, y_2, \ldots, y_n)$. Multiplication is then uniquely defined on equivalence classes, which form a quotient structure. A homomorphism is a function h such that for each operation f and all $x_1, x_2, \ldots, x_n$ in the domain $h[f(x_1, x_2, \ldots, x_n)] = f[h(x_1), h(x_2), \ldots, h(x_n)]$. Its image will be a substructure of the structure into which f is defined. An isomorphism is a 1–1 onto homomorphism. The direct product of algebraic structures with operations of the same type is the product set with componentwise operations.

A law for an algebraic structures S is a relation which holds identically for all elements of the structure. That is, it asserts that a certain composition of operations applied $x_1, x_2, \ldots, x_n$ equals a certain other composition for all $x_1, x_2, \ldots, x_n$ in the set. The commutative, associative, and distributive laws for commutative rings form an example. A variety is the class of all structures satisfying given laws. Any law valid for a class of structures will also hold for their substructures, homomorphic images, and direct products of them. Conversely, suppose a class $\mathfrak{C}$ of structures is closed under taking of substructures, product structures, and homomorphic images. For any set of generators $\mathfrak{g}$ we can construct an element of $\mathfrak{C}$ whose only relations are those implied by the general laws of $\mathfrak{C}$.

To do this we take one copy of each isomorphism class S_α of structues in $\mathfrak{C}$ that are either finite or have the same cardinality at most $|\mathfrak{g}|$. Replicate these copies until we have one copy $S_{\alpha\beta}$ for every map $f_{\alpha\beta} : \mathfrak{g} \to S_\alpha$. Take the direct product $\times S_{\alpha\beta}$ and the substructure $\mathbf{F}(\mathfrak{g})$ generated by the image of $\mathfrak{g}$ under each. Then by construction any relation among these generators must hold identically in all the S_α and thus be a law. It follows that the relations of $\mathfrak{g}$ conicide with laws of $\mathfrak{C}$.

If S is any structure in which the laws hold with $|S| = \mathfrak{g}$, then the the relations of $\mathbf{F}(\mathfrak{g})$, being laws of $\mathfrak{C}$, hold is S. In particular, there will then be an onto homomorphism from $\mathbf{F}(\mathfrak{g})$ to S. This proves that a class is a variety if and only if it is closed under taking substructures, product structures, and homomorphic images. Much work has been done on classifying varieties, especially groups.

Quasivarieties of relational structures on a set S also exist but must be defined a little differently, as classes closed under isomorphism, structures induced on subsets, and product structures. Laws are of three types. For a collection of variables $x_1, x_2, \ldots, x_n$ take a collection T_α of m-tuples from $x_1, x_2, \ldots, x_n$ for each m-ary relation R. For all replacements of x_i by elements of the set T either (1) if for all α, all members of T_α belong in R, then $(x_{i(1)}, x_{i(2)}, \ldots, x_{i(k)}) \in R$, where $i(1), i(2), \ldots, i(n) \in \mathbf{Z}^+$; (2) if for all α, all members of T_α belong in R, then $(x_{i(1)}, x_{i(2)}, \ldots, x_{i(k)}) \notin R$; or (3) if for all α, all members of T_α belong in R, then $x_{i(1)} = x_{i(2)}$ for some $i(1), i(2) \in \mathbf{Z}^+$. The classes of reflexive, irreflexive, symmetric, and transitive binary relations are all quasivarieties.

E. Categories and Topoi

If we take algebraic structures of a given type and ignore the internal structure but consider only the homomorphisms that exist between them, then we have assentially a category. To be precise, a category consists of (1) a class $\mathfrak{C}$ called the class of objects together with (2) a set $\mathrm{Hom}(x, y)$ for all $x, y \in \mathfrak{C}$ called the set of morphisms from x to y; (3) a special morphism 1_x for each x in $\mathfrak{C}$ called the identity morphism; and (4) an operation from $\mathrm{Hom}(x, y) \times \mathrm{Hom}(y, z)$ to $\mathrm{Hom}(x, z)$ called composition

of morphisms. It is required that composition be associative when defined and that 1_x act as the identity element whenever composition with it is defined. The most fundamental category consists of all sets and all functions, between sets. Sets and binary relations form a category. There are categories in which $\mathrm{Hom}(x, y)$ does not consist of functions, for example, let C be a poset and let $\mathrm{Hom}(x, y) = (x, y)$ if $x \leq y$ and $\mathrm{Hom}(x, y) = \emptyset$ otherwise.

The category of all modules over a ring is an additive category: Morphisms can be added in a natural way.

The direct product of objects $K_1 \times K_2$ can be described in terms of categories by its universal property. This is that there exist mappings $\pi_i : K_1 \times K_2$ to K_i such that for any object A and any mappings $f_i : A$ to K_i there exists a unique mapping $h : A \to K_1 \times K_2$ such that $\pi_i(h(x)) = f_i(x)$. There exists a dual concept of direct sum, an object $K_1 \times K_2$ with mapping $g_j : K_j \to K_1 \times K_2$ such that for any object A and any mappings $f_j : K_j$ to A there exists a unique mapping $h : K_1 \times K_2 \to A$ such that $h(g_j(x)) = f_j(x)$.

More generally, in category theory, one works with finite diagrams, or graphs, in which each vertex is an object and each arrow a morphism. The limit of a diagram cosists of an object A and a mapping g_j from each object a_j of the diagram to A such that if there is an arrow $f : a_k \to a_j$, then $g_j(f(x)) = g_k(x)$ and for any other object H and mappings m_j with the same properties there exists a unqiue mapping $h : A \to H$ such that $h(g_j(x)) = m_j(x)$ for all j. A colimit is the dual concept. Any variety of relational structures, as the variety of commutative rings, has products, coproducts, limits, and colimits. Moreover, there is the trivial structure $\{0\}$ where any set of operations on zero gives zero. Any object has a unique mapping to zero.

A topos is a category with limits, colimits, products, coproducts, and a few additional properties: There exists an exponential object X^Y for any objects X, Y such that for any object X there is a natural isomorphism from $\mathrm{Hom}(Z, X^Y)$ to $\mathrm{Hom}(Z \times Y, X)$. In addition, there exists a subobject classifier, that is, an object Ω and a mapping called true from a point p to Ω such that there is a 1–1 correspondence between substructures of any object X and mappings $X \to \Omega$. The category of sets acted on the left by a fixed monoid M is a topos. In the category of sets itself $\Omega = \{0, 1\}$, where subsets T of a set S are in 1–1 correspondence with mappings $f : S \to \{0, 1\}$ such that $f(x) = 1$ if x is in T and $x = 0$ if $x \notin T$. Topoi have important applications to models in mathematical logic such as in Boolean-valued models used to show the independence of the continuum hypothesis in Zermelo–Frankel set theory.

A functor from one category C_1 to another category C_2 consists of an assignment of a unique object of C_2 to each object of C_1, a unique morphism of C_2 to each morphism of C_1 such that the objects involved correspond, the mappings 1_x go to identities, and composition is preserved. Group rings over a fixed ring give a functor from the category of groups to the category of rings.

F. Algebraic Topology

Algebraic topology is the study of functors from subcategories (subsets of the sets and morphisms of a category forming a category under the same operations) of the category of topological spaces and continuous mappings to categories of algebraic structures. These functors often allow one to deduce the nonexistence or existence of possible topological spaces or continuous mappings. If for some functor F and two topological spaces X, Y, $F(X)$ is not isomorphic to $F(Y)$, then X, Y cannot be topologically equivalent. In topology, any mapping between spaces is meant to be a continuous mapping.

Let $A(B)$ be a subset of the topological space $X(Y)$. A mapping from X, A to Y, B is a mapping $f : X \to Y$ such that $f(A) \subset B$. Two mapings f and $g : X, A \to Y$, B are called homotopic if there exists a mapping $h : \mathfrak{U} \times X, \mathfrak{U} \times A \to Y, B$ where $\mathfrak{U} = [0, 1]$ such that $h(0, x) = f(x)$, and $h(1, x) = g(x)$. This is an equivalence relation. Equivalence classes of mappings under homotopy form a new category called the homotopy category.

The first important functor from a category associated to the category of topological spaces to the category of groups is the fundamental group. Let x_0 denote any point of X. Homotopy classes of mappings f, g from $\mathfrak{U}, \mathfrak{V}$ (where $\mathfrak{V} = \{0, 1\}$) to X, x_0 can be multiplied by defining $f * g = h$ such that $h(t) = g(2t)$ for $0 \leq t \leq 0.5$ and $h(t) = g(2t - 1)$ for $0.5 \leq t \leq 1$. The product depends only on homotopy classes, and on homotopy classes is a group, where $f(1 - t)$ is the inverse of $f(t)$.

Every group is the fundamental group of some topological space, and every group homomorphism can be realized by a continuous mapping of some topological spaces.

Higher homotopy groups are defined in a similar way using an n-cube and its boundary in place of $\mathfrak{U}, \mathfrak{V}$.

Suppose a space Y is a topological groupoid; that is, there exists a continuous mapping $Y \times Y \to Y$. Then the set $[X : Y]$ of homotopy classes of mappings from X to Y is a groupoid for any topological space Y. If we let Y be a topological space whose only nonzero homotopy group is an Abelian group G in dimension n, called an Eilenberg–MacLane space, then $[X; Y]$ is a group called the nth cohomology group of X. All the cohomology groups together form a ring. From the cohomology groups of Eilenberg–MacLane spaces themselves can be obtained cohomology operations, that is, mappings from one cohomology group to another preserved by continous mappings.

A vector bundle over a topological space X is a vector space associated to each point of X such that the union of the vector spaces forms a topological space. An example is the tangent vectors to a surface. The K-theory of a topological space is essentially a quotient of the semigroup of vector bundles over X under the operation of direct sum of vector spaces. The equivalence class of a vector bundle V is the set of vector bundles W such that $U \oplus V$ is equivalent to $U \oplus W$ for some vector bundle U. K-Theory of X also equals $[X; Y]$ for a space Y which can be constructed from quotient groups of groups of n-square matrices.

By analogy-with topological constructions, researchers have defined homology and K-theory for rings and modules.

G. Inclines and Antiinclines

We have defined incline above as a semiring satisfying the incline inequality. The set of two-sided ideals in a ring (semigroup) for example forms an incline under $I + J(I \cap J)$ and IJ. Many results on Boolean and fuzzy matrices generalize to inclines. Green's relation $\mathcal{L}$- ($\mathcal{R}$-) classes in the semiring of matrices over an incline are characterized by equality of row, column spaces. In many cases the row, column spaces will have a unique basis. For the fuzzy algebra [0, 1] under $\sup\{x, y\}$, $\inf\{x, y\}$, it is necessary to add the condition that if $c_i = \Sigma a_{ij} c_j$, $\{c_i\}$ the basis, then $a_{ii} c_i = c_i$. Matrices will have computable maximal subinverses $A \times A \leq A$, giving a way to test regularity.

In any finitely generated incline, no chain has an infinite nondecreasing subchain. Eigenvectors and eigenvalues can be described in terms of a power of a matrix. A particularly interesting incline is [0, 1] under $\sup\{x, y\}$ and xy (ordinary multiplication).

The asymptotic forms of finite matrices can be described by the existence of positive integers k, d, and a matrix C such that

$$A^{k+nd} = A \odot C \odot \cdots \odot C$$

where $\odot$ denotes entrywise product.

Antiinclines are defined by the reverse inequalities $xy \geq x$ and $yx \geq x$. They have the dual nonascending chain property.

H. Quadratic Forms

A *quadratic form* over a ring $\mathcal{R}$ is a function $f(x) = \Sigma a_{ij} x_i x_j$, $a_{ij} \in \mathcal{R}$ from $\mathcal{R}^n \times \mathcal{R}^n$ to $\mathcal{R}$. Quadratic forms occur often in statistics and optimization and as the first nonlinear term in a power series, as well as in pure mathematics.

Two quadratic forms are *isomorphic* if they become identical after a linear invertible change of variables. A form is *isotropic* if nonzero x gives $f(x) = 0$. It is defined to be *totally isotropic* if $f(x) = 0$ identically, *hyperbolic* if it is isomorphic to a direct sum of forms $f(x) = x_1 x_2$ over $\mathcal{R}^2$, and *anisotropic* if it is not isotropic.

Over a field every quadratic form is uniquely expressible as the direct sum of an anisotropic form, a hyperbolic form, and a totally isotropic form up to equivalence. Cancellation holds for direct sum: if $h(x) \oplus f(x)$ is isomorphic to $h(x) \oplus g(x)$ then $f(x)$ is isomorphic to $g(x)$. It is then convenient to work with a ring of isomorphism classes of forms and their additive inverses called the *Witt-Grothendieck ring* $\hat{\mathcal{W}}$. The *Witt ring* $\mathcal{W}$is generated by $\langle a \rangle$, the class of ax_{ij}^2 has operations induced by direct sum and tensor product, and is defined by relations $\langle 1 \rangle = 1$ $\langle a \rangle \langle b \rangle = \langle ab \rangle$, $\langle a \rangle + \langle b \rangle = \langle a + b \rangle (1 + \langle ab \rangle)$. Here $\hat{\mathcal{W}}$ is defined as $\hat{\mathcal{W}}/H$ where H is the ideal of hyperbolic forms. It is studied in terms of the filtration by powers of the ideal I generated by $\langle a \rangle - 1$.

Over the real numbers, rank [of the matrix (a_{ij})] and signature (the number of its positive eigenvalues, given $a_{ij} = a_{ji}$) are complete isomorphism invariants. Over the rational numbers the *strong Hasse principle* asserts that two forms are isomorphic if and only if they are isomorphic over each completion (real and p-adic) of the rationals. The *weak Hasse principle* asserts that a form is isotropic if and only if it is isotropic over each completion. A *complete set of invariants of nonsingular forms* is given by determinant $\det[(a_{ij})]$ and Hasse invariants at each prime. The Hasse invariant of a *quadratic form* $\Sigma a_{ii} x_i^2$ can be defined as

$$\prod_{i \leq j} (a_{ii}, a_{ij})$$

where $(a_{ii}, a_{jj}) = \pm 1$ according as $a_{ii} x^2 + a_{jj} y^2 = 1$ has a solution or not, over the p-adic numbers. The determinant, taken modulo squares is called the *discriminant*.

Milnor's theorem gives a description of the Witt ring of $\mathcal{K}(x)$, x transcendental in terms of the Witt ring of $\mathcal{K}$. The Tsen–Lang theorem asserts that if $\mathcal{K}$ has transcendance degree n over $\mathbf{C}$ then every quadratic form of dimension greater than 2^n is isotropic. The theory of quadratic forms is also related to that of division algebra.

I. Current and Recent Research

The greatest achievement in algebra itself since 1950 has been the classification of finite simple groups. Proofs are very lengthy and due to many researchers. Much progress has been made on decidability of algebraic problems, for example, the result that the word problem is unsolvable in general groups having a finite number of generators and

the solution of Hilbert's problem showing that polynomial equations in n variables over $\mathbf{Z}$ (Diophantine equation) are in general undecidable by Turning machines. The proof of Mordell's conjecture gives a positive step for equations of degree n in two variables over $\mathbf{Q}$. Another result in group theory was the construction of an infinite but finitely generated group G such that $x^m = e$ for all x in G and for a fixed m in $\mathbf{Z}^+$.

In algebraic geometry, a remarkable theory has been created leading to the proof of the Weil conjectures. This theory made it possible to prove that, for any Diophantine equation, it is decidable whether for every prime number p it has a solution modulo p. Much has been done with algebraic groups.

In coding theory, all perfect codes have been found. A pair of $n \times n$ orthogonal Latin squares have been constructed for all n except 2 and 6.

The entire subjects of homological algebra and algebraic K-theory have been developed. For a ring $\mathfrak{R}$, the following sequence

$$M_1 \xrightarrow{f_1} M_2 \xrightarrow{f_2} \cdots \xrightarrow{f_n} M_n$$

is called exact if $\mathrm{Im}(f_i) = \mathrm{Ker}(f_{i+1})$, where $\mathrm{Ker}(f)$ denotes the kernel of f. A free resolution of a module $\mathfrak{M}$ is an exact sequence $\cdots \to \mathfrak{C}_n \to \mathfrak{C}_{n-1} \to \cdots \mathfrak{C}_0 \to \mathfrak{M}$, where each $\mathfrak{C}_i$ is a free module. Another module $\mathfrak{N}$, gives a sequence

$$\cdots \leftarrow \mathrm{Hom}(\mathfrak{C}_n, \mathfrak{N}) \xleftarrow{g_n} \mathrm{Hom}(\mathfrak{C}_{n-1}, \mathfrak{N}) \xleftarrow{g_{n-1}} \cdots \xleftarrow{g_0} \mathrm{Hom}(\mathfrak{C}_0, \mathfrak{N})$$

The quotients $\mathrm{Ker}(g_{n+1})/\mathrm{Im}(g_n)$ are independent of the particular free resolution and are called $\mathrm{Ext}^n(\mathfrak{M}, \mathfrak{N})$. Whitehead's problem is whether if $\mathrm{Ext}^1(A, \mathbf{Z}) = 0$, then A is a direct sum of copies of $\mathbf{Z}$. S. Shelah proved this is independent of the axioms of Zermelo–Frankel set theory.

Considerable research has been done on order structures and on ordered algebraic structures in lattice theory and general algebra. Finite posets can be studied in ways similar to topological spaces, because they are equivalent to finite topological spaces. The theory of Boolean and fuzzy matrices has been developed with the advent of Green's relations classes. Inclines, semirings $(\mathfrak{R}, \circ, *)$ satisfying $x \circ x = x$, $x \circ (x * y) \circ (y * x) = x$ are a further generalization. The algebraic structure of semigroups, especially regular semigroups, has become well understood. In matrix theory and algebraic number theory, there have been numerous important developments such as the proof of the van der Waerden conjecture.

Mathematical linguistics and automata theory have become well-developed subjects.

Research is actively continuing in most of these areas as well as in category theory and the theory of varieties and combinatorial aspects of algebra.

In the 1990s quantum algebras has become a very active field. This deals with structures related to traditional algebraic structures in the way quantum physics is related to classical physics. In particular, a quantum group is a kind of Hopf algebra.

A somewhat related and active area is noncommutative algebraic geometry, in which a prominent place is occupied by the K-theoretic ideas of A. Connes.

SEE ALSO THE FOLLOWING ARTICLES

• ALGEBRAIC GEOMETRY • BOOLEAN ALGEBRA • GROUP THEORY • MATHEMATICAL LOGIC • SET THEORY • TOPOLOGY, GENERAL

BIBLIOGRAPHY

Cao, Z. Q., Kim, K. H., and Roush, F. W. (1984). "Incline Algebra and Application," Ellis Horwood, Chichester, England/Wiley, New York.
Childs, L. (1979). "A Concrete Introduction to Higher Algebra," Springer-Verlag, Berlin and New York.
Connes, Ah. (1994). "Noncommutative Geometry," Academic Press, New York.
Fraleigh, J. B. (1982). "A First Course in Abstract Algebra," Addison-Wesley, Reading, Massachusetts.
Gilbert, J., and Gilbert, L. (1984). "Applied Modern Algebra," Prindle, Weber & Schmidt, Boston, Massachusetts.
Kassel, C. (1995). "Quantum Groups," Springer-Verlag, Berlin and New.
Kim, K. H. (1982). "Boolean Matrix Theory and Applications," Marcel Dekker, New York.
Kim, K. H., and Roush, F. W. (1984). "Incline Algebra and Applications," Ellis Horwood, Chichester, England/Wiley, New York.
Lam, T. Y. (1973). "Algebraic Theory of Quadratic Forms," Benjamin, Reading, Massachusetts.
Laufer, H. B. (1984). "Applied Modern Algebra," Prindle, Weber & Schmidt, Boston, Massachusetts.
Lidl, R., and Pilz, G. (1984). "Applied Abstract Algebra," Springer-Verlag, Berlin and New York.
Pinter, C. C. (1982). "A Book of Abstract Algebra," McGraw-Hill, New York.



Algebraic Geometry

Rick Miranda
Colorado State University

I. Basic Definitions of Affine Geometry
II. Projective Geometry
III. Curves and Surfaces
IV. Applications

GLOSSARY

Affine space The space of n-tuples with coordinates in a field K.
Algebraic set The set of solutions to a collection of polynomial equations in affine or projective space.
Coordinate ring Ring of polynomial functions on an algebraic set.
Curve An algebraic variety of dimension one.
Genus The fundamental invariant of curves, equal to the dimension of the space of regular 1-forms.
Moduli space A topological space (usually an algebraic variety itself) whose points represent algebraic varieties of a specific type.
Projective space The space of $(n+1)$-tuples with coordinates in a field K, not all zero, up to scalar factor. The natural compactification of affine space, which includes points at infinity.
Quadric An algebraic variety defined by a single quadratic equation.
Rational function A ratio of polynomial functions defined on an algebraic set in affine or projective space.
Regular function A rational function whose denominator does not vanish at a subset of an algebraic set.

ALGEBRAIC GEOMETRY is, broadly speaking, the study of the solutions of polynomial equations. Since polynomials are ubiquitous in mathematics, algebraic geometry has played a central role in the development and study of many mathematical ideas. Its origins can be traced to ancient Greek mathematics, when the study of lines and conics in the plane began. From these relatively simple beginnings, modern algebraic geometry has grown into a subject which draws on and informs nearly every other discipline in mathematics, and modern practitioners develop and use some of the most sophisticated "mathematical machinery" (including sheaves, cohomology, etc.) available. However, the subject is still rooted in fundamental questions about systems of polynomial equations; the complexity and apparent abstraction of some of the modern tools of practicing algebraic geometers should not be confused with a lack of concern for basic and accessible problems.

I. BASIC DEFINITIONS OF AFFINE GEOMETRY

A. Affine Space

To be more precise about solutions to polynomial equations, one first specifies a field k of numbers where the

coefficients of the polynomials lie, and where one looks for solutions. A field is a set with two binary operations (addition and multiplication) in which all of the usual rules of arithmetic hold, including subtraction and division (by nonzero numbers). The main examples of interest are the field $\mathbb{Q}$ of rational numbers, the field $\mathbb{R}$ of real numbers, and the field $\mathbb{C}$ of complex numbers. Another is the finite field $\mathbb{Z}/p$ of the integers $\{0, 1, \ldots, p-1\}$ under addition and multiplication modulo a prime p.

Solutions to polynomials in n variables would naturally be n-tuples of elements of the field. If we denote the field in question by K, the natural place to find solutions is affine n-space over K, denoted by K^n, $\mathbb{A}^n_K$, or simply $\mathbb{A}^n$:

$$\mathbb{A}^n = \{\underline{\mathbf{z}} = (z_1, z_2, \ldots, z_n) \mid z_i \in K\}.$$

When $n = 1$ we have the affine line, if $n = 2$ we have the affine plane, and so forth.

B. Affine Algebraic Sets

Let $f(\underline{\mathbf{x}}) = f(x_1, \ldots, x_n)$ be a polynomial in n variables with coefficients in K. The zeros of f are denoted by $Z(f)$:

$$Z(f) = \{\underline{\mathbf{z}} \in \mathbb{A}^n \mid f(\underline{\mathbf{z}}) = 0\}.$$

For example, $Z(y - x^2)$ is a parabola in the plane, and $Z(x^2 - y^2 - z^2 - 4)$ is a hyperboloid in 3-space.

More complicated geometric objects in affine space must be defined by more than one polynomial. Let $\mathcal{S}$ be a set of polynomials (not necessarily finite) all having coefficients in the field K. The common zeros of all the polynomials in the set $\mathcal{S}$ is denoted by $Z(\mathcal{S})$:

$$Z(\mathcal{S}) = \{\underline{\mathbf{z}} \in \mathbb{A}^n \mid f(\underline{\mathbf{z}}) = 0 \quad \text{for all } f \in \mathcal{S}\}.$$

An algebraic set in $\mathbb{A}^n$, or simply an affine algebraic set, is a subset of $\mathbb{A}^n$ of the form $Z(\mathcal{S})$ for some set of polynomials $\mathcal{S}$ in n variables. That is, an affine algebraic set is a set which is exactly the set of common zeros of a collection of polynomials. Affine algebraic sets are the fundamental objects of study in affine algebraic geometry.

The empty set is an affine algebraic set [it is $Z(1)$] and so is all of affine space [it is $Z(0)$]. The intersection of an arbitrary collection of algebraic sets is an algebraic set. The union of a finite number of algebraic sets is also an algebraic set. Therefore the algebraic sets in affine space form the closed sets in a topology; this topology is called the Zariski topology.

C. Ideals Defining Algebraic Sets

The "algebraic" part of algebraic geometry involves the use of the tools of modern algebra to study algebraic sets. Algebra is the study of sets with operations on them, and the most common algebraic structure is the ring, which is a set with addition and multiplication, but not necessarily division. The ring of importance in affine algebraic geometry is the ring $K[\underline{\mathbf{x}}] = K[x_1, x_2, \ldots, x_n]$ of polynomials in the n variables $\underline{\mathbf{x}} = (x_1, \ldots, x_n)$. Sometimes this is called the affine coordinate ring, since it is generated by the coordinate functions x_i. It is from this ring that we draw the subset $\mathcal{S} \subset K[\underline{\mathbf{x}}]$ of polynomials whose zeros we want to study.

An ideal J in a ring R (like $K[\underline{\mathbf{x}}]$) is a subset of the ring R with the special properties that it is closed under addition and "outside multiplication": if f and g are in J then $f + g$ is in J, and if g is in J then fg is in J for any f in the ring. An example of an ideal is the collection of all polynomials that vanish at a particular point $p \in \mathbb{A}^n$.

If $\mathcal{S}$ is a subset of the ring, the ideal generated by $\mathcal{S}$ is the set of all finite "linear combinations" $\sum_i h_i f_i$ where the h_i's are in the ring and the f_i's are in the given subset $\mathcal{S}$. The reader may check that this set, denoted by $\langle \mathcal{S} \rangle$, is closed under addition and outside multiplication, and so is always an ideal.

Returning to the setting of algebraic sets, one can easily see that if $\mathcal{S}$ is any collection of polynomials in $K[\underline{\mathbf{x}}]$, and $J = \langle \mathcal{S} \rangle$ is the ideal in $K[\underline{\mathbf{x}}]$ generated by $\mathcal{S}$, then $\mathcal{S} \subset J$ and $Z(\mathcal{S}) = Z(J)$: the set $\mathcal{S}$ and the ideal J have exactly the same set of common zeros. This allows algebraic geometers to focus only on the zeros of ideals of polynomials: every algebraic set is of the form $Z(J)$ for an ideal $J \subset K[\underline{\mathbf{x}}]$. It is not the case that the ideal defining the algebraic set is unique, however: it is possible that two different ideals J_1 and J_2 have the same set of common zeros, so that $Z(J_1) = Z(J_2)$.

D. Algebraic Sets Defining Ideals

We saw in the last paragraph that ideals may be used to define all algebraic sets, but that the defining ideal is not unique. In the construction of algebraic sets, we look at a collection of polynomials and take their common zeros, which is a set of points in $\mathbb{A}^n$. Now we turn this around, and look instead at a subset $X \subset \mathbb{A}^n_K$ of points in affine space, and consider all the polynomials that vanish on X. We denote this by $I(X)$:

$$I(X) = \{f \in K[\underline{\mathbf{x}}] \mid f(\underline{\mathbf{z}}) = 0 \quad \text{for all } \underline{\mathbf{z}} \in X\}.$$

No matter what kind of subset X we start with, this is always an ideal of polynomials. Again there is the same possibility of non-uniqueness: it may be that two different subsets X and Y of $\mathbb{A}^n$ have the same ideal, so that $I(X) = I(Y)$. This would happen if every polynomial that vanished on X also vanished on Y and vice versa.

The ideal $I(X)$ for a subset $X \subset \mathbb{A}^n$ has a special property not shared by all ideals: it is closed under roots. That is, if f is a polynomial, and f^m vanishes on all the points of X, then it must be the case that f itself vanishes. This means that if $f^m \in I(X)$, then $f \in I(X)$. An ideal with this special property is called a radical ideal.

E. The *Z–I* Correspondence

The two operations of taking an ideal of polynomials and using Z to get a subset of affine space, and taking a subset of affine space and using I to get an ideal of polynomials form the theoretical foundation of affine algebraic geometry. We have the following basic facts:

- If $J_1 \subset J_2 \subset K[\underline{\mathbf{x}}]$ are ideals, then $Z(J_1) \supset Z(J_2)$.
- If $X_1 \subset X_2 \subset \mathbb{A}^n$, then $I(X_1) \supset I(X_2)$.
- If $J \subset K[\underline{\mathbf{x}}]$ is an ideal, then $I(Z(J)) \supset J$.
- If $X \subset \mathbb{A}^n$, then $Z(I(X)) \supset X$.

This last statement can be sharpened to give a criterion for when a subset of $\mathbb{A}^n$ is algebraic, as follows. If X is algebraic, equal to $Z(J)$ for some ideal J, then $I(X) = I(Z(J)) \supset J$, so that $Z(I(X)) = Z(I(Z(J))) \subset Z(J) = X$, which forces $Z(I(X)) = X$. Conversely if $Z(I(X)) = X$, then X is obviously algebraic. Hence,

X is an algebraic subset of $\mathbb{A}^n$ if and only if $Z(I(X)) = X$.

This statement addresses the question of which points are additional zeros of all the polynomials which vanish on X, stating that if X is algebraic there are no additional such points.

F. The Nullstellensatz and the Basis Theorem

The Nullstellensatz turns the previous question around and asks which additional polynomials always vanish at the points which are common zeros of a given ideal of polynomials. This is a much more subtle question, and the answer is the fundamental theorem in elementary affine geometry. It uses the concept of the radical of an ideal J, $\mathrm{rad}(J)$, which is the set of all polynomials for which some power is in J:

$$\mathrm{rad}(J) = \{f \in K[\underline{\mathbf{x}}] \mid f^m \in J \quad \text{for some } m \geq 1\}.$$

It is always the case that $\mathrm{rad}(J) \supset J$, and that $\mathrm{rad}(J)$ is a radical ideal. Clearly $I(Z(J)) \supset \mathrm{rad}(J)$: if $f^m \in J$, then f will certainly vanish wherever everything in J (in particular wherever f^m) vanishes.

An algebraically closed field is a field that contains all roots of all polynomials in one variable. The field $\mathbb{C}$ of complex numbers is algebraically closed; the field $\mathbb{Q}$ of rational numbers, the field $\mathbb{R}$ of real numbers, and all finite fields are not algebraically closed.

Theorem 1 (Hilbert's Nullstellensatz). If K is an algebraically closed field, then for any ideal $J \subset K[\underline{\mathbf{x}}]$, $I(Z(J)) = \mathrm{rad}(J)$.

The power of the Nullstellensatz occurs because in many applications one has access to a polynomial f which is known to be zero when other polynomials $g_1, \ldots, g_k$ are. The conclusion is then that f is in the radical of the ideal generated by the g_i's. Therefore there is a power of f which is a linear combination of the g_i's, and so there is an explicit equation of the form

$$f^m = \sum_i h_i g_i.$$

The Nullstellensatz permits a more detailed correspondence of properties between the algebraic set X and its ideal $I(X)$. Typical statements when K is algebraically closed are as follows:

- There is a one-to-one correspondence between algebraic subsets of $\mathbb{A}^n$ and radical ideals in $K[\underline{\mathbf{x}}]$.
- X is empty if and only if $I(X) = K[\underline{\mathbf{x}}]$.
- $X = \mathbb{A}^n$ if and only if $I(X) = \{0\}$.
- X consists of a single point if and only if $I(X)$ is a maximal ideal.
- X is an irreducible algebraic set (that is, it is not a union of two proper algebraic subsets) if and only if $I(X)$ is a prime ideal [that is, $I(X)$ has the property that if $fg \in I(X)$ then either $f \in I(X)$ or $g \in I(X)$].

All these statements relate geometric properties of the algebraic set X to algebraic properties of the ideal $I(X)$. This is the basic theme of algebraic geometry in general.

The other main theorem concerning these general ideas is the basis theorem, which addresses the question of the finiteness of the set of equations which may define an algebraic set:

Theorem 2 (Hilbert's basis theorem). Every ideal J in the ring of polynomials $K[\underline{\mathbf{x}}]$ is finitely generated: there is a finite set $\mathcal{S}$ of polynomials such that $J = \langle \mathcal{S} \rangle$. As a consequence, every algebraic set $X \subset \mathbb{A}^n$ may be defined as the set of common zeros of a finite set of polynomials.

G. Regular and Rational Functions

Any polynomial $f \in K[\underline{\mathbf{x}}]$ can be considered as a K-valued function on affine space $\mathbb{A}^n$. If $X \subset \mathbb{A}^n$ is an algebraic set, such a polynomial may be restricted to X

to obtain a function on X. Functions on X which are restrictions of polynomial functions are called polynomial functions. The polynomial functions on X form a ring, denoted by $K[X]$ and called the affine coordinate ring of X. For example, the affine coordinate ring of affine space is the entire polynomial ring: $K[\mathbb{A}^n] = K[\underline{\mathbf{x}}]$.

There is an onto ring homomorphism (restriction) from $K[\underline{\mathbf{x}}]$ to $K[X]$, whose kernel is the ideal $I(X)$. Therefore $K[X]$ is isomorphic as a ring to the quotient ring $K[\underline{\mathbf{x}}]/I(X)$.

Irreducible algebraic sets are often referred to as algebraic varieties. If X is an algebraic variety, then $I(X)$ is a prime ideal, so that the coordinate ring $K[X]$ is an integral domain (if $fg = 0$ then either $f = 0$ or $g = 0$). The field of fractions of $K[X]$, denoted by $K(X)$, represent rational functions on the variety X. Rational functions are more useful than polynomial functions, but they have the drawback that any given rational function may not be defined on all of X (where the denominator vanishes). If a rational function is defined at a point p, one says that the function is regular at p.

The rational function field $K(X)$ of an algebraic variety X is an extension field of the base field K, and as such has a transcendence degree over K: this is the largest number of algebraically independent rational functions. This transcendence degree is the dimension of the variety X. Points have dimension zero, curves have dimension one, surfaces have dimension two, and so forth.

Polynomial and rational functions are used to define maps between algebraic sets. In particular, maps between two affine spaces are simply given by a vector of functions. Maps between affine algebraic sets are given by restriction of such a vector of functions. Depending on the type of functions used, such maps are called polynomial or rational maps. Again, rational maps have the property that they may not be defined everywhere. A map is called regular if it is given by rational functions but is defined everywhere.

H. Examples

The most common example of an affine algebraic variety is an affine subspace: this is an algebraic set given by linear equations. Such a set can always be defined by an $m \times n$ matrix A, and an m-vector $\underline{\mathbf{b}}$, as the vanishing of the set of m equations given in matrix form by $A\underline{\mathbf{x}} = \underline{\mathbf{b}}$. This gives the geometric viewpoint on linear algebra. One consequence of the geometric point of view is a deeper understanding of parallelism for linear spaces: geometrically, one begins to believe that parallel lines might be made to intersect "at infinity." This is the germ of the development of projective geometry.

No doubt the first nonlinear algebraic sets to be studied were the conics. These are the algebraic sets given by a single quadratic equation in two variables, and they define a curve in the affine plane. The classification and study of conics date to antiquity. Most familiar is the classification of nonempty irreducible conics over the real numbers $\mathbb{R}$: we have either an ellipse, a parabola, or a hyperbola. Again an appreciation of the points "at infinity" sheds light on this classification: an ellipse has no real points at infinity, a parabola has one real point at infinity, and a hyperbola has two real points at infinity.

Over any field, if an irreducible conic C [given by $f(x, y) = 0$] is nonempty, then it may be parametrized using a rational function of one variable. This is done by choosing a point $p = (x_0, y_0)$ on C, writing the line L_t through p with slope t [given by $y - y_0 = t(x - x_0)$], and intersecting L_t with C. This intersection will consist of the point p and one other point p_t which depends on t. One may solve easily for the coordinates of p_t as rational functions of t, giving a rational parametrization for the conic C. From the point of view of maps, this gives a rational map $\phi : \mathbb{A}^1 \to C$ which has an inverse: for any point (x, y) on C, the t-value is $t = (y - y_0)/(x - x_0)$. Invertible rational maps are called birational, and most of the classification efforts of algebraic geometry are classifications up to birational maps.

Applying this procedure to the unit circle C (defined by $x^2 + y^2 = 1$) and choosing the point p to be the point $(-1, 0)$, one finds the parametrization

$$p_t = \left(\frac{1 - t^2}{1 + t^2}, \frac{2t}{1 + t^2} \right).$$

We see here the origins of algebraic number theory, in particular the formulas for the Pythagorean triples. If we look at the case when $t = p/q$ is a rational number, we find (clearing the denominator q) that $((q^2 - p^2)/(q^2 + p^2), 2pq/(q^2 + p^2))$ is a point on the unit circle. Again clearing the denominator $q^2 + p^2$, we see that this means

$$(q^2 - p^2)^2 + (2pq)^2 = (q^2 + p^2)^2,$$

which gives the standard parametrization of Pythagorean triples. This style of argument in elementary number theory dates the ancient Greeks, in particular to Appollonius and Diophantus.

In higher dimensions, an algebraic variety given by a single quadratic equation is called a quadric. Although the classification of affine quadrics is not difficult, it becomes clearer that the use of projective techniques simplifies the matter considerably.

Algebraic varieties defined by either more equations or equations of degree larger than two present a much greater challenge. Even the study of cubic curves in the plane [given by a single equation $f(x, y) = 0$ where f

has degree three] is still a subject of modern research, especially over number fields.

I. Affine Schemes

The theory exposed above was developed in the hundred years ending in the middle of the twentieth century. In the second half of the twentieth century a different foundation to algebraic geometry was developed, which more closely follows the algebra of the rings and ideals in question. Recall that there is a correspondence between algebraic subsets of affine space and radical ideals in the polynomial ring $K[\underline{\mathbf{x}}]$. If the ground field K is algebraically closed, points correspond to maximal ideals, and irreducible algebraic sets to prime ideals. The ring of polynomial functions on X is naturally the ring $K[\underline{\mathbf{x}}]/I(X)$. This ring is a finitely generated K-algebra, with no nilpotent elements (elements $f \neq 0$ such that $f^k = 0$ for some k). From this ring one can recover X, as the set of maximal ideals.

The idea of affine schemes is to start with an arbitrary ring R and to construct a geometric object X having R as its natural ring of functions. Grothendieck's theory, developed in the 1950s and 1960s, uses $\mathrm{Spec}(R)$, the set of prime ideals of R, as the set X. This gives the notion of an affine scheme.

II. PROJECTIVE GEOMETRY

A. Infinity

Consider the affine line $\mathbb{A}^1$, which is simply the ground field K as a set. What should it mean to approach infinity in $\mathbb{A}^1$? Consider a ratio $x/y \in K$; if $y \neq 0$ this is an element of K, but as y approaches zero for fixed x, this element will "approach infinity." However, we cannot let y equal zero in this ratio, since one cannot divide by zero in a field. It makes sense then to separate the numerator and denominator in this ratio and consider the ordered pair: let us write $[y : x]$ for this. Since we are thinking of this ordered pair as representing a ratio, we should maintain that $[y : x] = [ay : ax]$ for any nonzero $a \in K$. The ordered pair $[1 : x]$ will represent the number $x \in K$; the ordered pair $[0 : 1]$ will represent a new point, at infinity. We remove the ordered pair $[0 : 0]$ from the discussion, since this would represent an undefined ratio.

B. Projective Space

The construction above generalizes to n-dimensional affine space $\mathbb{A}^n$, as follows. Let us define projective space $\mathbb{P}^n$ to be the set of all ordered $(n+1)$-tuples $[x_0 : x_1 : \ldots : x_n]$, with each $x_i \in K$, not all equal to zero, subject to the equivalence relation that

$$[x_0 : x_1 : \ldots : x_n] = [\lambda x_0 : \lambda x_1 : \ldots : \lambda x_n]$$

for any nonzero $\lambda \in K$. The x_i's are called the homogeneous coordinates of the point $[\underline{\mathbf{x}}]$; note that they are not individually well-defined, because of the scaling condition. However, it does make sense to say whether $x_i = 0$ or not.

If $x_0 \neq 0$, we can scale by $\lambda = 1/x_0$ and assume $x_0 = 1$; then all the other n coordinates are well-defined and correspond to a unique point in affine n-space $\mathbb{A}^n$:

$$p = (x_1, \ldots, x_n) \in \mathbb{A}^n \text{ corresponds to}$$
$$[1 : x_1 : \ldots : x_n] \in \mathbb{P}^n.$$

However, we have a host of new points where $x_0 = 0$ in $\mathbb{P}^n$; these are to be thought of as points "at infinity" of $\mathbb{A}^n$. In this way the points at infinity are brought into view and in fact become no different than any other point, in projective space.

C. Projective Algebraic Sets

Let $K[\underline{\mathbf{x}}] = K[x_0, x_1, \ldots, x_n]$ be the polynomial ring generated by the homogeneous coordinates. We can no longer view these polynomials as functions since (because of the scaling issue) even the coordinates themselves do not have well-defined values. However, suppose that a polynomial $F(\underline{\mathbf{x}})$ is homogeneous; that is, every term of F has the same degree d. Then $F(\lambda\underline{\mathbf{x}}) = \lambda^d F(\underline{\mathbf{x}})$ for any nonzero λ, and hence whether $F = 0$ or not at a point $[\underline{\mathbf{x}}] \in \mathbb{P}^n$ is well-defined.

We therefore define a projective algebraic set to be the set of common zeros of a collection $\mathcal{S}$ of homogeneous polynomials:

$$Z(\mathcal{S}) = \{[\underline{\mathbf{x}}] \in \mathbb{P}^n \mid F(\underline{\mathbf{x}}) = 0 \quad \text{for all } F \in \mathcal{S}\}.$$

Correspondingly, given a subset $X \subset \mathbb{P}^n$, we define the homogeneous ideal of X, $I(X)$, to be the ideal in $K[\underline{\mathbf{x}}]$ generated by all homogeneous polynomials F which vanish at all points of X:

$$I(X) = \langle \{\text{homogeneous } F \in K[\underline{\mathbf{x}}] \mid F|_X = 0\} \rangle.$$

The reader will see the possibility of developing exactly the same type of theory, complete with a $Z-I$ correspondence, a basis theorem, and a projective version of the Nullstellensatz, all properly interpreted with the extra attention paid to the homogeneity conditions at every turn. This is in fact what happens, and projective geometry attains an algebraic foundation as solid as that of affine geometry.

D. Regular and Rational Functions

In the context of projective geometry, the polynomial ring $K[\underline{\mathbf{x}}]$ is called the homogeneous coordinate ring. It is a graded ring, graded by degree: $K[\underline{\mathbf{x}}] = \oplus_{d\geq 0} V_d$ where V_d is the vector space of homogeneous polynomials in $\underline{\mathbf{x}}$ of degree exactly d. If $X \subset \mathbb{P}^n$ is a projective algebraic set, the homogeneous ideal $I(X)$ is also graded: $I(X) = \oplus_d I_d$, and the graded quotient ring $K[X] = K[\underline{\mathbf{x}}]/I(X)$ is called the homogeneous coordinate ring of X.

As we have noted above, a polynomial (even a homogeneous one) F does not have well-defined values at points of projective space. However, a ratio of polynomials $r = F/G$ will have a well-defined value at a point p if both F and G are homogeneous of the same degree, and $G(p) \neq 0$. Such a ratio is called a rational function of degree zero, and these functions form the foundation of the function theory on projective algebraic sets. A rational function whose denominator does not vanish at p is said to be regular at p.

E. Homogenization

Let us suppose now that we have an affine algebraic set, given by the vanishing of a polynomial $f(x_1, \ldots, x_n)$. We want to investigate how this algebraic set looks at infinity, in projective space. Let d be the largest degree of terms of f. To each term of f, we multiply in the appropriate power of the new variable x_0 to make the term have degree exactly d. This produces a homogeneous polynomial $F(x_0, x_1, \ldots, x_n)$, which for $x_0 = 1$ has the same zeros as the original polynomial f. However, now for $x_0 = 0$ we have new zeros, at infinity, in projective space.

For a general affine algebraic set X defined by more than one equation, we do this process for every polynomial in $I(X)$, producing a collection of homogeneous polynomials $\mathcal{S}$. Then $Z(\mathcal{S}) \subset \mathbb{P}^n$ is the projective closure $\bar{X}$ of X and exactly adds the minimal set of points at infinity to X to produce a projective algebraic set.

For example, let us consider the hyperbola in the affine plane defined by $f(x, y) = x^2 - y^2 - 1 = 0$. We homogenize this to $F(z, x, y) = x^2 - y^2 - z^2$, a homogeneous polynomial of degree two. For $z = 0$, we recover the two points $[0:1:1]$ and $[0:1:-1]$ at infinity (where $x^2 - y^2 = 0$), up to scalar factors.

F. Bezout's Theorem

One of the consequences of adding in the points at infinity to form projective space is that algebraic sets which did not intersect in affine space now tend to intersect, at infinity, in projective space. The classic example is that of two parallel lines in the affine plane: these intersect at one point at infinity in the projective plane.

In general, one expects that each time one adds a new equation that must vanish, the dimension of the set of zeros goes down by one. Therefore, in projective space $\mathbb{P}^n$, which has dimension n, one expects that the projective algebraic set defined by exactly n homogeneous polynomials $F_1 = F_2 = \cdots = F_n = 0$ will have dimension zero; this means that it should be a finite set of points.

Bezout's theorem deals with the number of such intersections. It may happen that a point of intersection is counted with multiplicity; this is the same phenomenon as when a polynomial in one variable has a double root: it counts for two when the number of roots is being enumerated. In more than one variable, there is a corresponding notion of multiplicity; this is always an integer at least one (for a common isolated root).

Theorem 3 (Bezout's theorem). Suppose the ground field K is algebraically closed. Let F_i, $i = 1, \ldots, n$, be homogeneous polynomials in the $n+1$ homogeneous variables of projective n-space $\mathbb{P}^n$. Suppose that F_i has degree d_i. Then,

(a) *The common zero locus $X = Z(\{F_1, \ldots, F_n\})$ is nonempty.*
(b) *If X is finite, the cardinality of X is at most the product of the degrees $d_1 d_2 \cdots d_n$.*
(c) *If each point of X is counted according to its multiplicity, the sum of the multiplicities is exactly the product of the degrees $d_1 d_2 \cdots d_n$.*

For example, three quadrics (each of degree two) in $\mathbb{P}^3$ always intersect. If the intersection is finite, it is at most eight points; if the points are counted according to their multiplicity, the sum is exactly eight.

As another example, a line and a cubic in the plane will intersect three times, counted with multiplicity. This is the basis for the famous "group law" on a plane projective cubic curve X: given two points p and q on X, the line joining p and q will intersect the curve X in a third point, and this operation forms the basis for a group structure on X.

III. CURVES AND SURFACES

A. Singularities

Let X be an irreducible algebraic variety, of codimension r, in $\mathbb{A}^n$. If x_i are local affine variables at a point $p \in X$, and the ideal of X is generated near p by polynomials $f_1, \ldots, f_k$, then the Jacobian matrix $J = (\partial f_i / \partial x_j)$ will have rank at most r at every point of X and will have maximal rank r at all points of X away from a subalgebraic set $\mathrm{Sing}(X)$. At points of $X - \mathrm{Sing}(X)$, the variety

is said to be nonsingular, or smooth; at points of $\mathrm{Sing}(X)$, the variety is said to be singular, and $\mathrm{Sing}(X)$ is called the singular locus of X. Over the complex numbers, the nonsingular points are those points of X where X is a complex manifold, locally analytically isomorphic to an open set in $\mathbb{C}^d$, where d is the dimension of X. A common situation is when $\mathrm{Sing}(X)$ is empty; then the variety is said to be smooth. At smooth points of algebraic varieties, there are local "analytic" coordinates $y_1, \ldots, y_d$ equal in number to the dimension of the variety.

B. 1-Forms

A 1-form on a smooth algebraic variety is a collection of local expressions of the form $\sum_i f_i(\mathbf{y})\, dy_i$, where $\{y_i\}$ are local coordinates on the variety and the f_i's are rational functions; this collection of expressions must be all equal under changes of coordinates, and at every point of the variety at least one of the expressions must be valid. For a curve, where there is only one local coordinate y, a local 1-form expression has the simple form $f(y)\, dy$. The set of 1-forms on a variety form a vector space.

C. The Genus of a Curve

Let X be a smooth projective curve. Let ω be a 1-form on X. If at every point of X there is a local expression for ω of the form $f(y)\, dy$ where y is a local coordinate and f is a regular function, then we say that ω is a regular 1-form. The set of regular 1-forms on a smooth projective curve form a finite-dimensional vector space over K, and the number of linearly independent regular 1-forms, which is the dimension of this vector space, is the most important invariant of the curve, called the genus.

If K is the complex numbers, the genus has a topological interpretation also. A smooth projective curve is a compact complex manifold of dimension one, which is therefore a compact orientable real manifold of dimension two. As such, it is topologically homeomorphic to a sphere with g handles attached; this g is the topological genus, which is equal to the genus. If $g = 0$, the curve is topologically a sphere; if $g = 1$, the curve is topologically a torus; if $g \geq 2$, the curve is topologically a g-holed torus.

The simplest smooth projective curve is the projective line itself, $\mathbb{P}^1$. It has genus zero.

D. Plane Curves and Plücker's Formula

The most common projective curves studied over the centuries are the plane curves, which are defined by a single irreducible polynomial $f(x, y) = 0$ in affine 2-space, and then closed up with points at infinity to a projective curve defined by the homogenization $F(x, y, z) = 0$ in the projective plane. Plücker's formula addresses the question of the genus of this curve in relation to the degree of the polynomial F:

Theorem 4 (Plücker's formula). *Suppose X is a smooth projective plane curve defined by an irreducible polynomial $F(x, y, z)$ of degree d. Then the genus of X is equal to $(d-1)(d-2)/2$.*

Plücker's formula has been generalized to curves with singularities, in particular with simple singularities like nodes (locally like $xy = 0$ or $y^2 = x^2$) and cusps (locally like $y^2 = x^3$). Then the formula gives the genus of the desingularization of the curve: if a plane projective curve of degree d has a finite number of singularities which are all either nodes or cusps, and there are ν nodes and κ cusps, then the genus of the desingularization is

$$g = \frac{(d-1)(d-2)}{2} - \nu - \kappa.$$

E. Elliptic and Hyperelliptic Curves

Using Plücker's formula, we see that smooth plane projective curves of degree one or two have genus zero. We have in fact seen above that conics may be parametrized by lines, and this is possible precisely because they have the same genus. Smooth projective plane cubic curves have genus one, and they cannot be parametrized by lines. Every smooth projective plane curve over $\mathbb{C}$ has exactly nine inflection points, and if an inflection point is put at the projective point $[0:1:0]$, and if the line of inflection is the line at infinity, the affine equation of the cubic curve can be brought into Weierstrass form

$$y^2 = x^3 + Ax + B$$

for complex numbers A and B, such that $4A^3 + 27B^2 \neq 0$ (this is the smoothness condition). The form of this equation shows that a cubic curve is also representable as a double cover of the x-line: for every x-value there are two y-values each giving points of the curve.

In general, there are curves of every genus representable as double covers, using similar affine equations of the form

$$y^2 = f_d(x),$$

where f_d is a polynomial with distinct roots of degree d. Such a curve is called hyperelliptic and has genus $g = [(d-1)/2]$. In particular these constructions show the existence of curves of any genus $g \geq 0$.

Higher-degree coverings of the projective line, such as a curve given by an equation of the form $y^3 = f(x)$, are also important in the classification of curves. Coverings of degree three are called trigonal curves, tetragonal curves are covers of degree four, and so forth.

F. Rational Functions, Forms, and the Riemann-Roch Theorem

The most celebrated theorem in the theory of curves is the theorem of Riemann-Roch, which gives precise information about the rational functions on a smooth projective curve X. The space of rational functions $K(X)$ on X forms a field of transcendence degree one over K, and as such is an infinite-dimensional vector space over K. In order to organize these functions, we concentrate our attention on the zeros (where the numerator vanishes) and the poles (where the denominator vanishes). Specifically, given a point $p \in X$ and a positive integer m, we may look at all the rational functions on X with a zero at p, no zero or pole at p, or a pole at p of order at most m. We denote this space by $\mathcal{L}(mp)$. If $m = -n$ is a negative integer, we define $\mathcal{L}(mp) = \mathcal{L}(-np)$ to be the space of rational functions with a zero at p of order at least n.

A divisor on X is a function from the points of X to the group of integers $\mathbb{Z}$, such that all but finitely many values are zero. For any divisor D, we define the vector space $L(D)$ to be the space $L(D) = \cap_{x \in X} \mathcal{L}(D(x) \cdot x)$. In plain English this is the space of rational functions with restricted poles (to the set of points with positive D-values) and prescribed zeros (as the set of points with negative D-values). Part of the Riemann-Roch theorem is that these spaces are finite-dimensional.

We can make the same construction with rational 1-forms also. Let us recall that a rational 1-form has local expression $f(y)\,dy$ and use the function part f to define the zeros and poles. Given a divisor E, we may then consider the space $\Omega^1(E)$ of rational 1-forms with restricted poles and prescribed zeros as E indicates. Again, this is a finite-dimensional space of forms.

Theorem 5 (Riemann-Roch). Let X be a smooth projective curve of genus g. Let D be a divisor on X, and denote by d the sum of the values of D: $d = \sum_x D(x)$. Then,

$$\dim(L(D)) = d + 1 - g + \dim(\Omega^1(-D)).$$

The inequality $\dim(L(D)) \geq d + 1 - g$ was proved by Riemann; Roch supplied the "correction term" related to 1-forms. One of the main uses of the Riemann-Roch theorem is to guarantee the existence of rational functions with prescribed zeros and poles, in order to intelligently embed the curve in projective space: given rational functions $f_0, \ldots, f_n$ on a curve X, the mapping sending $x \in X$ to the point $[f_0(x), f_1(x), \ldots, f_n(x)]$ will always be well-defined on all of X and under certain circumstances will embed X as a subvariety of $\mathbb{P}^n$.

Using this technique, we can show for example that (a) all curves of genus zero are isomorphic to the projective line $\mathbb{P}^1$; (b) all curves of genus one are isomorphic to smooth plane cubic curves; (c) all curves of genus two are hyperelliptic, given by an equation of the form $y^2 = f_6(x)$, where f_6 is a sextic polynomial; (d) all curves of genus three are either hyperelliptic or smooth plane quartic curves; (e) all curves of genus four are either hyperelliptic or the intersection of a quadric surface and a cubic surface in $\mathbb{P}^3$; and (f) all curves of genus five are hyperelliptic, trigonal, or the intersection of three quadric threefolds in $\mathbb{P}^4$. Of special interest is the so-called canonical embedding of a smooth curve, which shows that a curve of genus g either is hyperelliptic or can be embedded as a curve of degree $2g - 2$ in $\mathbb{P}^{g-1}$.

G. The Moduli Space M_g

Much of the research in algebraic geometry since 1960 has focused on the study of the moduli spaces for algebraic varieties. In general, a moduli space is a topological space M whose points represent particular geometric objects; the topology on M is such that points which are close in M represent geometric objects which are also "close" to each other. Of particular interest has been the moduli space $\mathcal{M}_g$ for smooth projective curves of genus g. For example, since every curve of genus zero is isomorphic to the projective line, $\mathcal{M}_0$ consists of a single point. The moduli space $\mathcal{M}_1$ classifying curves of genus one is isomorphic to the affine line $\mathbb{A}^1$: the famous j-invariant of elliptic curves classifies curves of genus one by a single number. [For a plane cubic curve of genus one in Weierstrass form $y^2 = x^3 + Ax + B$, $j = 4A^3/(4A^3 + 27B^2)$.] For $g \geq 2$, $\mathcal{M}_g$ is itself an algebraic variety of dimension $3g - 3$.

Of particular interest has been the construction and study of meaningful compactifications of various moduli spaces. For $\mathcal{M}_g$, the most natural compactification $\overline{\mathcal{M}_g}$ was constructed by Deligne and Mumford and the additional points represent stable curves of genus g, which are curves without continuous families of automorphisms, and having only nodes as singularities. Even today the construction and elementary properties of moduli spaces for higher-dimensional algebraic varieties (e.g., surfaces) is a challenge. More recently attention has turned to moduli spaces for maps between algebraic varieties, and it is an area of very active research today to compactify and understand such spaces of maps.

H. Surfaces and Higher Dimensions

The construction and understanding of the moduli spaces $\mathcal{M}_g$ for smooth curves is tantamount to the successful classification of curves and their properties. The classification of higher-dimensional varieties is not anywhere near as

complete. Even surfaces, for which a fairly satisfactory classification exists due to Enriques, Kodaira, and others, presents many open problems.

The Enriques classification of smooth surfaces essentially breaks up all surfaces into four categories. The first category consists of those surfaces with a family of genus zero curves on them. Since genus zero curves are all isomorphic to lines, such surfaces are known as ruled surfaces, and a detailed understanding of them is possible. The prototype is a product surface $X \times \mathbb{P}^1$ for a curve X. The second category consists of surfaces with a nowhere-vanishing regular 2-form, or finite quotients of such surfaces. These are the so-called abelian surfaces, $K3$ surfaces, Enriques surfaces, and hyperelliptic surfaces. The third category consists of surfaces with a family of genus one curves on them. Some techniques similar to those used in the study of ruled surfaces are possible, and since genus one curves are very well understood, again a rather detailed description of these surfaces, called elliptic surfaces, is available. The last category is the surfaces of general type, and most surfaces are in this category. Moduli spaces have been constructed, and many elementary invariants are known, but there is still a lot of work to do to understand general-type surfaces. An exciting current application area has to do with the connection of algebraic surfaces (which over the complex numbers are real four-dimensional objects) and the study and classification of 4-manifolds.

For varieties of dimension three or more, there are essentially no areas of complete classification. Basic techniques available for curves and surfaces begin to break down for threefolds, in particular because several fundamental constructions lead immediately to singular varieties, which are much more difficult to handle. However, since the 1980s, starting with the groundbreaking work of Mori, steady progress has been made on fundamental classification constructions.

IV. APPLICATIONS

The origins of algebraic geometry, and the development of projective geometry in the Renaissance, were driven in large part by applications (of the theory of multivariable polynomials) to a variety of problems in geography, art, number theory, and so forth. Later, in the 1800s, newer problems in differential equations, fluid flows, and the study of integrals were driving development in algebraic geometry. In the 1900s the invention of algebra as we know it today caused a rethinking in the foundations of algebraic geometry, and the energies of working researchers were somewhat siphoned into the development of new structures and techniques: these have culminated in the widespread use of general schemes, sheaves, and homological algebra, as well as slightly more general notions of algebraic spaces and stacks.

After this period of partial introspection, a vigorous return to application areas in algebraic geometry began in the 1980s. We will close this article with brief discussions of a sampling of these.

A. Enumeration

One of the fundamental questions of geometry, and of many other subjects in mathematics, is the "how many" question: in algebraic geometry, this expresses itself in counting the number of geometric objects with a given property. In contrast to pure combinatorics, often the geometric approach involves counting objects with multiplicity (e.g., roots of a polynomial).

Typical enumerative questions (and answers) ask for the number of flexes on a smooth cubic curve (9); the number of lines on a smooth cubic surface (27); the number of conics tangent to five given conics (3264); and the number of lines on a smooth quintic threefold (2875).

Recent breakthrough developments in intersection theory have enabled algebraic geometers to compute such numbers of enumerative interest which have stumped prior generations. New excitement has been brought by totally unexpected relationships with string theory in theoretical physics, where the work of Witten and others has found surprising connections between computations relating elementary particles and generating functions for enumerative problems of the above sort in algebraic geometry.

B. Computation

Computation with polynomials is a fundamentally algorithmic process which permits in many cases the design of explicit algorithms for calculating a multitude of quantities of interest to algebraic geometers. Usually these quantities either are related to enumerative questions such as those mentioned above or are the dimensions of vector spaces (typically of functions or forms) arising naturally either in elementary settings or as cohomology spaces.

With the advent of sufficient computing power as represented by modern computers and computer algebra software packages, it has become more possible to actually perform these types of computations by means of computer. This has led to an explosion of activity in designing efficient and effective algorithms for the computations of interest. These algorithms usually rely in some way on the theory of Gröbner bases, which build on a multivariable version of the division algorithm for polynomials.

Software is now widely available to execute many calculations, and it is typically user-customizable, which has enabled researchers around the world to make

fundamental contributions. Software packages which are widely used include Macaulay, CoCoA, and Schubert.

C. Mechanics

Many problems in mechanical engineering and construction involve the precise understanding of the position of various machine parts when the machine is in motion. Robotic analysis is especially concerned with the position and velocity of robot extremities given various parameters for joint motion and extension. There are a few basic motions for machine joints, and these are all describable by simple polynomials [e.g., circular motion of radius r occurs on the circle $(x^2 + y^2 = r^2)$]. It is not difficult to see therefore that in suitable coordinate systems, virtually all such problems can be formulated by polynomial systems.

However, mechanical devices with many joints can correspond to algebraic sets in affine spaces with many variables and many equations, which makes the geometric analysis rather complicated. It is therefore useful to be able to apply more sophisticated techniques of the theory to reduce the complexity of the problem, and this is where the power of algebraic geometry can come into play.

A specific example of the type of problem is the following so-called n-bar configuration. Let us consider a cyclical arrangement of n stiff rods, joined at the ends by joints that can actuate only in a planar way locally. For what initial positions of this arrangement does the configuration actually move and flex? In how many different ways can it be made to flex? When $n = 3$, the configuration must lie in a plane, and no flexing or motion is possible; for $n = 4$, the flexing configurations, although known, have not been completely analyzed.

D. Coding Theory

A code is a collection of vectors in K^n for some n, where K is a field; each vector in the collection is a code word. The Hamming distance between two code words is the number of positions where the code words differ. If each code word is intended to represent some piece of irreducible information, then desirable properties of a code are as follows:

(a) *Size*: There should be many code words so that a large amount of information can be represented.
(b) *Distinctness*: The Hamming distance between any two code words should be large, so that if a code word is corrupted in some way, the original code word can be recovered by changing the corrupted code word in a small number of positions.
(c) *Efficiency*: The ambient dimension n, which is the number of positions, should be as small as possible.

These three properties tend to act against one another, and the theory of error-correcting codes is directed toward the classification and analysis of possible codes and coding schemes.

Algebraic geometry has found an application in this area, by taking the field K to be a finite field, and taking the code to be certain natural spaces of functions or forms on an algebraic variety. Most successful have been attempts to use algebraic curves; this was initiated by Goppa and has been successful in producing codes with desirable properties and in aiding in the classification and uniform treatment of several families of previously known codes.

E. Automatic Theorem Proving

It is often the case, especially in elementary geometry, that geometric statements can be expressed by having some polynomial vanish. For example, three points (x_1, y_1), (x_2, y_2), and (x_3, y_3) in the plane are collinear if and only if the determinant of the matrix

$$\begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{pmatrix},$$

which is the polynomial $x_2y_3 - x_3y_2 - x_1y_3 + x_3y_1 + x_1y_2 - x_2y_1$, is zero.

A typical theorem therefore might be viewed as a collection of hypothesis polynomials $h_1, \ldots, h_k$, and a conclusion polynomial g. The truth of the theorem would be equivalent to saying that wherever the hypothesis polynomials all vanish, the conclusion polynomial is also zero. This exactly says that $g \in I(Z(\{h_1, \ldots, h_k\}))$, using the language of the $Z-I$ correspondence of affine algebraic geometry.

If the field is algebraically closed, the Nullstellensatz says that the conclusion polynomial g is in this ideal if and only if some power of g is a linear combination of the h_i's; that is, there is some equation of the form

$$g^r = \sum_i f_i h_i,$$

where f_i are other polynomials. A "proof" of the theorem would be given by an explicit collection of polynomials f_i, which exhibited the equation above for some r, and this can be easily checked by a computer. Indeed, algorithms exist which check for the existence of such an expression for g and are therefore effective in determining the truth of a given proposed theorem.

F. Interpolation

A general problem in approximation theory is to construct easy-to-evaluate functions with specified behavior.

Typically this behavior involves the values of the function at prescribed loci, the values of the derivatives of the functions, or both. Polynomial interpolation is the method whereby polynomial functions are used in this manner, and algebraic geometry has found recent applications and new problems of interest in this field in recent decades.

Lagrange interpolation involves finding polynomial functions with specified values c_k at a specified set of points $p_k \in \mathbb{A}^n$. In one variable, there is a relatively easy formula for writing down the desired polynomial, and this is taught in most first-year calculus courses. In higher dimensions, no such formula exists. However, the problem is a linear problem, and so it is a straightforward application of linear algebra techniques.

Hermite interpolation involves finding polynomial functions with specified derivative values. This is a significantly more complicated generalization, and open questions exist even for polynomials in two variables.

Spline interpolation involves finding piecewise polynomial functions, which stitch together to have a certain degree of global smoothness, but which also have specified behavior at given points. Cubic splines are the most popular in one variable, and with the advent of computer graphics, two- and three-variable splines are becoming more widely known and used in elementary applications.

In all these settings, techniques of algebraic geometry are brought to bear in order to determine the dimension of the space of suitable interpolating functions. Usually it is a simple matter to find a lower bound for such dimensions, and the more difficult problem is to find sharp upper bounds.

G. Graphics

The origins of projective geometry lie in the early Renaissance, when artists and architects began to be interested in accurately portraying objects in perspective. The correct treatment of horizon lines and vanishing points in artwork of the period led directly to a new appreciation of points at infinity and eventually to a mathematically sound theory of projective geometry.

With the recent explosion of capabilities in computer speed, resolution of displays, and amount and interest of visualizing data, an appreciation of projective geometry, homogeneous coordinates, and projective transformations has lately become standard fare for computer graphics specialists and novices alike.

Although initially the application of the ideas of projective geometry were rather elementary, more sophisticated techniques are now being brought to bear, especially those involving problems in computer vision and pattern recognition. In particular it now seems feasible to use subtle projective invariants to discriminate between scene objects of either different types at the same perspective or the same type at different perspectives.

SEE ALSO THE FOLLOWING ARTICLES

ABSTRACT ALGEBRA • BOOLEAN ALGEBRA • CALCULUS

BIBLIOGRAPHY

Beauville, A. (1996). "Complex Algebraic Surfaces," 2nd ed., Cambridge Univ. Press, Cambridge, UK.

Cox, D., Little, J., and O'Shea, D. (1997a). "Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra," 2nd ed., Springer-Verlag, New York.

Cox, D., Little, J., and O'Shea, D. (1997b). "Using Algebraic Geometry," Graduate Texts in Mathematics, Vol. 185, Springer-Verlag, New York.

Eisenbud, D. (1996). "Commutative Algebra with a View toward Algebraic Geometry," Graduate Texts in Mathematics, Vol. 150, Springer-Verlag, New York.

Fulton, W. (1997). "Intersection Theory," 2nd ed., Springer-Verlag, New York.

Griffiths, P., and Harris, J. (1978). "Principles of Algebraic Geometry," Wiley (Interscience), New York.

Harris, J. (1992). "Algebraic Geometry: A First Course," Springer-Verlag, New York.

Harris, J., and Morrison, I. (1998). "Moduli of Curves," Springer-Verlag, New York.

Hartshorne, R. (1979). "Algebraic Geometry," Springer-Verlag, New York.

Kirwan, F. (1992). "Complex Algebraic Curves," Cambridge Univ. Press, Combridge, UK.

Miranda, R. (1995). "Algebraic Curves and Riemann Surfaces," Am. Math. Soc., Providence.

Reid, M. (1988). "Undergraduate Algebraic Geometry," Cambridge Univ. Press, Combridge, UK.

Shafarevich, I. (1994). "Basic Algebraic Geometry," 2 vols., Springer-Verlag, New York.

Sturmfels, B. (1996). "Grobner Bases and Convex Polytopes," Am. Math. Soc., Providence.

Ueno, K. (1999). "Algebraic Geometry I: From Algebraic Varieties to Schemes," Am. Math. Soc., Providence.



Approximations and Expansions

Charles K. Chui
University of Missouri–St. Louis

I. Background
II. Best Approximation
III. Bivariate Spline Functions
IV. Compact Operators and *M* Ideals
V. Constrained Approximations
VI. Factorization of Biinfinite Matrices
VII. Interpolation
VIII. Multivariate Polyhedral Splines
IX. Quadratures
X. Smoothing by Spline Functions
XI. Wavelets

GLOSSARY

Center Let A be a bounded set in a normed linear space X. An element x_0 in X is called a center (or Chebyshev center) of A if $\sup\{\|x_0 - y\| : y \in A\} = \inf_{x \in X} \sup\{\|x - y\| : y \in A\}$.

***L* Summand** A closed subspace S of a Banach space X is said to be an L summand (or L ideal) if there exists a closed subspace T of X such that $X = S \otimes T$ and $\|x + y\| = \|x\| + \|y\|$ for all $x \in S$ and $y \in T$.

***M* Ideal** A closed subspace S of a Banach space X is called an M ideal in X if its annihilator $S^{\perp}$ is an L summand of the dual space X^*.

Modulus of smoothness Let f be in $L_p[0, 1]$, $1 \leq p \leq \infty$, and $\Delta_t^r f$ denote the rth forward difference of f with step size t, where $rt \leq 1$. The L_p rth modulus of smoothness of f is defined by $w_r(f, h)_p = \sup\{\|\Delta_t^r f\|_p : 0 < t < h\}$, where the L_p norm is taken over all of $[0, 1]$ for the periodic case and over $[0, 1 - rt]$ for the nonperiodic setting. Also, denote $w(f, h)_p = w_1(f, h)_p$.

Natural spline Let $a = t_0 < \cdots < t_{n+1} = b$. A function s in $C^{2k-2}[a, b]$ is called a natural (polynomial) spline of order $2k$ and with knots at $t_1, \ldots, t_n$, if the restriction of s to each $[t_i, t_{i+1}]$ is a polynomial of degree $2k - 1$ for $i = 1, \ldots, n - 1$, and a polynomial of degree $k - 1$ on each of the intervals $[a, t_1]$ and $[t_n, b]$.

Normalized *B* spline Let $t_i \leq \cdots \leq t_{i+k}$ with $t_i < t_{i+k}$. The ith normalized B spline of order k and with knots at $t_i, \ldots, t_{i+k}$, is defined by $N_{i,k}(x) = (t_{i+k} - t_i)[t_i, \ldots, t_{i+k}](\cdot - x)_+^{k-1}$, where the kth order divided difference of the truncated $(k - 1)$st power is taken at $t_i, \ldots, t_{i+1}$.

***n* Widths** Let A be a subset of a normed linear space X. The Kolmogorov n width of A in X is the quantity $d_n(A; X) = \inf_{X_n} \sup_{x \in A} \inf_{y \in X_n} \|x - y\|$, where X_n is any subspace of X with dimension at most n. The linear

n width of A in X is the quantity $\delta_n(A; X) = \inf_{P_n} \sup_{x \in A} \|x - P_n x\|$, where P_n is any set of continuous linear operators of X into itself of rank at most n. The Gel'fand n width of A in X is the quantity $d^n(A; X) = \inf_{L_n} \sup_{x \in A_n \cap L_n} \|x\|$, where L_n is any closed subspace of X with codimension at most n. The Bernstein n width of A in X is the quantity $b_n(A; X) = \sup_{X_{n+1}} \sup\{t : tB(X_{n+1}) \subseteq A\}$, where X_{n+1} is any subspace of X with dimension at least $n+1$ and $B(X_{n+1})$ is the unit ball $\{x \in X_{n+1} : \|x\| \leq 1\}$ in X_{n+1}.

Padé approximants The $[m, n]$ Padé approximant of a formal power series $c_0 + c_1 z + c_2 z^2 + \cdots$ is the (unique) rational function p_m/q_n, where $p_m(z) = a_0 + \cdots + a_m z^m$ and $q_n(z) = b_0 + \cdots + b_n z^n$ are determined by the algebraic quantity $(c_0 + c_1 z + c_2 z^2 + \cdots)(b_0 + \cdots + b_n z^n) - (a_0 + \cdots + a_m z^m) = d_1 z^{m+n+1} + d_2 z^{m+n+2} + \cdots$ for some $d_1, d_2, \ldots$.

Total positivity An $n \times n$ square matrix $A = [a_{ij}]$ is said to be totally positive (TP) if the determinant of $[a_{i_m j_n}]$ is nonnegative for all choices of integers $1 \leq i_1 < \cdots < i_k \leq n$ and $1 \leq j_1 < \cdots < j_k \leq n$ and any integer $k = 1, \ldots, n$. It is said to be strictly totally positive (STP) if each of the above determinants is positive. A kernel $K(x, y)$, where $a \leq x \leq b$ and $c \leq y \leq d$, is TP or STP if the matrices $K(x_i, y_j)$ are TP or STP, respectively, for all choices of x_i and y_j with $a \leq x_1 < \cdots < x_m \leq b$ and $c \leq y_1 < \cdots < y_m \leq d$, and all $m = 1, 2, \ldots$.

Wavelets A function, with sufficiently fast decay at infinity, can be used as a wavelet to define the wavelet transform, if its Fourier transform vanishes at the origin.

APPROXIMATION of discrete data, complicated functions, solutions of certain equations, and soon by functions from a simple and usually finite dimensional space is often done in every field of science and technology. Approximation theory, also known as approximations and expansions, serves as an important bridge between pure and applied mathematics and is intimately related to numerical analysis. This field has no well-defined boundary and overlaps with various branches of mathematics. Among the most important areas of study are possibility of approximation, quality of approximation, optimal approximation, families of approximants, approximation schemes, and computational algorithms. Approximation of bounded operators by compact operators on a Banach space and factorization of biinfinite matrices are also important areas in the field of approximation theory. Recent development of approximation theory includes computer-aided geometric design (CAGD) and wavelets.

I. BACKGROUND

Approximation theory, which is often classified as approximations and expansions, is a subject that serves as an important bridge between pure and applied mathematics. Although its origin is traced back to the fundamental work of Bernstein, Chebyshev, Haar, Hermite, Kolmogorov, Lagrange, Markov, and others, this branch of mathematics was not fully established until the founding of the *Journal of Approximation Theory* in 1968. Now with the rapid advance of digital computers, approximation theory has become a very important branch of mathematics. Two other journals on this subject, *Approximation Theory and Its Applications and Constructive Approximation,* were founded in late 1984.

There is no well-defined boundary of the field of approximation theory. In fact, it overlaps with both classical and modern analysis, as well as numerical analysis, linear algebra, and even various branches of applied mathematics. We may nevertheless divide it into five areas: (1) possibility of approximation, (2) quality of approximation, (3) optimal approximation, (4) families of approximants, and (5) approximation schemes and computational algorithms. The first area is mainly concerned with density and completeness problems. The typical classical results in this area are the theorems of Stone-Weierstrass, Müntz-Szasz, Mergelian, Korovkin, and others. The subject of qualitative approximation includes mainly the study of degrees of approximations and inverse theorems. Jackson's theorems and the so-called saturation theorems are typical results in this area. Optimal approximation is probably the backbone of the field of approximation theory. The typical problems posed in this area are existence, uniqueness, characterization, and computation of best approximants, and the most basic result is probably the so-called alternation theorem. This area of mathematics is extremely important and includes the subjects of least-squares methods, minimal projections, orthogonal polynomials, optimal estimation, and Kalman filtering. Depending on the mathematical models or the nature of the approximation problems, certain families of approximants are to be used. The most familiar families of approximants are algebraic polynomials, trigonometric polynomials, rational functions, and spline functions. In this direction, there is a vast amount of literature, including, in particular, the study of total positivity. Unfortunately, most of the beautiful properties of these classical families of approximants do not carry over to the two- and higher-dimensional settings. The important subject of multivariate spline functions, for instance, has been a very active field of research in recent years. Again depending on the approximation problems, appropriate schemes of approximation are used. Computational algorithms are very important for the purpose of

yielding good approximations efficiently. This area of approximation theory includes the subjects of interpolation, least-squares methods, and approximate quadratures.

It must be noted that the separation of the field of approximation theory into five areas should not be taken very strictly since these areas obviously have to overlap. For instance, the elementary and yet important method of least-squares is an approximation scheme with efficient computational algorithms and yields an optimal approximation. In addition, many mathematicians and practitioners may give a broader or even different definition of approximation theory. Hence, only selected topics of this field are treated in this article and the selection of topics and statements of results has to be subjective. In addition, many interesting areas, such as approximation theory in the complex domain, trigonometric approximation and interpolation, numerical approximation and algorithms, inequalities, orthogonal polynomials, and asymptotic approximation, are not included.

II. BEST APPROXIMATION

Approximation theory is built on the theory of best approximation. Suppose that we are given a sequence of subsets $P_1 \subset P_2 \subset \cdots$ in a normed linear space X of functions with norm $\| \ \|$. For each function f in X, consider the distance

$$E_n(f) = \inf\{\|f - g\| : g \in P_n\}$$

of f from P_n. If there exists a function p_n in P_n such that $\|f - p_n\| = E_n(f)$, we say that p_n is a best approximant of f in X from P_n. Existence of best approximants is the first basic question in the theory of best approximation. Compactness of P_n, for instance, guarantees their existence. The second basic question is uniqueness, and another basic problem is to characterize them. Characterization is fundamental to developing algorithms for the construction of best approximants.

Suppose that the function f that we wish to approximate lies in some subset G of X. Let B be the unit ball of all f in G (with $\|f\| \leq 1$). Then the quantity

$$E_n(B) = \sup\{E_n(f) : f \in B\}$$

is called the order of approximation of functions in G from P_n. Knowing if, and if so how fast, $E_n(B)$ tends to zero is also one of the most basic problems in best approximation.

A. Polynomial and Rational Approximation

Consider the Banach space $C[0, 1]$ with the uniform (or supremum) norm. Let π_n be the space of all polynomials of degree at most n. Although π_n is not compact, a compactness argument still applies to yield the existence of a best approximant to each f in $C[0, 1]$ from π_n. It is also well known that best approximation from π_n has a unique solution. This can be proved by using the following characterization theorem, which is better known as the alternation theorem.

Theorem 1. p_n is a best approximant of f in $C[0, 1]$ from π_n if and only if there exist points

$$0 \leq x_1 < \cdots < x_{n+2} \leq 1$$

such that $(f - p_n)(x_i) = a(-1)^i \|f - p_n\|$ for $i = 1, \ldots, n + 2$, where $a = 1$ or -1.

More can be said about the uniqueness of best approximants as stated in the following so-called strong unicity theorem.

Theorem 2. Let f be in $C[0, 1]$ and p_n its best approximant from π_n. Then there exists a constant C_f such that

$$\|f - g\| \geq E_n(f) + C_f \|p_n - g\|$$

for all g in π_n.

There are results concerning the behavior of the constant C_f in the literature. All the above results on approximation by algebraic polynomials are also valid for any Chebyshev system.

Results on the orders (or degrees) of approximation by algebraic or trigonometric polynomials are called Jackson's theorems. Favard also gave sharp constants for best trigonometric polynomial approximation.

Although rational functions do not form a linear space, analogous results on existence, uniqueness, and characterization still hold. Let $R_{m,n}$ be the collection of all functions p/q, where $p \in \pi_m$ and $q \in \pi_n$, such that p and q are relatively prime and $q(x) > 0$ for all x on $[0, 1]$. The last condition is not restrictive since, for approximation purposes, p/q must be finite on $[0, 1]$. A more careful compactness argument shows that every function in $C[0, 1]$ has at least one best rational approximant from $R_{m,n}$. The following alternation theorem characterizes best rational approximants. We will use the notation $d(p)$ to denote the degree of a polynomial p.

Theorem 3. p_m/q_n is a best approximant of f from $R_{m,n}$ if and only if there exist points

$$0 \leq x_1 < \cdots < x_r \leq 1$$

where $r = 2 + \max\{m + d(q_n), n + d(p_m)\}$ such that $(f - p_m/q_n)(x_i) = a(-1)^i \|f - p_m/q_n\|$ for $i = 1, \ldots, r$, where $a = 1$ or -1.

Again uniqueness of best rational approximants can be shown using the alternation theorem, and in fact an analogous strong unicity result can be obtained. Unfortunately, although rational functions form a larger class than polynomials, they do not improve the orders of approximation in general. For instance, the approximation order of the class of all lip α functions by R_{nn} is $O(1/n^\alpha)$, which is the same as the order of approximation from π_n. However, while best approximation from π_n is saturated [e.g., $E_n(f) = O(1/n^\alpha)$ for $0 < \alpha < 1$ if and only if $f \in \text{lip } \alpha$, and $E_n(f) = O(1/n)$ if and only if f is in the Zygmund class], this is certainly not the case in rational approximation. A typical example is that the order of best approximation of the important function $|x - \frac{1}{2}|$ from R_{nn} is $O(e^{-\sqrt{n}})$, which is much better than $O(1/n)$.

B. Best Local Approximation

Because of the alternation properties of best polynomial and rational approximants, these approximants are believed to behave like Taylor polynomials and Padé approximants, respectively, when the interval of best approximation is small. This leads to the study of best local approximation.

Suppose we are given a set of discrete data $\{a_{jk}\}$, which are the values of some (unknown) function f and its derivatives at the sample points $x_1, \ldots, x_n$, say

$$a_{jk} = f^{(j)}(x_k), \qquad k = 1, \ldots, n$$

The problem is to give a "best" approximation of f using the available discrete data from some class P of functions (π_n, R_{mn}, etc.).

Let us assume, for the time being, that f is known in a neighborhood of the sample points. It will be clear that the best local approximant of f (or more precisely of the data $\{a_{jk}\}$) does not depend on f, except on its interpolatory conditions mentioned above. For each $k = 1, \ldots, n$, let $V_k = x_k + \varepsilon_k A_k$, where $\varepsilon_k = \varepsilon_k(\delta) > 0$ and tends to zero as $\delta \to 0$ and $A_k = A_k(\delta)$, be uniformly bounded measurable sets with unit measure, and let $V(\delta)$ be the union of $V_1, \ldots, V_n$. Now, for each δ, let $g_\delta = g_{\delta,p}$ be a best approximant of f from P in $L_p[V(\delta)]$, where $1 \le p \le \infty$ (and L_∞ will be replaced by C, as usual). Then if $g_\delta \to g_0$ as $\delta \to 0$, we say that g_0 is a best L_p local approximant of f or of the data $\{a_{jk}\}$.

We will discuss only the one-variable setting, although some of these results have multivariate extensions. We will also say that an n tuple $(i_1, \ldots, i_n)$ of integers is p-balanced (or simply balanced), if for each j with $i_j > 0$,

$$\sum_{k=1}^{n} \varepsilon_k^{i_k + 1/p} = O\left(\varepsilon_j^{i_j - 1 + 1/p}\right)$$

as $\delta \to 0$. If this n tuple is balanced, then their sum is called a balanced integer. There is an algorithm to generate all balanced integers.

Theorem 4. Let $(i_1, \ldots, i_n)$ be balanced, N its sum, and P_N a fully interpolating set of functions in $L_p[V(\delta)]$, in the sense that, for any discrete data $\{a_{jk}\}$, there exists a unique g in P_n such that

$$g^{(j)}(x_k) = a_{jk}$$

for $j = 0, \ldots, i_k - 1$ and $k = 1, \ldots, n$. Then for any f that is i_k-times differentiable around x_k, $k = 1, \ldots, n$, the best local L_p approximant of f from P_N exists and is the unique function g_0 that satisfies the Hermite interpolation condition

$$(f - g_0)^{(j)}(x_k) = 0$$

for $j = 0, \ldots, i_k - 1$ and $k = 1, \ldots, n$. Conversely, to any n tuple of positive integers $(c_1, \ldots, c_n)$ with sum M, there is some scaling $\varepsilon_k'(\delta) > 0$ such that this n tuple is balanced with respect to $\{\varepsilon_k'(\delta)\}$, $k = 1, \ldots, n$; consequently, any Hermite interpolation is a best L_p local approximation scheme.

The general case where $(i_1, \ldots, i_n)$ is not balanced is much more delicate, and the best result in this direction in the literature is stated as follows:

Theorem 5. Suppose $(i_1, \ldots, i_n)$ is not balanced and N is its sum. Assume that each A_k is either an interval for each δ or is independent of δ. Also assume that

$$\varepsilon_k(\delta)^{i_k' + 1/p} \left[\sum_{j=1}^{n} \varepsilon_j(\delta)^{i_j' + 1/p} \right]^{-1} = e_k + o(1)$$

as $\delta \to 0$, where $i_1' + \cdots + i_n'$ is the largest balanced integer that does not exceed N. Let $J_A(i, p)$ denote the minimum L_p norm on a measurable set A of the polynomial in π_i with unit leading coefficient. Suppose that f is i_k'-times differentiable around x_k for $k = 1, \ldots, n$ and $1 < p \le \infty$. Then the best L_p local approximant of f exists and is the unique solution of the constrained l_p minimization problem:

$$\min_{g \in P_N} \left\| \left\{ e_k J_{A_k}(i_k', p)(f - g)^{(i_k')}(x_k) i_k! \right\} \right\|_{l_p}$$

$$\text{subject to } \begin{cases} (f - g)^{(j)}(x_k) = 0 \\ j = 0, \ldots, i_k' - 1; \qquad k = 1, \ldots, n \end{cases}$$

We remark that if A_k is an interval, then the subscript A_k of J can be replaced by $[0, 1]$ and that for $p = 1$, the l_1 minimization may not have a unique solution, but if it does, it is the best L_1 local approximant of f.

In the special case when there is only one sample point, say the origin, then the best L_∞ local approximant to a

sufficiently smooth function f from R_{mn} is the $[m, n]$ Padé approximant of f. We remark, however, that the convergence of the net of best uniform approximants on small intervals to the Padé approximant is in general not uniform, but only in measure.

C. Padé Approximants

As just mentioned, Padé approximants are best local approximants and include Taylor polynomials, which are just $[m, 0]$ approximants. What is interesting is that Padé approximants to a formal power series are defined algebraically, so that they are certainly well defined even if the series has zero radius of convergence. In this case, taking suitable limits of the approximants on the Padé table can be considered a summability scheme of the divergent series. The most interesting example is, perhaps, the Stieltjes series,

$$a_0 + a_1 z + a_2 z^2 + \cdots$$

where

$$a_n = \int_0^\infty t^n \, d\mu$$

with μ being a positive measure on $[0, \infty)$. It has been shown that the diagonal Padé approximants $[n + j, n]$, where j is fixed, converge uniformly, in each compact set in the complex plane that does not intersect the interval $[0, \infty)$, to the Stieltjes integral

$$\int_0^\infty \frac{d\mu}{1 - zt}$$

provided that the coefficients a_n do not tend to infinity too fast, say

$$a_n = O[(2n + 1)! R^{2n}]$$

for some $R > 0$. A simple example is $a_n = n!$, where

$$d\mu = e^{-t} \, dt.$$

It has been proved, however, that the Padé approximants to a Stieltjes series diverge along any ray of the Padé table that is not parallel to the diagonal.

It is also interesting that, while Padé approximants along the diagonals behave so nicely for Stieltjes series, they do not necessarily converge even for some entire functions. One reason is that the distribution of the poles of the approximants cannot be controlled. Suppose that

$$f(z) = a_0 + a_1 z + \cdots$$

is actually a meromorphic function. Then along each row of the Padé table, where the denominators of the approximants have a fixed degree n, the poles of f usually attract as many poles of the approximants as possible. Results of this type are related to a classical theorem of de Montessus de Ballore. The following inverse result has been obtained. Let q_{mn} denote the denominator of the $[m, n]$ Padé approximant of f. Since n is fixed, we use any norm that is equivalent to a norm of the $(n + 1)$-dimensional space π_n.

Theorem 6. Suppose that there exist nonzero complex numbers z_i and positive integers p_i with $p_1 + \cdots + p_k = n$, such that

$$\lim_{m \to \infty} \sup \left\| q_{mn}(z) - \prod_{j=1}^{k} (z - z_i)^{P_i} \right\|^{1/m} = r < 1$$

Then f is analytic in

$$|z| < r^{-1} \max\{|z_j| : j + 1, \ldots, k\}$$

with an exception of the poles at $z_1, \ldots, z_k$ with multiplicities $p_1, \ldots, p_k$, respectively.

Without the geometric convergence assumed in Theorem 6, the following inverse result still holds.

Theorem 7. Suppose that

$$\lim_{m \to \infty} q_{mn}(z) = \prod_{j=1}^{n} (z - z_i)$$

where $0 < |z_1| \leq \cdots \leq |z_{n-1}| < |z_n|$. Then f is analytic in $|z| < |z_n|$ with an exception of the poles at $z_1, \ldots, z_{n-1}$, counting multiplicities, and has a singular point at z_n.

Let us return to the diagonal approximants. In particular, consider the main diagonal of the Padé table: $r_n = p_n / q_n$. Since there exist entire functions whose diagonal Padé approximants r_n diverge everywhere in the complex plane except at the origin, it is interesting to investigate the possibility of convergence when the poles of r_n can be controlled. In this direction, the following result has been obtained.

Theorem 8. Let $f(z) - r_n(z) = bz^{2n+1} + \cdots$ for each $n = 1, 2, \ldots$. Suppose that r_n is analytic in $|z| < R$ for all sufficiently large n. Then $\{r_n\}$ converges uniformly on every compact subset of $|z| < R$ to f, so that f is analytic in $|z| < R$.

It should be noted that no *a priori* analytic assumption was made on f. This shows that, once the location of the poles of the Padé approximants can be estimated, we can make a conclusion about the analyticity of the formal power series. This result actually holds for a much more general domain in the complex plane.

D. Incomplete Polynomials

Loosely speaking, if certain terms in an algebraic polynomial do not appear, it is an incomplete polynomial. We

shall discuss best approximations by incomplete polynomials. in $C[0, 1]$. Let H be a finite set of nonnegative integers, and consider the best approximation problem

$$E(f, H) = \inf_{c_i} \left\| f - \sum_{i \in H} c_i x^i \right\|$$

Note that $E(f, H) = E_n(f)$ if $H = \{0, \ldots, n\}$. We also denote

$$E_n^k(f) = E(f, \{0, \ldots, k-1, k+1, \ldots, n\})$$

for any $0 \le k \le n$. The following result has been obtained:

Theorem 9. Let $0 < c < 1$. There exists a positive constant M_k independent of n such that, if f is in $C^k[0, 1]$ and $f^{(k)}(c) \ne 0$, then

$$E_n^k(f) \ge M_k n^{-k}[|f^{(k)}(c)| + o(1)]$$

as $n \to \infty$.

Hence, since $E_n(f)$ is of order

$$O\left[n^{-k} E_n(f^{(k)})\right] = o(n^{-k})$$

we have $E_n^k(f)/E_n(f) \to \infty$ for each f in $C^k[0, 1]$ whose kth derivative does not vanish identically in $[0, 1]$. So even one term makes some difference in studying the order of approximation.

If only finitely many terms are allowed, the natural question is which are the optimal ones. Let $0 < n < N$ and $H_n + \{t_1, \ldots, t_n\}$ be a set of integers with $0 \le t_1 < \cdots < t_n < N$. The problem is to find a set H_n, which we call an optimal set, denoted by H_n^*, such that

$$E\left(f, H_n^*\right) = \inf\{E(f, H_n) : H_n\}$$

An analogous problem for trigonometric approximation has interesting applications to antenna design where certain frequencies are to be received by using a specified number of antennas. Unfortunately, the general problem is unsolved. We have the following result:

Theorem 10. For $f(x) = x^M$, where $M \ge N$, $H_n^* = \{N - n, \ldots, N - 1\}$ and is unique.

This result can be generalized to a Descartes system and even holds for any Banach function space with monotone norm. In a different direction, the analogous problem of approximating a given incomplete polynomial by a monomial has also been studied.

Next, consider $H_n = \{0, t_1, \ldots, t_n\}$, where $0 < t_1 < t_2 <, \ldots$ with

$$\sum_{i=1}^{\infty} \frac{1}{t_i} = \infty$$

and set $E_n(f) = E(f, H_n)$. The so-called Müntz–Szasz's theorem guarantees that $E_n(f) \to 0$ for every f in $C[0, 1]$. Hence, the order of approximation is of interest. Results in this area are called Müntz–Jackson theorems.

We end our discussion of incomplete polynomials by considering the density of the polynomials.

$$p_{n,a}(x) = \sum_{k=m(a)}^{n} c_k x^k, \qquad m(a) \ge na$$

where $0 < a < 1$.

Theorem 11. If $\|p_{n,a}\|$ is uniformly bounded, then

$$\lim_{n \to \infty} p_{n,a}(x) = 0, \qquad 0 \le x < a^2$$

and the convergence is uniform on every interval $[0, r]$, where $r < a^2$.

Thus, the best we can do is to consider approximation on $[a^2, 1]$:

Theorem 12. Let $f \in C(a^2, 1)$. Then there exists a sequence of polynomials $p_{n,a}$ such that

$$\lim_{n \to \infty} p_{n,a}(x) = f(x)$$

uniformly on $[r, 1]$ for any $r > a^2$.

E. Chebyshev Center

Suppose that, instead of approximating a single function f in a normed linear space X from a subset P of X, we are interested in approximating a bounded set A of functions in X simultaneously from P. Then the order of approximation is the quantity

$$r_P(A) = \inf\left\{ \sup_{f \in A} \|f - g\| : g \in P \right\}$$

Of course, if $P = P_n$ and A is a singleton $\{f\}$, then $r_P(A)$ reduces to $E_n(f)$, introduced at the beginning of this section. In general, $r_P(A)$ is called the Chebyshev radius of A with respect to P. A best simultaneous approximant x_0 of A from P, defined by

$$\sup_{f \in A} \|f - x_0\| = r_P(A)$$

is called a Chebyshev center (or simply center) of A with respect to P. We denote the (possibly empty) set of such x_0 by $c_P(A)$. In particular, we usually drop the indices if $P = X$; that is,

$$r(A) = r_X(A) \qquad \text{and} \qquad c(A) = c_X(A)$$

which are simply called the Chebyshev radius and the set of centers of A, respectively.

Most of the literature in this area is concerned with the existence problem; namely, $c_P(A)$ is nonempty. If $c_P(A) \ne \emptyset$ for all bounded nonempty sets A in X, we

say that P admits centers. A classical result is that if X is the range of a norm-one projection defined on its second dual, then X admits centers. In addition, if X admits centers and P is a norm-one complemented subspace, then P also admits centers. An interesting example is the Banach space of real-valued continuous functions on a paracompact space. Not only does it admit centers; the set of centers of any bounded nonempty set A of functions in this space is also relatively easy to describe in terms of A. If X is a uniformly rotund Banach space, then we even have unique best approximants in the sense that $c(A)$ is a singleton for any bounded nonempty set A in X. The following result has been obtained:

Theorem 13. Let X be a Banach lattice such that the norm is additive on the positive cone, and Q be a positive linear nonexpansive mapping of X into itself. Then the set of fixed points of Q admits centers.

The proof of this result depends on the fact that the Banach lattice X described here admits centers itself.

Now suppose that X is a Hilbert space and B its unit ball. Then a hypercircle in X is the (nonempty) intersection of some translate of a closed linear subspace of X with rB for some $r > 0$.

Theorem 14. The Chebyshev center of any hypercircle in a Hilbert space is the unique element of minimum norm.

This result is basic in the study of optimal estimation.

F. Optimal Estimation and Recovery

Suppose that a physical object u is assumed to be adequately modeled by an unknown element x_u of some normed linear space X, and observations or measurements are taken of u that give a limited amount of information about x_u. The data accumulated, however, are inadequate to determine x_u completely. The optimal estimation problem is to approximate x_u from the given data in some optimal sense. Let us assume that the data describe some subset A of X. Then the maximum error in estimating x_u from A is

$$\sup_{f\in A}\|f - x_u\|$$

In order for this quantity to be finite it is necessary and sufficient that A is a bounded subset of X, and we shall make that assumption. To optimize our estimate with only the additional knowledge that x_u must lie in some subset P of X we minimize the error, and this leads to the Chebyshev radius

$$r_P(A) = \inf\left\{\sup_{f\in A}\|f - x\| : x \in P\right\}$$

of A with respect to the "model" set P. A (Chebyshev) center, if it exists, is an optimal estimation of x_u.

A related but somewhat different problem in best approximation is optimal recovery. Let x be in a normed linear space X, and U an operator from X into another normed linear space Z. The problem is to recover Ux from a limited amount of information about x. Let I be the information operator mapping X into another normed linear space Y. Suppose Ix is known. To estimate Ux, we need an operator A from Y to Z, so that AIx approximates Ux. A is called an algorithm.

Usually, we assume that x is restricted to a balanced convex subset K of X. Here, K is said to be balanced if $v \in K$ implies that $-v \in K$. If the information Ix of x may not be precise, say with some tolerance $\varepsilon \geq 0$, then an algorithm A that is used to recover Ux has the maximum error

$$E(A) = \sup\{\|Ux - Ay\| : x \in K,\ \|Ix - y\| \leq \varepsilon\}$$

Hence, to recover Ux optimally we have to determine an algorithm A^*, called an optimal algorithm, such that

$$E(A^*) = \inf_A E(A)$$

It should be remarked that an optimal algorithm may be nonlinear, although linear optimal algorithms exist in a fairly general situation. A lower bound of $E(A^*)$ usually helps, and it can be found in

$$E(A^*) \geq \sup\{\|Ux\| : x \in K, \|Ix\| \leq \varepsilon\}$$

There has been some progress in optimal recovery of nonlinear operators, and results on the Hardy spaces have also been obtained. It is not surprising that (finite) Blaschke products play an essential role here in the H^p problems.

G. *n* Widths

Let A be a set in a normed linear space X that we wish to approximate from an n-dimensional subspace X_n of X. Then the maximum error (or order of approximation if A is the unit ball) is the quantity

$$E(A, X_n) = \sup_{x\in A}\inf\{\|x - y\| : y \in X_n\}$$

It was Kolmogorov who proposed finding an optimal n-dimensional subspace $\bar{X}_n$ of X. By this, we mean

$$E(A, \bar{X}_n) = \inf_{X_n} E(A, X_n)$$

where the infimum is taken over all subspaces X_n of X with dimension at most n. This optimal quantity is called the Kolmogorov n width of A and is denoted by

$d_n(A) = d_n(A; X)$. Knowing the value of $d_n(A)$ is important in identifying an optimal (or extremal) approximating subspace $\bar{X}_n$ of X.

Replacing the distance of $x \in A$ from X_n in Kolmogorov's definition by the distance of x from its image under a continuous linear operator P_n of rank at most n of X into itself yields the notion of the linear n width of A in X, defined by

$$\delta_n(A) = \delta_n(A; X) = \inf_{P_n} \sup_{x \in A} \|x - P_n x\|$$

It is clear from the two definitions that $\delta_n(A) \geq d_n(A)$.

A dual to the Kolmogorov n width is the Gel'fand n width defined by

$$d^n(A) = d^n(A; X) = \inf_{L_n} \sup_{x \in A \cap L_n} \|x\|$$

where L_n is a closed subspace of X with codimension at most n. It has also been shown that $d^n(A)$ is also bounded above by the linear n width $\delta_n(A)$. To provide a lower bound for both $d_n(A)$ and $d^n(A)$, the Bernstein n width defined by

$$b_n(A) = b_n(A; X) = \sup_{X_{n+1}} \sup\{t : tB(X_{n+1}) \subseteq A\}$$

could be used. Here, X_{n+1} is a subspace of X with dimension at least $n+1$, and $B(X_{n+1})$ is the unit ball in X_{n+1}.

An important setting is that A is the image of the unit ball in a normed linear space Y under some $T \in L(Y, X)$, the class of all continuous linear operators from Y to X. Although the set A is *not* TY, the commonly used notations in this case are $d_n(TY; X)$, $\delta_n(TY; X)$, $d^n(TY; X)$, and $b_n(TY; X)$. Let $C(Y, X)$ be the class of all compact operators in $L(Y, X)$. Then the duality between d_n and d^n can be seen in the following theorem:

Theorem 15. Let $T \in L(Y, X)$ and $T^* \in L(X^*, Y^*)$ be its adjoint. Suppose that $T \in C(Y, X)$ or that X is a reflexive Banach space. Then

$$d_n(TY; X) = d^n(T^*X^*; Y^*)$$

and

$$\delta_n(TY; x) = \delta_n(T^*X^*; Y^*)$$

Let us now consider some important examples. First, let A be the unit ball B_p^k of functions in the Sobolev space $H_p^k = H_p^k[0, 1]$ with $\|f^{(k)}\|_p \leq 1$, where $1 \leq p \leq \infty$. Also, let q be the conjugate of p. The following result describes the asymptotic estimates of d_n, d^n and δ_n. We shall use the notation $d_n \approx n^s$ to mean $C_1 n^s \leq d_n \leq C_2 n^s$ for some positive constants C_1 and C_2 independent of n.

Theorem 16. Let $k \geq 2$. Then

$$d_n\left(B_p^k; L_r\right) \approx \begin{cases} n^{-k}, & 1 \leq r \leq p \leq \infty \\ & \text{or} \quad 2 \leq p \leq r \leq \infty \\ n^{-k+1/p-1/2}, & 1 \leq p \leq 2 \leq r \leq \infty \\ n^{-k+1/p-1/r}, & 1 \leq p \leq r \leq 2 \end{cases}$$

$$d^n\left(B_p^k; L_r\right) \approx \begin{cases} n^{-k}, & 1 \leq r \leq p \leq \infty \\ & \text{or} \quad 1 \leq p \leq r \leq 2 \\ n^{-k+1/2-1/r}, & 1 \leq p \leq 2 \leq r \leq \infty \\ n^{-k+1/p-1/r}, & 2 \leq p \leq r \leq \infty \end{cases}$$

and

$$\delta_n\left(B_p^k; L_r\right) \approx \begin{cases} n^{-k}, & 1 \leq r \leq p \leq \infty \\ n^{-k+1/p-1/r} & 1 \leq p \leq r \leq 2 \\ & \text{or} \quad 2 \leq p \leq r \leq \infty \\ n^{-k+1/p-1/2}, & 1 \leq p \leq 2 \leq r \leq \infty \\ & \text{and} \quad q \geq r \\ n^{-k+1/2-1/r}, & 1 \leq p \leq 2 \leq r \leq \infty \\ & \text{and} \quad q \leq r \end{cases}$$

For the discrete case R^m, let B_p denote the unit ball using the $l_p(R^m)$ norm, and set $X = l_\infty(R^m)$. The following estimates have been obtained:

Theorem 17. There exist constants C and C_a independent of m and n such that

$$d_n(B_1; X) \leq C_m^{1/2} \Big/ n \tag{1}$$

$$d_n(B_1; X) \leq 2[(\ln m)/n]^{1/2} \tag{2}$$

$$d_n(B_1; X) \leq C_a n^{-1/2} \qquad \text{if} \qquad n < m < n^a \tag{3}$$

$$d_n(B_1; X) \leq C\left[\left(1 + \ln \frac{n}{m}\right) \Big/ n\right]^{1/2} \tag{4}$$

$$d_n(B_1; X) \leq 8\left(\frac{\ln m}{\ln n}\right) \Big/ n^{1/2} \tag{5}$$

$$d_n(B_1; X) \geq \left(1 + \frac{(m-1)n}{m-n}\right)^{-1/2} \tag{6}$$

$$d_n(B_2; X) \leq C\left(1 + \ln \frac{m}{n}\right)^{3/2} \Big/ n^{1/2} \tag{7}$$

III. BIVARIATE SPLINE FUNCTIONS

Spline functions in one variable have proved to be very rich in theory and extremely useful in applications. However, not much was done in the multivariable setting until the 1980s, and at this writing, many basic questions concerning dimensions, bases, minimum supported elements, interpolation, approximation order, shape-prescribed or

shape-preserving approximation schemes, computational algorithms, and so on are still unanswered. In fact, many of the very fundamental problems do not seem to have satisfactory solutions. This is an area in approximation theory that requires much research effort. In this section we discuss only some selected basic results in the two-variable setting, although most of the results here could be obtained in higher dimensions. A different aspect of the multivariable theory is discussed in Section VIII.

Since spline functions in one variable (or univariate splines) are piecewise polynomials separated by points, spline functions in two variables (of bivariate splines) are piecewise polynomials separated by curves. If the bivariate splines are to be continuous, these curves, which we call grids, are necessarily algebraic. In fact, if the restrictions of s to D_1 and D_2 are polynomials p_1 and p_2, respectively, and D_1, D_2 are separated by an (algebraic) curve C represented by an irreducible polynomial equation $l(x, y) = 0$, then a necessary and sufficient condition that s is in C^k on $D_1 \cup C \cup D_2$ is the existence of a polynomial q_{12} with $p_1 - p_2 = q_{12} l^{k+1}$. Since the factor l^{k+1} determines the smoothing condition of the bivariate spline s, it is called the smoothing factor of s across C, and the polynomial q_{12} is called a smoothing cofactor of s. Note that $q_{21} = -q_{12}$ and q_{12} uniquely determines s on D_2 if s is already known on D_1. If $C_1, \ldots, C_n$ are algebraic curves with a common point of intersection A, called a vertex, and separate the domains $D_1, \ldots, D_n$, and if the restriction of a C^k bivariate spline s on D_j is a polynomial p_j, then the smoothing cofactors of s across $C_1, \ldots, C_n$ must satisfy the following conformality condition.

$$q_{12}(l_1)^{k+1} + \cdots + q_{n-1,n}(l_{n-1})^{k+1} + q_{n,1}(l_n)^{k+1} = 0$$

where $l_j(x, y) = 0$ is an irreducible polynomial equation of C_j. Hence, a bivariate spline s on a region D in R^2 that is partitioned by algebraic curves is uniquely determined by all smoothing cofactors that satisfy the conformality conditions around each interior vertex as long as one polynomial piece is prescribed. This simple observation allows us to study the dimensions of bivariate spline spaces and to construct basis elements, particularly locally supported bivariate splines. However, even the dimension on a very simple grid partition is "unstable" in the sense that a perturbation would change the dimension. Among the "stable" bivariate spline spaces are those with ray (or quasi-cross-cut) partitions. Let D be a simply connected domain in R^2. A ray partition Δ of D is one that each (polynomial) curve C must have at least one end point that lies on the boundary of D. The other end point must either be an interior vertex or also lie on the boundary of D. In the latter case, C is called a cross-cuts of D. Suppose that there are m cross-cuts with (irreducible) degrees $d_1, \ldots, d_m$ and n interior vertices $A_1, \ldots, A_n$, such that the rays through A_i are represented by irreducible polynomials $l_{i,1}, \ldots, l_{i,n_i}$ with degrees $f_{i,1}, \ldots, f_{i,n_i}$, respectively. Then we have the following result on the dimension of the space S_d^k of all C^k bivariate splines on D with degree d and grid partition Δ.

Theorem 18.

$$\dim S_d^k = \binom{d+2}{2} + \sum_{j=1}^{m} \binom{d-(k+1)d_j+2}{2}_+ + \sum_{j=1}^{n} b_j$$

where $b_j = \dim V_j$ and

$$V_j = \left\{ (q_1, \ldots, q_{n_j}) : q_i \in \pi_{d-(k+1)f_{j,i}}, \sum_{i=1}^{n_j} q_i (l_{j,i})^{k+1} = 0 \right\}$$

However, the dimension b_j is difficult to determine in general. In the particular case where all $f_{j,i}$'s are equal to 1, then we have, using the notation $N = n_j$,

$$b_j = \dim V_j = \frac{1}{2}\left[d - k - \left[\frac{k+1}{N-1} \right] \right]_+ \times \left[(N-1)d - (N+1)k + (N-3) + (N-1)\left[\frac{k+1}{N-1} \right] \right]$$

with $[x]$ denoting the integral part of x.

A basis of the above bivariate spline space can be constructed, but there may not be any locally supported functions in general. Much work has been done in the three and four-directional meshes. Let us assume that two of the directions in each case are parallel to the x and y axes, and to be more specific let the horizontal and vertical grids be $x = i$ and $y = i$ with $i \in Z$. Then a three-directional mesh is a unidiagonal triangulation and a four-directional mesh is a crisscross triangulation of a uniform rectangular partition. They are also called type I and type II triangulations, respectively. Even in these special cases many interesting and unexpected results have been obtained. For instance, minimum-supported splines may be linearly dependent, some functions in S_d^k cannot be locally reproduced, and even the optimal approximation orders are somewhat unexpected.

We discuss only the three-directional mesh here. The approximation order of S_d^k is an integer m such that dist $(f, S_h) = O(h^m)$ for all sufficiently smooth functions f and dist$(g, S_h) \neq o(h^m)$ for some C^∞ function g. Here, S_h is simply the space of all $s_h(x, y) = s(x/h, y/h)$, where

$s \in S_d^k$. Let $m(d) = \min\{2(d-K), d+1\}$. It is known that, if $k \le (2d-2)/3$, then $m(d) - 2 \le m \le m(d)$. For instance, for S_3^1, $m(d) = 4$ and m is known to be 3. It is also known that if $k > (2d-2)/3$, then $m = 0$. These results were proved using box splines, which are discussed in Section VIII.

While the exact value of m is still unknown even for the three-directional mesh that we discuss here, the controlled approximation order with respect to box splines has been determined. Let $\{B^1, \ldots, B^N\}$ be the collection of all box splines in S_d^k. Then the controlled approximation order of S_h with respect to box splines is the largest integer n such that

$$\left\| f - \sum_{i=1}^{N} \sum_{z \in Z^2} w_i^h(j) B^i\left(\frac{\cdot}{h} - j\right) \right\|_\infty \le C h^n \| D^n f \|_\infty$$

for some sequence $w_i^h(j)$ satisfying

$$\left\| w_i^h(\cdot) \right\|_{l_\infty} \le C \| f \|_\infty, \qquad i = 1, \ldots, N$$

where C is an absolute constant, and f and C^∞ function. In Section VIII, it will be seen that ∞ can be replaced by any p, $1 \le p \le \infty$, even when the controlled condition on the weights is replaced by a "local" one.

Theorem 19. Let S_h be the scaled three-directional mesh of S_d^k. Then the controlled approximation order of S_h with respect to all box splines of S_d^k is

$$2d - 2k \quad \text{for} \quad 2d - 3k = 2 \tag{1}$$

$$2d - 2k - 1 \quad \text{for} \quad 2d - 3k = 3 \quad \text{or} \quad 4 \tag{2}$$

$$d + 1 \quad \text{for} \quad k = 0 \tag{3}$$

$$\min\{2d - 2k - 2, d\} \quad \text{for} \quad k \ge 1 \quad \text{and} \quad 2d - 3k \ge 5 \tag{4}$$

However, the controlled approximation order of S_h with respect to all minimum-supported splines is still unknown.

When the rectangular partitions are nonuniform, the unidiagonal and crisscross triangulations are no longer three- and four-directional meshes. The dimensions of these spaces are still unknown except for the cases when $d - k$ is sufficiently large or when $k = 1, 2$. It is also known that, while the support of the S_2^1 minimum-supported bivariate splines on the crisscross triangulation is independent of the uniformity of the rectangular partition, the corresponding statement for S_3^1 on the unidiagonal triangulation is false. In fact, even the support of the box splines, which are larger in this case, is not preserved under nonuniform perturbation of the rectangular partition, and if one minimum-supported spline in the perturbed case here is used, the totality of all its "translates" does not necessarily produce constants. There is still no general result in the nonuniform setting.

When a triangular grid partition is considered, it is usually more convenient to use Bézier (or Bernstein) representations of the polynomial pieces. Smoothing conditions on the adjacent polynomial pieces are expressed in terms of certain relations on the Bézier coefficients. Many interesting formulas have been recently obtained. These formulas have applications to constructing Hermite interpolants and quasi interpolants to scattered data and also have nice applications to computer-aided geometric designs.

IV. COMPACT OPERATORS AND *M* IDEALS

Since the mid-1970s, a branch of approximation theory called operator approximation has come into vogue. This area is concerned with the approximation of an operator from a family of operators with some nice structures, namely, positive operators, self-adjoint operators, normal operators, compact operators, or operators with more than one of these properties. Since the majority of the research papers in this area have been concerned with compact operator approximation, we limit ourselves to the discussion of this topic. Let $B(X)$ denote the class of bounded linear operators on a Banach space X, and $C(X)$ the subcollection of compact operators on X. The general question of interest is the existence of a best approximant to a given operator T in $B(X)$ from $C(X)$. If every T in $B(X)$ has at least one best approximant from $C(X)$, we say that $C(X)$ is proximal in $B(X)$. It is well known that, if X is a Hilbert space, then $C(X)$ is proximinal in $B(X)$. The following result on Banach spaces is worth stating:

Theorem 20. Let $1 < p < \infty$. Then $C(l_p)$ is proximinal in $B(l_p)$. However, $C(X)$ is not proximinal in $B(X)$ for $X = C[0, 1]$ or $X = L_p[0, 1]$, $1 < p < \infty$, $p \ne 2$.

Nevertheless, certain bounded linear operators in $B(L_p)$ may still have best compact approximants. Let $B_1(L_p)$ be the collection of T in $B(L_p)$ such that $\|Tx_n\|_p \to 0$ for every uniformly bounded weakly null sequence $\{x_n\}$ in L_p. Also, let $B_2(L_p)$ be the integral operators in $B_1(L_p)$. The following result holds:

Theorem 21. $C(L_p)$ is proximinal in $B_1(L_p)$ for $2 < p < \infty$ and is proximinal in $B_2(L_p)$ for $1 < p < \infty$.

If H is a Hilbert space, every Hankel operator in $B(H)$ has a best compact Hankel approximant. Using some function theoretic arguments, the following result can then be obtained. Here, H^∞ will denote the Hardy space of bounded analytic functions and C the space of continuous functions.

Theorem 22. $H^\infty + C$ is a proximinal subspace in L^∞.

It should be remarked that, although this theorem does not seem to be a result in operator approximation, it is indeed related to best approximation by compact Hankel operators. For the l_p spaces, if we let P_n be the norm-one projection of l_p onto the first n coordinate vectors, the following distance formula is obtained:

Theorem 23. Let T be in $B(l_p)$, where $1 < p < \infty$. Then $\|P_n^\perp T P_n^\perp\| \to \text{dist}[T, C(l_p]$ as n tends to infinity.

For $p = 2$, P_n could be chosen as a norm-one projection onto the first n vectors of any orthonormal basis. It is important to remark, however, that nothing has been mentioned about uniqueness. In fact, if T is a noncompact bounded linear operator in l_p, where $1 < p < \infty$, T has "many" compact best approximants. This observation follows from the fact that $C(l_p)$ is an M ideal in $B(l_p)$.

The notion of an M ideal in a Banach space was introduced in the early 1970s. A closed subspace Y of a Banach space X is an M ideal in X if its annihilator $Y^\perp$ is an L summand of the dual space X^*. This in turn means that $Y^\perp$ is the range of an L projection defined on X^*, that is, a projection $Q : X^* \to Y^\perp$ with the property that $\|f\| = \|Qf\| + \|f + Qf\|$ for every f in X^*. The importance of M ideals in approximation theory is that M ideals are proximinal subspaces with certain special approximation properties. For instance, (1) the metric projection P_Y onto Y satisfies the Lipschitz condition $d_H[P_Y(x), P_Y(y)] \le 2\|x - y\|$ for all x, y in X, where d_H denotes the Hausdorff distance, and (2) there exists a continuous homogeneous selection for the metric projection P_Y. Perhaps the most remarkable approximation characteristic of M ideals is the following result:

Theorem 24. Let Y be an M ideal in a Banach space X and x be in $X \backslash Y$. Then $P_Y(x)$ algebraically spans Y.

We have already seen in Theorem 20 that $C(l_p)$ is proximinal in $B(l_p)$, where $1 < p < \infty$. It is also known that $C(l_p)$ is an M ideal in $B(l_p)$, $1 < p < \infty$. In fact, $C(l_p)$ is the only M ideal in $B(l_p)$.

Theorem 25. Let Z be a subspace of l_p, where $1 < p < \infty$. Then the compact operators on Z form an M ideal in the space of bounded operators on Z if and only if Z has the compact approximation property.

In the same direction as Theorem 22, the following result can be proved:

Theorem 26. The compact Hankel operators form an M ideal in the space of Hankel operators on a Hilbert space.

M Ideals have other important structures. For instance, an M ideal of a Banach algebra must be an algebra, and if Ω is a compact Hausdorff space and A a function algebra contained in $C(\Omega)$, then the M ideals of A are precisely the closed ideals with a bounded approximate identity. If, in particular, A is the disc algebra, then the M ideals of A are exactly the two-sided ideals with norm-one approximate identity. In addition, the following interpolation result has been obtained:

Theorem 27. Let f be in the disc algebra A and for each $n = 1, 2, \ldots, E_n$ be a closed subset of the unit circle with linear measure zero such that $\gamma_n = \log(\|f\| T' \|f\|_{E_n}) \to 0$. Then there exist minimal norm interpolants s_n of f satisfying $(f - s_n) \in Y_n = \{g \in A : g(E_n) = 0\}$ such that $\|f - s_n\| = O(\gamma_n)$. Furthermore, the rate of convergence $O(\gamma_n)$ cannot be replaced by $o(\gamma_n)$ in general.

V. CONSTRAINED APPROXIMATIONS

In many approximation and interpolation problems, the approximants, which may also be interpolants, are required to satisfy certain constraints. The constraints may be explicit conditions imposed on the approximation problems, or they may be certain specific properties of the mathematical model or data the approximants are supposed to preserve. In general, constrained approximation problems are nonlinear, and some important problems do not even have analytic solutions. We mention a few such problems. The examples we have chosen should not be construed as the most important work in this area but should be thought of as illustrating the results of the subject.

We first limit ourselves to approximation by polynomials and splines in one variable. For polynomial approximation in the uniform (or supremum) norm $\| \|_\infty$, the following result is known:

Theorem 28. Let k be any nonnegative integer. Then there exists a constant C such that for any function f in $C^k[0, 1]$ satisfying $f' \ge 0$ and any integer $n \ge k + 1$

$$\inf\{\|f - p\|_\infty : p' \ge 0,\ p \text{ a polynomial of degree} \le n\} \le C n^{-k} w\left(f^{(k)}, n^{-1}\right)_\infty$$

where $w(\cdot, n^{-1})_\infty$ denotes, as usual, the modulus of continuity in the uniform norm.

This result essentially says that monotone approximation by polynomials retains the same order of approximation as unconstrained approximation. For spline approximation, an analogous result has been obtained, and this

result holds even in the L_p setting. Let $S^*_{k,N}$ denote the space of all kth-order splines s with knots at i/N, $i = 0, \ldots, N$, such that $s' \geq 0$. We have the following result:

Theorem 29. Let $1 \leq p \leq \infty$ and k be a nonnegative integer. Then there exists a positive constant C such that for any monotonically nondecreasing function f with $f^{(j)} \in L_p[0, 1]$, if $1 \leq p > \infty$, or $f^{(j)} \in C[0, 1]$, if $p = \infty$, where $0 \leq j \leq k-1$, and for any $N = 1, 2, \ldots$

$$\inf\{\|f - s\|_p : s \in S^*_{k,N}\} \leq CN^{-j} w\big(f^{(j)}, N^{-1}\big)_p$$

We now turn to the interpolation problem but concentrate only on spline interpolation. Let $0 = t_0 \leq t_1 \leq \cdots t_n \leq t_{n+1} = 1$. Defining t_j for $j < 0$ and $j > n+1$ arbitrarily as long as $t_j < t_{j+k}$ for all j, we can use $\{t_j\}$ as the knot sequence of the normalized B splines $N_{j,k}$ of order k. Let $H^k_p = H^k_p[0, 1]$ denote, as usual, the Sobolev space of functions f in $C^{k-1}[0, 1]$ with $f^{(k-1)}$ absolutely continuous and $f^{(k)}$ in $L_p[0, 1]$. The "optimal" interpolation problem can be posed as follows. Let $\{g_i\}$, $i = 1, \ldots, n$, and

$$\big\|s^{(k)}\big\|_p = \inf\big\{\big\|f^{(k)}\big\|_p : f \in H^k_p,\ f(t_i) = g_i,\ i = 1, \ldots, n\big\}$$

if $1 < p < \infty$, s always exists, and if $1 < p < \infty$ and $n \geq k$, s is also unique. It is also known that, for $1 < p < \infty$ and $n \geq k$,

$$s^{(k)} = |h|^{q-1} \operatorname{sgn} h$$

where h is some linear combination of the normalized B splines $N_{i,k}$ and q the conjugate of the index p. Suppose that the data $\{g_i\}$ are taken from some function g in $C^k[0, 1]$. Then the constrained version of the above problem is determining an s in H^k_p such that $s(t_i) = g(t_i)$ for $i = 1, \ldots, n$, $s^{(k)} \geq 0$, and

$$\big\|s^{(k)}\big\|_p = \inf\big\{\big\|f^{(k)}\big\|_p : f \in H^k_p,\ f(t_i) = g(t_i),\ f^{(k)} \geq 0\big\}$$

Again for $1 < p < \infty$ and $n \geq k$, s exists and is unique. In fact, the following result is known. To formulate the result, we set

$$d_i = [t_i, \ldots, t_{i+k}]g$$

where the divided difference notation has been used. Also, let χ_A denote the characteristic function of a set A.

Theorem 30. The unique function s described above is characterized by

$$\int_0^1 s^{(k)} N_{i,k} = d_i$$

for $i = 1, \ldots, n-k$, and

$$s^{(k)} = \left(\sum_{j=1}^{n-k} \alpha_j N_{j,k}\right)_+^{q-1} \chi_A$$

where q is the conjugate of p and

$$A = [0, 1] \Big\backslash \bigcup_{j=1}^{n-k} \{(t_j, t_{j+k}) : d_j = 0\}$$

We observe that if $p = 2$, so that $q - 1 = 1$, then s is in C^{k-1} and is a piecewise polynomial of order $2k$. Recently, an analog to the perfect spline solution of the unconstrained problem has been obtained for case $p = \infty$. However, with the exception of some simple cases, no numerical algorithm to determine s is known.

To describe computational methods for shape-prescribed splines, it is easier to discuss quadratic spline interpolation. It should be clear from the above result that extra knots are necessary. Let these knots be $x_1, \ldots, x_{n-1}$ with $t_i < x_i < t_{i+1}$ and let $\{y_i, m_i\}$, $i = 1, \ldots, n$, be a given Hermite data set. Then a quadratic spline s with knots at $\{t_1, \ldots, t_n, x_1, \ldots, x_{n-1}\}$ is uniquely determined by the interpolation conditions $s(t_i) = y_i$ and $s'(t_i) = m_i$ for $i = 1, \ldots, m$. The slopes of s at the knots x_i can be shown to be

$$s'(x_i) = \frac{2(y_{i+1} - y_i) - (x_i - t_i)m_i - (t_{i+1} - x_i)m_{i+1}}{t_{i+1} - t_i}$$

for $i = 1, \ldots, n-1$. Hence, the knot x_i is not active if and only if

$$\frac{m_{i+1} + m_i}{2} = \frac{y_{i+1} - y_i}{t_{i+1} - t_i}$$

In general, the slopes m_i are not given and can be considered parameters to ensure certain shape. For instance, if $m_i m_{i+1} \geq 0$, a necessary and sufficient condition that $m_i s'_{(t)} \geq 0$ for $t_i \leq t \leq t_{i+1}$ is $|(x_i - t_i)m_i + (t_{i+1} - x_i)m_{i+1}| \leq 2\,|y_{i+1} - y_i|$.

Similar conditions can be obtained for comonotonicity and convexity preservation. For practical purposes, the slopes m_i could be selected as a weighted central difference formula for $2 \leq i \leq n-1$ and a weighted noncentral formula for $i = 1$ and n, so that the artificial knots x_i can be determined using the corresponding necessary and sufficient condition such as the inequality mentioned above. Such an algorithm has the advantage of being one-pass.

Cubic splines could be easily used for monotone approximation. In fact, if $n \geq 2$, there is a uique function s such that s'' satisfies $s' \geq 0$, $s(t_i) = g_i$, and

$$\|s''\|_2 = \inf\big\{\|f''\|_2 : f \in H^2_2,\ f(t_i) = g_i,\ f' \geq 0\big\}$$

The unique solution s is a natural cubic spline with at most $2[n/2] + 2$ extra knots in addition to the knots at $t_i, \ldots, t_n$.

Much more can be said when the Hilbert space H^k_2 is considered. For instance, let C be a closed convex subset of H^k_2. This may be a set of functions that are positive, monotone, convex, and so on. The following result has been obtained:

Theorem 31. Let $g \in C \subset H_2^k$ and $n \geq k+1$. Then there is a unique function s_n in C such that $s_n(t_i) = g(t_i)$ for $i = 1, \ldots, n$, and

$$\left\| s^{(k)} \right\|_2 = \inf\left\{ \left\| f^{(k)} \right\|_2 : f \in H_2^k,\ f(t_i) = g(t_i),\ f \in C \right\}$$

Furthermore, s_n is a piecewise polynomial of order $2k$, and if s_n^* denotes the (unconstrained) natural spline of order $2k$ interpolating the same data, then

$$\left\| (s_n - g)^{(k)} \right\|_2 \leq \left\| \left(s_n^* - g\right)^{(k)} \right\|_2$$

Consequently, the rate of convergence of the constrained interpolants s_n to g is established. Another consequence is that the convergence rate for splines interpolating derivatives at the end points remains the same as the unconstrained ones.

We now give a brief account of the multivariable setting. Minimizing the L_2 norm over the whole space R^s of certain partial derivatives gives the so-called thin-plate splines. So far, the only attempt to preserve shape has been preserving the positivity of the data. Since the notation has to be quite involved, we do not go into details but remark that the convergence rate does not differ from the unconstrained thin-plate spline interpolation. It must be mentioned, however, that thin-plate splines are not piecewise polynomials. Piecewise polynomials satisfying certain smoothness joining conditions are discussed in the sections of bivariate splines for two variables and multivariate polyhedral splines for the general n-dimensional setting. Results on monotone and convex approximation by piecewise polynomials, however, are not quite complete and certainly not yet published.

VI. FACTORIZATION OF BIINFINITE MATRICES

There has been much interest in biinfinite matrices. We will discuss this topic only from the view of an approximation theorist, although biinfinite matrices have also important applications to signal processing, time series, and so on. The importance of biinfinite matrices in approximation theory can best be explained via the following (polynomial) spline interpolation problem.

Let $\mathbf{t}: \cdots < t_i < t_{i+1} < \cdots$ be a knot sequence and $\mathbf{x}: \cdots < x_i < x_{i+1} < \cdots$ a sequence of sample points. If a certain sequence of data $\{(x_i, y_i)\}$ is given where $\{y_i\} \in l_\infty$, we would like to know the existence and uniqueness of a kth-order spline s with knots at $\mathbf{t}$ such that $s(x_i) = y_i$ for all i and $s \in L_\infty$. If we write s as a linear combination of the normalized B splines $N_j = N_{j,k,t}$, the interpolation problem can be expressed as a biinfinite linear system $A\alpha = \mathbf{y}$ where $\mathbf{y} = [\cdots y_i, y_{i+1}, \ldots]^T$ and $A = [a_{ij}]$, with $a_{ij} = N_j(x_i)$. Hence, the biinfinite (coefficient) matrix A can be considered an operator from l_∞ into itself. If A is surjective, then a bounded spline interpolant s exists, and if A is injective, then s is uniquely determined by the interpolation condition $s(x_i) = y + i$, where $i = \ldots 0, 1, \ldots$. It is clear that A cannot be invertible unless it is banded, and using the properties of B splines, it can be shown that A is totally positive.

Generally, certain biinfinite matrices $A = [a_{ij}]$ can be considered as operators on l_p as follows. Let $A\{x_i\} = \{Ax_i\}$, with $Ax_i = \sum_j a_{ij}x_j$, where $\{x_j\}$ is a finitely supported sequence. Then the definiton can be extended to all of l_p using the density of finitely supported sequences. If the extension is unique, we denote the operator also by A and use the usual operator norm $\|A\|_p$. Let $\{\mathbf{e}_j\}$ be the natural basis sequence with $e_j(i) = \delta_{ij}$, and denote by P_n the projection defined by $P_n\mathbf{e}_j = \mathbf{e}_j$ for $|j| \leq n$ and zero otherwise. Therefore, the biinfinite matrix representation of P_n is $[p_{ij}]$, where $p_{ij} = \delta_{ij}$ for $|j| \leq n$ and zero otherwise. For any biinfinite matrix A, we can consider its truncation $A_n = P_nAP_n$. Also, let S be the shift operator defined by $S\mathbf{e}_j = \mathbf{e}_{j+1}$. Then $(S^rA)_n(i, j) = 0$ if $|i| > n$ or $|j| > n$ and $= a_{i-rj}$ otherwise. That is, S^r shifts the diagonals of A down by r units.

We use the notation $A \in B(l_p)$ if A, considered an operator on l_p, is a bounded operator. We also say that A is boundedly invertible if, as an operator on l_p, A is both injective and surjective. In order to extend the important results in finite matrices, it is important to be able to identifya "main diagonal" of A. Perhaps the following is a good definition of such diagonal. The rth diagonal of A is main if

$$\limsup \left\| (S^rA)_n^{-1} \right\|_p < \infty$$

Clearly, the zeroth (or central) diagonal of the identity operator I on l_p is main, and in fact if both $I + K$ and its inverse are in $B(l_p)$, the zeroth diagonal is also main. Another example is that, if $\|K\|_p < 1$, then the zeroth diagonal of $I + K$ is a main diagonal. For the Hilbert space l_2, in particular, it can be verified that, if A is a positive definite symmetric matrix such that both A and A^{-1} are in $B(l_2)$, then the zeroth diagonal of A is main. For l_1, if A is column diagonally dominant such that A and A^{-1} are in $B(l_1)$, then again the zeroth diagonal of A is a main diagonal.

The following result is important in spline interpolation:

Theorem 32. Any totally positive biinfinite matrix A such that both A and A^{-1} are in $B(l_\infty)$ has a unique main diagonal.

This theorem allows us to conclude that A^{-1} is checkerboard with the main diagonal containing only positive entries, so that a bound on the local mesh ratio in spline

interpolation can be obtained in terms of the norm of A^{-1}. We remark that there are many other possible and plausible definitions for a main diagonal of a biinfinite matrix.

In solving a biinfinite linear system as discussed earlier, a Gauss elimination procedure quite often factors the coefficient matrix A as $A = LU$, where L is a unit lower triangular matrix with the zeroth diagonal as its rightmost (nontrivial) band, and U is an upper triangular matrix. This is called an LU factorization of A. We say that an LU factorization is invertible if each factor is bounded and boundedly invertible, with L^{-1} and U^{-1} being again lower and upper triangular, respectively.

Theorem 33. Let A be totally positive with both A and A^{-1} in $B(l_\infty)$. Then there exists a unique r such that $S^r A = LU$, where L, L^{-1}, U, U^{-1} are in $B(l_\infty)$, L, L^{-1} being unit lower triangular and U, U^{-1} being upper triangular biinfinite matrices. Furthermore, if we write

$$(S^r A)_n = L_n U_n$$

for each n, then $L_n \to L$ and $U_n \to U$ entrywise.

It is clear that the LU factorization above is unique. However, it would be even better if U were the transpose of L. Such a factorization is called a Cholesky decomposition. We have the following result:

Theorem 34. Let A be a positive definite symmetric matrix with A and A^{-1} in $B(l_2)$. Then A has a unique Cholesky decomposition; that is, $A = LL^T$, where L and L^{-1} are lower triangular biinfinite matrices in $B(l_2)$ with $l_{ii} > 0$. Furthermore, L is unique, and writing $A_n = L_n L_n^T$, we have $L_n \to L$ and $n \to \infty$.

The following result for l_1 is also of some interest.

Theorem 35. Let A be a column diagonally dominant biinfinite matrix with A and A^{-1} in $B(l_1)$. Then there is a unique factorization $A = LU$, where L, L^{-1} are $B(l^1)$ unit lower triangular, and U, U^{-1} are $B(l^1)$ upper triangular. Furthermore, writing $A_n = L_n U_n$, we have $L_n x \to Lx$ and $U_n x \to Ux$ for all $x \in l_1$.

Certain biinfinite matrices have more refined factorizations. We will call a one-banded biinfinite matrix $R = [r_{ij}]$ elementary if $r_{ii} = 1$ and $r_{ij} = 0$ for all i and j with $j \le i - 2$ or $j > i$.

Theorem 36. Any strictly m-banded totally positive unit lower triangular biinfinite matrix A has a factorization $A = R_1 \cdots R_m$ where each $R_i, i = 1, \ldots, m$, is elementary.

Hence, if A is not necessarily unit but is m-banded totally positive and lower triangular, then $A = R_1 \cdots R_m B$, where B is a diagonal matrix with positive diagonal entries.

We conclude this section by stating a factorization result of block Toeplitz matrices.

Theorem 37. Let A be totally positive biinfinite block Toeplitz, with block size at least 2, and unit lower triangular. Write $A = [A_{ij}]$, where $A_{ij} = A_k$ for all i and j with $i = j + k$, and denote

$$A(z) = \sum_{k=0}^{\infty} A_k z^k$$

Then

$$A(z) = \prod_{k=1}^{\infty} [I + a_k(z)] c(z) \left(\prod_{k=1}^{\infty} [I - b_k(z)] \right)^{-1}$$

where $I + a_k(z)$ and $I - b_k(z)$ are the symbols of one-banded block Toeplitz matrices with $a_k(1), b_k(1) \ge 0$ and $\sum [a_k(1) + b_k(1)] < \infty$. Furthermore, $c(z)$ is the symbol of a totally positive block Toeplitz matrix with $c(z)$ and $c^{-1}(z)$ entire, and $\det c(z) = 1$.

VII. INTERPOLATION

The theory and methods of interpolation play a central role in approximation theory and numerical analysis. This branch of approximation theory was initiated by Newton and Lagrange. Lagrange interpolation polynomials, formulated as early as 1775, are still used in many applications. To be more explicit, let

$$\Delta_n : -1 \le t_{n-1} < \cdots t_{nn} \le 1$$

be a set of interpolation nodes (or sample points) and

$$l_{nk}(t) = \prod_{i \ne k} (t - t_{ni}) / (t_{nk} - t_{ni})$$

Then for an given set of data $Y = \{y_i, \ldots, y_n\}$, the polynomial

$$p_n(t) = p_n(t; Y, \Delta_n) = \sum_{k=1}^{n} y_k l_{nk}(t)$$

interpolates the data Y on Δ_n. It is interesting that, even for uniformly spaced Δ_n, there exists a continuous function g on $[-1, 1]$ such that the (Lagrange) interpolating polynomials $p_n(t)$ to $y_i = g(t_{ni})$, $i = 1, \ldots, n$, do not converge uniformly to $g(t)$ on $[-1, 1]$. This leads to the study of the Lebesgue function and constant defined, respectively, by

$$l_n(t) = l_n(t; \Delta_n) = \sum_{k=1}^{n} |l_{nk}(t)|$$

and

$$l_n = l_n(\Delta_n) = \max\{l_n(t) : -1 \le t \le 1\}$$

The following result indicates the divergence property of $l_n(t)$:

Theorem 38. There exists a positive constant C such that, for any sequence of positive numbers e_n and any Δ_n, there is a set $H_n(e_n, \Delta_n)$ of measure no greater than e_n with

$$l_n(t) > Ce_n \ln n$$

for $t \in [-1, 1] \backslash H_n(e_n, \Delta_n)$ and all $n = 1, 2, \ldots$.

Hence, $l_n = O(\ln n)$ is the best we can expect for various choices of Δ_n. A natural (and almost best) choice of Δ_n is the set

$$T_n = \left\{ \cos \frac{(2k-1)\pi}{2n} : k = 1, \ldots, n \right\}$$

of roots of the nth-degree Chebyshev polynomial of the first kind. In this case, the asymptotic expression of the Lebesgue constant is given in the following expression:

Theorem 39.

$$l_n(T_n) - \frac{2}{\pi}\left(\ln n + \gamma + \ln \frac{8}{\pi} \right) \sim \frac{8}{\pi} \sum_{s=1}^{\infty} (-1)^{s-1} \times (2^{2s-1} - 1)^2 \pi^{2s} B_{2s}^2 \big/ (2s)!(2s)(2n)^{2s}$$

where γ is the Euler constant and B_{2s} are the Bernoulli numbers.

The best choice of Δ_n could be defined by Δ_n^*, where

$$l_n\left(\Delta_n^*\right) = \inf\{l_n(\Delta_n) : \Delta_n\}$$

and we set $l_n^* = l_n(\Delta_n^*)$. It is known that Δ_n^* exists and is unique. In fact, it is characterized by the localization condition

$$\max\{l_n(t; \Delta_n) : t_{n,i} \le t \le t_{n,i+1}\} = l_n(\Delta_n)$$

for $i = 1, \ldots, n-1$. Let us denote the above localized maximum by $l_{ni}(\Delta_n)$. Then the (modified) roots of the nth-degree Chebyshev polynomials are known to satisfy

$$\max_i l_{ni}(\hat{T}_n) - \min_i l_{ni}(\hat{T}_n) < 0.0196$$

for all $n \ge 70$, where the modification is just a linear scaling, mapping the largest and smallest roots to the end points of $[-1, 1]$; namely,

$$T_n = \left(\cos \frac{\pi}{2n} \right) \hat{T}_n$$

This is the reason that modified roots of the Chebyshev polynomials are believed to be an "almost best" choice of Δ_n.

For each choice of Δ_n, there is some continuous function f such that the corresponding Lagrange polynomial interpolation to f does not even converge in L_p. Some extra interpolating conditions are required to increase the chance of convergence. We first discuss interpolation at the roots $x_{nk} = x_{nk}(w)$ of the (generalized) Jacobi orthogonal polynomials corresponding to a "smooth" weight w. Let $x_{no} = -1$, $x_{n,n+1} = 1$, and

$$L_n = L_n^{(r,s)}(\cdot, f) = L_n^{(r,s)}(w, f, \{x_{nk}\})$$

denote the interpolating polynomial of f satisfying the interpolating conditions

$$L_n(x_{nk}) = f(x_{nk}), \qquad k = 0, \ldots, n+1$$
$$D^l L_n(1) = 0, \qquad l = 1, \ldots, r-1$$

and

$$D^l L_n(-1) = 0, \qquad l = 1, \ldots, s-1$$

Theorem 40. Let u be a nonnegative function not identically equal to zero on $[-1, 1]$ and be in $(L \log^+ L)_p$, where $0 < p < \infty$. Then

$$\lim_{n\to\infty} \left\| \left[L_n^{(r,s)}(\cdot, f) - f(\cdot) \right] u(\cdot) \right\|_p = 0$$

for all f in $C[-1, 1]$ if and only if both

$$(1-x)^{-r+1/4}(1+x)^{-s+1/4}[w(x)]^{1/2}$$

is in L_1 and

$$u(x)(1-x)^{r-1/4}(1+x)^{s-1/4}[w(x)]^{-1/2}$$

is in L_p. Furthermore, the condition that u is in $(L \log^+ L)_p$ cannot be replaced by $u \in L_p$.

It should be mentioned, however, that to every continuous function f there exists a sequence of interpolation nodes Δ_n such that the Lagrange interpolation polynomials of f at Δ_n converge uniformly to f on $[-1, 1]$. Fejér arranged the interpolation nodes for Hermite interpolation and gave an easy proof of the Weierstrass theorem. His polynomials, which are usually called Hermite–Fejér interpolation polynomials, are defined by the interpolation condition $H_n(x_{nk}, f) = f(x_{nk})$ and $DH_n(x_{nk}, f) = 0$ for $k = 1, \ldots, n$, where $\{x_{nk}\} = T_n$ are the roots of the nth-degree Chebyshev polynomial. Later, other choices of Δ_n were also studied, and results on approximation orders, asymptotic expansions, and so on were obtained. For instance, the following result is of some interest:

Theorem 41. Let C_n denote the nth-degree Chebyshev polynomial of the first kind and $H_n(\cdot, f)$

denote the Hermite-Fejér interpolation polynomial of a continuous function f at the roots T_n of C_n. Then

$$|H_n(x, f) - f(x)| = O(1)\left\{ \sum_{i=1}^{n} \frac{1}{i^2} w\left[f, \frac{i}{n}(1-x^2)^{1/2} \times |C_n(x)| \right] + \sum_{i=1}^{n} \frac{1}{i^2} w\left[f, \left(\frac{i}{n}\right)^2 |xC_n(x)| \right] \right\}$$

and

$$|H_n(x, f) - f(x)| = O(1)\left\{ \frac{1}{n} \sum_{i=1}^{n} w\left[f, \frac{1}{i}(1-x^2)^{1/2} \times |C_n(x)| + i^{-2}|xC_n(x)| \right] \right\}$$

There has been much interest in the so-called Birkhoff polynomial interpolation problem. The interpolation conditions here are governed by an incidence matrix $E = [e_{ij}]$, where $e_{ij} = 1$ or 0, namely,

$$D^j p(x_i) = y_{ij} \qquad \text{if} \qquad e_{ij} = 1$$

We now study interpolation by spline functions. Spline interpolation is much more useful than polynomial interpolation. In the first place there is no need to worry about convergence. Another great advantage is that a spline interpolation curve does not unnecessarily oscillate as a polynomial interpolation curve does when a large number of nodes are considered. The most commonly used is a cubic spline. If interpolation data are given at the nodes $\Delta_m : 0 = x_0 < \cdots < x_m = 1$, the interior nodes can be used as knots of the cubic spline. When the first derivatives are prescribed at the end points, the corresponding interpolation cubic spline, called a complete cubic spline, is uniquely determined. If no derivative values at the end points are given, a natural cubic spline with zero second derivatives at the end points could be used. Spline functions can also adjust to certain shapes. This topic is discussed in Section V on constrained approximation. There is a vast amount of literature on various results on spline interpolations: existence, uniqueness, error estimates, asymptotics, variable knots, and so on. We shall discuss only the following general interpolation problem:

Let $E = [e_{ij}]$ and $F = [f_{ij}]$ be $m \times n$ incidence matrices and Δ_n be a partition of [0, 1] as defined above. We denote by $S_n(F, \Delta_m)$ the space of all piecewise polynomial functions f of degree n with the partition points x_i of Δ_m as break points such that $D^{n-j} f(x_i^-) = D^{n-j} f(x_i^+)$ for all pairs (i, j) with $f_{ij} = 0$, where $0 < i < m$. The following problem is called an interpolation problem, I.P.(E, F, Δ_m):

Given a set of data $\{y_{ij} : e_{ij} = 1\}$, find an f in $S_n(F, \Delta_m)$ satisfying $D^j f(x_i) = y_{ij}$, where $e_{ij} = 1$ and $D^j f(x_i) = \frac{1}{2}[D^j f(x_i^-) + D^j f(x_i^+)]$.

The problem I.P.(E, F, Δ_m) posed here is said to be poised if it has a unique solution for any given data set. It is known that I.P.(E, F, Δ_m) is poised if and only if I.P.(E, F, Δ_m) is poised. By using the Pólya conditions on (E, F) and the Budan–Fourier theorem, the following result can be shown. We say that E is quasi Hermite if for each $i = 1, \ldots, m-1$, there exists an M_i such that $e_{ij} = 1$ if and only if $j < M_i$.

Theorem 42. Suppose that (E, F) satisfies the Pólya conditions and the $e_{ij} = 1$ implies $f_{i,n-j} = 0$ for all (i, j). Suppose further that one of the matrices E and F is quasi Hermite and the other has no supported odd blocks. Then the problem I.P.(E, F, Δ_m) is poised for any Δ_m.

For cardinal interpolation, we impose the extra conditions $e_{0j} = e_{mj}$ and $f_{0j} = f_{mj}$ for $j = 0, \ldots, m$ on the incidence matrices. Of course, $S_n(F, \Delta_m)$ has to be extended to be piecewise polynomial functions f of degree n with break points at $q + x_1$, where $i = 1, \ldots, m$ and $q = \cdots 0, 1, \ldots$, such that

$$D^{n-j} f\left(q + x_i^-\right) = D^{n-j} f\left(q + x_i^+\right)$$

for all q and (i, j) with $f_{ij} = 0$. The following problem will be called a cardinal interpolation problem, C.I.P.(E, F, Δ_m):

Given a set of data $\{y_{ijq}\}$, for any $0 \le i < m$ and $e_{ij} = 1$, find an f in $S_n(F, \Delta_m)$ satisfying $D^j f(q + x_i) = y_{ijq}$, where $D^j f(q + x_i)$ is the average of $D^j f(q\ x_i^-)$ and $D^j f(q + x_i^+)$.

The problem C.I.P.(E, F, Δ_m) is said to be poised if whenever y_{ijq} is arbitrarily given and satisfies $y_{ijq} = O(|q|^p)$ for some p, as $q \to \pm\infty$, for all (i, j), the problem C.I.P.(E, F, Δ_m) has a unique solution f satisfying $f(x) = O(|x|^p)$ as $x \to \pm\infty$. Let C_o denote the space of all solutions of C.I.P.(E, F, Δ_m) with $y_{ijq} = 0$ and denote its dimension by d. A function f in C_o is called an eigenspline with eigenvalue λ if $f(x+1) = \lambda f(x)$ for all x. The following results have been obtained:

Theorem 43. The C.I.P.(E, F, Δ_m) has eigenvalue Λ if and only if the C.I.P.(E, F, Δ_m) has eigenvalue λ^{-1}.

Theorem 44. Let the C.I.P.(E, F, Δ_m) have d distinct eigenvalues. Then the problem is poised if and only if none of the eigenvalues lies on the unit circle.

Error estimates on interpolation by these splines have also been obtained.

We next discuss multivariate interpolation. First it should be mentioned that not every choice of nodes in R^s,

$s \geq 2$, admits a unique Lagrange interpolation. Hence, it is important to classify such nodes. We will use the notation

$$N_n(s) = \binom{n+s}{s}$$

The following result gives such a criterion, which can be used inductively to give an admissible choice of nodes in any dimension:

Theorem 45. Let $\Delta_n = \{x^i\}_{1,\ldots,N_n(s)}$ be a set of nodes in R^s. If there exists $n+1$ distinct hyperplanes $S_0, \ldots, S_n$ in R^s and $n+1$ pairwise disjoint subsets $A_0, \ldots, A_n$ of the set of nodes Δ_n such that for each $j = 0, \ldots, n$, A_j is a subset of $S_j \backslash \{S_{j+1} \cup \cdots \cup S_n\}$, has cardinality $N_j(s-1)$, and admits a unique Lagrange polynomial interpolation of degree j in $(s-1)$ variables, then Δ_n admits a unique Lagrange polynomial interpolation of degree n in R^s.

On the other hand, we remark that an arbitrary choice of n nodes in R^s admits an interpolation from some n-dimensional incomplete polynomial space if we neglect the degree of the polynomials.

We next discuss a Birkhoff-type interpolation problem by polynomials in two variables. It will be clear that an analogous problem can be posed in any dimension. Let S be a lower set of a finite number of pairs of integers; that is, whenever $i' \leq i$, $j' \leq j$ and $(i, j) \in S$, then $(i', j') \in S$. Denote by P_S the space of all polynomials $\sum a_{ij} x^i y^j$, where $a_{ij} = 0$ if $(i, j) \notin S$. Note that the dimension of P_S is simply $|S|$, the cardinality of S. Let $E = [e_{qik}]$ be a set of 0 and 1 defined for $q = 1, \ldots, m$ and $(i, k) \in S$. We shall call E an interpolation matrix. The interpolation problem we discuss here is to find a polynomial p in P_S such that

$$\frac{\partial^{i+k}}{\partial x^i \partial y^k} p(z_q) = c_{qik}$$

for all q, i, k with $e_{qik} = 1$, where $Z = \{z_1, \ldots, z_m\}$ is a given set of nodes in R^2 and $C = \{c_{qik}\}$ is a given set of data values.

An interpolation matrix E is said to be regular if for any sets Z and C the above interpolation problem is solvable. It is said to be an Abel matrix if, for each pair (i, k) in S, there is one and only one q for which $e_{qik} = 1$. The following result has been obtained:

Theorem 46. Let $E = [e_{qik}]$ be a normal interpolation matrix with respect to a lower index set S; that is,

$$\sum_{q=1}^{m} \sum_{(i,k)\in S} e_{qik} = |S|$$

Then E is regular if and only if it is an Abel matrix.

We now consider the problem of Lebesgue constants in the multivariate setting. Let V_i, $i = 1, \ldots, s+1$, be the vertices of a simplex T^s in R^s, and for each x in T^s let $(u_1, \ldots, u_{s+1})$ be its barycentric coordinate; that is,

$$x = \sum_{i=1}^{s+1} u_i V_i; \qquad \sum_{i=1}^{s+1} u_i = 1, \qquad u_i \geq 0$$

Let $a = (a_1, \ldots, a_{s+1})$, where $a_1, \ldots, a_{s+1}$ are nonnegative integers such that $|a| = a_1 + \cdots + a_{s+1} \leq n$. It is known that the set of equally spaced nodes $\{x_a\}$, $|a| \leq n$, where $x_a = (a_1 n^{-1}, \ldots, a_{s+1} n^{-1})$, admits a unique Lagrange interpolation. For this set of equally spaced knots $\{x_a\}$, we define

$$l_a(x) = n^n \prod_{i=1}^{s+1} \prod_{j=0}^{a_i - 1} \left(u_i - \frac{j}{n}\right) \Bigg/ \prod_{k=1}^{s+1} a_k!$$

$$L_n(x) = \sum_{|a| \leq n} |l_a(x)|$$

and

$$L_n = \max\{L(x) : x \in T^s\}$$

$L_n(x)$ and L_n are called the Lebesgue function and Lebesgue constant, respectively, corresponding to the knots $\{x_a\}$. The following estimate of the Lebesgue constant is known:

Theorem 47. For the above equally spaced nodes $\{x_a\}$, the Lebesgue constant satisfies

$$L_n \leq \binom{2n-1}{n}$$

for each $n = 1, 2, \ldots$. Furthermore,

$$L_n \to \binom{2n-1}{n}$$

as the dimension s tends to infinity.

Another problem in Lebesgue constants is concerned with optimality of the nodes. However, not much is known about this subject. We simply quote a result on its growth order, which again turns out to be logarithmic as expected. Let

$$n = \prod_{i=1}^{s} (n_i + 1)$$

and $E_n = \{x^1, \ldots, x^n\}$ be an admissible set of nodes in the unit cube I^s in the sense that for each function f in $C(I^s)$ there exists a unique polynomial

$$p_n(x, f) = \sum_{\substack{0 \leq k_i \leq n_i \\ i=1,\ldots,s}} b_k x^k$$

where $k = (k_1, \ldots, k_s)$, such that $p_n(x^i, f) = f(x^i)$, $i = 1, \ldots, n$. We define the "optimal" Lebesgue constant Λ_n by

$$\Lambda_n = \inf_{E_n \subset I^s} \sup_{f \in C(I^s)} \frac{\|p(\cdot, f)\|}{\|f\|}$$

where the supremum norm is being used. The following result has been obtained:

Theorem 48. There exist constants A_s and B_s depending only on the dimension s, such that

$$A_s \prod_{i=1}^{s} \ln(n_i + 2) \le \Lambda_n \le B_s \prod_{i=1}^{s} \ln(n_i + 2)$$

VIII. MULTIVARIATE POLYHEDRAL SPLINES

Spline functions in one variable can be considered linear combinations of B splines. Let $X = \{t_0, \ldots, t_n\}$, where $t_0 \le \cdots \le t_n$ with $t_0 < t_n$. Then the univariate B spline, with X as its set of knots and order n, is defined by

$$M(x, X) = n[t_0, \ldots, t_n](\cdot - x)_+^{n-1}$$

An application of the Hermite–Gennochi formula for divided differences gives the interesting relation

$$\int_R M(x, X)g(x)\,dx = n! \int_{S^n} g(u_0 t_0 + \cdots + u_n t_n)\,du_1 \cdots du_n$$

for any continuous function g, where S^n is the n simplex

$$\sum_{i=0}^{n} u_i = 1, \qquad u_i \ge 0$$

The development of multivariate polyhedral splines is based on a certain geometric interpretation of this relation. First, by simply replacing the knots $\{t_0, \ldots, t_n\}$ by a set $X = \{t^0, \ldots, t^n\}$ of $n+1$ points (which are not necessarily distinct) in R^s such that the algebraic span of X, denoted by $\langle X \rangle$, is all of R^s, we arrive at the formula

$$\int_{R^s} S(x, X)g(x)\,dx = n! \int_{S^n} g\left(u_0 t^0 + \cdots + u_n t^n\right) du_1 \cdots du_n$$

for any test function g in C_0, the class of all continuous functions in R^s with compact supports. The function $S(\cdot, X)$, called the simplicial spline with X as its set of knots, has a very nice geometric interpretation. Let $y^0, \ldots, y^n$ be in R^n such that the set

$$\sigma = \left\{ \sum_{i=0}^{n} u_i y^i : u_0 + \cdots + u_n = 1;\ u_0, \ldots, u_n \ge 0 \right\}$$

has positive volume and $Py^i = t^i$, $i = 0, \ldots, n$, where P is the projection of any y in R^s onto an x in R^s consisting of its first s coordinates. Then for each x in R^s, we have

$$S(x, X) = \mathrm{vol}_{n-s}\{y \in \sigma : Py = x\}/\mathrm{vol}_n \sigma$$

Next, let $X_n = \{x^1, \ldots, x^n\}$, where each x^i is a nonzero vector in R^s such that $\langle X_n \rangle = R^s$. Suppose that $I = [0, 1]$ and $B(\cdot, X_n)$ is defined, again in the distribution sense, by

$$\int_{R^s} B(x, X_n)g(x)\,dx = \int_{I^n} g\left(u_1 x^1 + \cdots + u_n x^n\right) du_1 \cdots du_n$$

for all test functions g in C_0. We call $B(\cdot, X_n)$ a box spline with directions $X_n = \{x^1, \ldots, x^n\}$. Box splines also have a nice geometric interpretation. Let

$$\eta = \left\{ \sum_{i=1}^{n} u_i y^i : 0 \le u_i \le 1,\ Py^i = x^i \right\}$$

where, and hereafter, P denotes an arbitrary orthogonal projection from R^n onto R^s. Then we can write

$$B(x, X_n) = \mathrm{vol}_{n-s}\{y \in \eta : Py = x\}/\mathrm{vol}_n \eta$$

If we consider the formula

$$\int_{R^s} C(x, X_n)g(x)\,dx = \int_{R_+^n} g\left(u_1 x^1 + \cdots + u_n x^n\right) du_1 \cdots du_n$$

for all g in C_0, we arrive at a truncated power (or cone) spline.

Box splines can be derived from truncated power splines, and conversely truncated power splines can be expressed in terms of box splines. Let

$$\Delta_y f(\cdot) = f(\cdot) - f(\cdot - y) \qquad \text{for any} \qquad y \text{ in } R^s$$

and

$$\Delta_{X_n} = \prod_{y \in X_n} \Delta_y$$

In addition, let $[X_n]$ be the $n \times n$ matrix whose ith column is x^i, and let

$$L(X_n) = \left\{ [X_n]b : b \in Z^n \right\}$$

and $t(a, X_n)$ denote the number of different solutions $b \in Z_+^n$ to $a = [X_n]b$. Here, Z and Z_+ denote, as usual, the set of all integers and nonnegative integers, respectively. Then we have the following relationships:

Theorem 49.

$$B(\cdot, X_n) = \Delta_{X_n} C(\cdot, X_n)$$

and

$$C(\cdot, X_n) = \sum_{a \in L(X_n)} t(a, X_n) B(\cdot - a, X_n)$$

A more general notion is the so-called multivariate polyhedral spline (or P shadow) $M_B(\cdot, P)$, where B is an arbitrary convex polyhedral body in R^n, defined in the distribution sense by

$$\int_{R^s} M_B(x, P) g(x)\, dx = \int_B g(Pu)\, du$$

for all g in C_0. Some important properties are included in the following two theorems. Let

$$D_y = \sum_{i=1}^{n} y_i \frac{\partial}{\partial x_i}$$

where $y = (y_1, \ldots, y_n)$, and if B is a polyhedral body in R^n, then B_i will denote an $(n-1)$-dimensional flat of B with unit outer normal n_i.

Theorem 50. For each z in R^n,

$$D_{Pz} M_B(\cdot, P) = -\sum_i z \cdot n_i M_{B_i}(\cdot, P) \tag{1}$$

$$M_B(Pz, P) = \frac{1}{n-s} \sum_i (b_i - z) \cdot n_i M_{B_i}(Pz, P) \tag{2}$$

$$D_x M_B(x, P) = (n-m) M_B(x, P) - \sum_i b_i \cdot n_i M_{B_i}(x, P) \tag{3}$$

where b_i lies on B_i.

Theorem 51. Let $n \geq m \geq s$, and P and Q be any orthogonal projections of R^n and R^m, respectively, onto R^s. If B and C are any two proper convex polyhedral bodies in R^n and R^m respectively, then

$$\int_{R^s} M_B(x - y, P) M_C(y, Q)\, dy = M_{B \times C}(x, P \otimes Q) \tag{1}$$

$$\int_{R^s} M_B(x, P) M_C(x, Q)\, dx = \frac{1}{n+m-s} \times \left\{ \sum_i (b_i - z) \cdot n_i \int_{R^s} M_{B_i}(x) M_C(x)\, ds + \sum_i (c_i - z') \cdot m_i \int_{R^s} M_B(x) M_{C_i}(x)\, dx \right\} \tag{2}$$

where b_i and c_i lie on B_i and C_i, respectively, z in R^n, with z' in R^m consisting of the first m components of z.

It is obvious that $M_B(\cdot, P)$ is nonnegative and vanishes outside PB. From the recurrence relations, it is also clear that it is a piecewise polynomial of degree $n-s$. Let us now specialize on box splines $B(\cdot, X_n)$. In addition to being a piecewise polynomial of degree $n-s$, it is in $C^{d-1}(R^s)$, where

$$d = d(X_n) = \max\{m : |Y| = m, Y \subset X_n, \langle X_n \backslash Y \rangle = R^s\}$$

The Fourier transform of $B(\cdot, X_n)$ is particularly simple, being

$$\hat{B}(x, X_n) = \prod_{y \in X_n} \frac{1 - e^{-ix \cdot y}}{ix \cdot y}$$

This immediately yields the following result, which is a special case of the above theorem:

$$B(\cdot, X_n \cup Y_m) = B(\cdot, X_n) * B(\cdot, Y_m)$$

Let $S(B)$ denote the linear span

$$\sum_{a \in L(X_n)} c_a B(\cdot - a, X_n)$$

of the translates of $B(\cdot, X_n)$. We first state the following result:

Theorem 52. Let Y be a subset of X_n with $\langle Y \rangle = R^s$. Then

$$\sum_{b \in Z^s} B(\cdot - Yb, X_n) = \frac{1}{|\det Y|}$$

Hence, appropriate translates of box splines form a partition of unity, so that $S(B)$ can at least approximate continuous functions. More important is the following result on linear independence of translates of a box spline.

Theorem 53. Let $X_n \subset Z^s \backslash \{0\}$. Then

$$\{B(\cdot - a, X_n) : a \in L(X_n)\}$$

is linearly independent if and only if

$$|\det Y| = 1$$

for all $Y \subset X_n$ with $\langle Y \rangle = R^s$.

As usual, let π_k denote the space of all polynomials in R^s with total degree at most k. Then the following result is obtained:

Theorem 54. Let $X_n \subset Z^s \backslash \{0\}$ and $d = d(X_n)$. Then there exists a linear functional λ on π_d such that

$$p(x) = \sum_{a \in Z^s} \lambda[p(\cdot + a)] B(x - a, X_n)$$

for all $p \in \pi_d$.

In fact, if we write

$$[\hat{B}(x, X_n)]^{-1} = \sum_j a_j x^j$$

where $j = (j_1, \ldots, j_s)$, and $x^j = x_1^{j_1} \cdots x_s^{j_s}$ then, using the notation

$$D^j = \left(\frac{\partial}{\partial x_1}\right)^{j_1} \cdots \left(\frac{\partial}{\partial x_s}\right)^{j_s}$$

we have

$$\lambda(p) = \sum_j a_j (-1)^{|j|} D^j p(0)$$

Let F be an extension of the linear functional $\lambda(p)$ on $C(R^s)$, such that,

$$|F(f)| \le \|F\| \|f\|, \qquad f \in C(R^s)$$

and $F(p) = \lambda(p)$ for any polynomial p. For any $h > 0$, define

$$(Q_h f)(x) = \sum_a F\{f[h(\cdot + a)]\} B\left(h^{-1}x - a, X_n\right)$$

Then the following estimate is established:

Theorem 55. There exist absolute constants C and r such that

$$|f(x) - (Q_h f)(x)| \le C h^{d+1-s/p} \sum_{|j|=d+1} \|D^j f\|_{LP(A_{x,h})}$$

for all $x \in R^s$ and $f \in H_p^d(R^s)$, where $d = d(X_n)$, and

$$A_{x,h} = x + rh[-1, 1]^s$$

At least for $p = \infty$, the order $O(h^{d+1})$ cannot be improved. Let us now replace B by a finite collection Φ of locally supported splines ϕ_i on R^s. Hence, $S(\Phi)$ will denote the linear span of the translates $\phi_i(\cdot, -a)$, $a \in Z^s$, where $i = 1, \ldots, N$, say. Let $H_{p,c}^m$ be the subspace of the Sobolev space H_p^m of compactly supported functions f with norm

$$\|f\|_{p,m} = \left\{ \sum_{|j| \le m} \|D^j f\|_p^p \right\}^{1/p}$$

We say that $S(\Phi)$ provides local L_p approximation of order k, or $S(\Phi) \in A_{p,k}$, if for any $f \in H_{p,c}^m$, there exist weights $w_{i,a}^h$ such that

$$\left\| f - \sum_{i=1}^N \sum_{a \in \mathbf{Z}^s} w_{i,a}^h \phi_i\left(\frac{\cdot}{h} - a\right) \right\|_p \le C h^k \|f\|_{p,m}$$

and $w_{i,a}^h = 0$ whenever $\operatorname{dist}(ah, \operatorname{supp} f) > r$ holds, where C and r are positive constants independent of h and f.

Theorem 56. The following statements are equivalent:

1. There exists a sequence $\{\psi_a\}$, $|a| < k$, in $S(\Phi)$ that satisfies

$$\hat{\psi}_0(a) = \delta_{0,a}, \qquad a \in 2\pi Z^s$$

and

$$\sum_{c \le b} \frac{(-i)^{|c|}}{c!} D^c \hat{\psi}_{b-c}(a) = 0$$

for $a \in 2\pi Z^s \backslash \{0\}$ and $1 \le |b| < k$.

2. There exists a sequence $\{\psi_a\}$, $|a| < k$, in $S(\Phi)$ such that

$$\frac{x^b}{b!} - \sum_{c \le b} \sum_{a \in Z^s} \psi_{b-c}(x - a) \frac{a^c}{c!}$$

is in $\pi_{|b|-1}$ for $|b| < k$.

3. There exist some finitely supported $w_{i,a}$ so that

$$\psi = \sum_{i=1}^N \sum_{a \in Z^s} w_{i,a} \phi_i(\cdot - a)$$

satisfies

$$\frac{x^b}{b!} - \sum_{a \in Z^s} \psi(x - a) \frac{a^b}{b!}$$

is in $\pi_{|b|-1}$ for $|b| < k$.

4. $S(\Phi) \in A_{p,k}$ for all p, $1 \le p \le \infty$.

5. $S(\Phi) \in A_{\infty,k}$.

In the two-dimensional setting, and particularly the three-directional mesh, the approximation order of $S(\Phi)$, where Φ is a collection of box splines, has been extensively studied. It is interesting that, although C^1 cubics locally reproduce all cubic polynomials, the optimal approximation rate is only $O(h^3)$. However, not much is known on the approximation order of $S(\Phi)$ for R^s with $s > 2$.

Although multivariate polyhedral splines are defined in the distributional setting, they can be computed by using the recurrence relations. The computational schemes, however, are usually very complicated. For R^2, there is an efficient algorithm that gives the Bézier coefficients for each polynomial piece of a box spline. In general, one could apply the subdivision algorithms, which are based on discrete box splines. Under certain mild assumptions, the convergence rate could be shown to be quadratic.

IX. QUADRATURES

This section is devoted to approximation of integrals. Let D be a region in R^s and $C(D)$ the space of all continuous

functions on D. The approximation scheme is to find linear functionals L_{nk}, $k = 1, \ldots, n$, on $C(D)$ such that the error functional r_n, defined by

$$\int_D f(x)w(x)\,dx = \sum_{k=1}^{n} L_{nk} f + r_n(f) = Q_n(f) + r_n(f)$$

where $w(\cdot)$ is some positive weight function, satisfying $r_n(f) \to 0$ as $n \to \infty$ for all f in $C(D)$. In one dimension, the above formula is called an integration quadrature formula. Multivariable quadrature formulas are sometimes called cubature formulas. There is a vast amount of literature on this subject. The most well-known quadrature formulas are, perhaps the trapezoidal and Simpson rules. These rules are obtained using certain interpolatory linear functionals L_{nk}. In addition to various other interpolatory-type formulas, there are automatic quadrature, Gauss-type quadrature, and integral formulas obtained by the adaptive Simpson (or Romberg) method. Of course, error analysis is very important. There is literature on algebraic precision, optimal quadrature, asymptotic expansion, and so on. In one variable, the most commonly used weights are perhaps 1, $(1-x)^\alpha x^\beta$, $\log(1/x)$ for the interval $[0, 1]$, e^{-x} for $[0, \infty)$ and e^{-x^2} for R. It is impossible to discuss many important results on this subject in a short section. We give only a brief description of three approaches in the one-variable setting and briefly discuss optimal and multivariable quadratures at the end of the section. For simplicity we will assume that $D = [-1, 1]$ in the following discussion.

A. Automatic Quadratures

An integration method that can be applied in a digital computer to evaluate definite integrals automatically within a tolerable error is called an automatic quadrature. The trapezoidal, Simpson, and adaptive Simpson rules are typical examples. We may say that an automatic quadrature is based on interpolatory-type formulas. An ideal interpolatory quadrature is one whose linear functionals L_{nk} are defined by $L_{nk} f = A_k^n(x_k)$, $k = 1, \ldots, n$, where $-1 \le x_1 < \cdots < x_n \le 1$, such that $A_k^n > 0$, the function values $f(x_k)$, can be used again in future computation (i.e., for $n+1, n+2, \ldots$), and that the quadrature has algebraic precision of degree $n-1$ [i.e., $r_n(f) = 0$ for $f(x) = 1, \ldots, x^{n-1}$]. Using Lagrange interpolation by polynomials of degree $n-1$ guarantees the algebraic precision. Here, we have

$$A_k^n = \int_{-1}^{1} \prod_{i \ne k} (x - x_i)/(x_k - x_i)\,dx$$

and the quadrature is sometimes attributed to Newton. However, the sequence $\{A_k^n\}$, $k = 1, \ldots, n$, is usually quite oscillatory. Since it is known that $r_n(F) \to 0$ for all $f \in C[-1, 1]$ if and only if

$$\sum_{k=1}^{n} |A_k^n| \le M < \infty$$

for all n, the sample points x_k should be chosen very carefully. For instance, equally spaced sample points yield

$$\sum_{k=1}^{n} |A_k^n| \to \infty$$

The following scheme is quite efficient. Let $t_1 = \frac{1}{4}$, $t_{2i} = t_i/2, t_{2i+1} = t_{2i} + \frac{1}{2}, i = 1, 2, \ldots,$ and $x_k = \cos 2\pi t_k, k = 1, 2, \ldots$. Fix an $N = 2^m$ for some $m > 0$, and let $T_k(x)$ and $U_k(x)$ denote the Chebyshev polynomials of the first and second kinds, respectively. Consider the polynomial

$$p_n(x) = \sum_{k=1}^{N-1} b_{0k} U_{k-1}(x) + \sum_{i=1}^{n} w_i [T_N(x)] {\sum_{k=0}^{N-1}}' b_{ik} T_k(x)$$

where $\sum'$ indicates that the first term under this summation is halved, and

$$w_m(x) = 2^m \prod_{k=1}^{m} (x - x_k)$$

For each f in $C[-1, 1]$, the polynomial $p_n(x)$ is then uniquely determined by the interpolation conditions $p_n(x_i) = f(x_i)$ for $i = 1, \ldots, (n+1)N - 1$. Write $b_{ik} = b_{ik}(f)$ for this interpolation polynomial. We then arrive at the quadrature

$$Q_{(n+1)N-1}(f) = \sum_{k=1}^{N-1} A_{0k} b_{0k}(f) + \sum_{i=1}^{n} \sum_{k=0}^{N-1} A_{ik} b_{ik}(f)$$

where

$$A_{0k} = \int_{-1}^{1} U_{k-1}(x)\,dx$$

and

$$A_{1k} = \int_{-1}^{1} w_i [T_N(x)] T_k(x)\,dx$$

for $i = 1, \ldots, n$. It is important to point out that $b_{ik}(f)$ and A_{ik} are independent of n and that to proceed from $Q_{(n+1)N-1}(f)$ to $Q_{(n+2)N-1}(f)$ we need only the extra terms $b_{n+1,k}(F)$ and $A_{n+1,k}$, for $k = 1, \ldots, N$, and their recurrence relations are available.

B. Gauss Quadrature

Let us return to the integration quadrature formula

$$\int_{-1}^{1} f(x)w(x)\,dx = Q_n(f) + r_n(f)$$

where

$$Q_n(f) = \sum_{k=1}^{n} A_k^n f(x_{nk})$$

with $-1 \le x_{n1} < \cdots < x_{nn} \le 1$. Gauss approached this problem by requiring $r_n(f) = 0$ for as many polynomials f with degrees as high as possible and proved that the highest degree is $2n-1$. In fact, this polynomial precision condition uniquely determines x_{nk} and the coefficients A_k^n. We call this $Q_n(f)$ a Gauss quadrature formula. The points x_{nk} turn out to be the roots of the orthonormal polynomials p_n with respect to the weight function w, and the coefficients A_k^n, known as Christoffel numbers, are

$$A_k^n = \left(\sum_{j=0}^{n-1} p_j^2(x_{nk}) \right)^{-1}$$

If $w(x) = 1$, then the polynomials p_n are the Legendre polynomials, and if $w(x) = (1-x^2)^{-1/2}$, the p_n's are (constant multiples of) Chebyshev polynomials of the first kind. The corresponding quadrature formulas are called Gauss–Legendre and Gauss–Chebyshev quadratures. The following estimate has been obtained:

Theorem 57. Let $1 \le s < 2n$ and $(1-x^2)^s f^{(s)}(x)$ be integrable on $[-1, 1]$. Then the error $r_n(f)$ of the Gauss–Legendre quadrature of f satisfies

$$|r_n(f)| \le C_s \int_{-1}^{1} \left| f^{(s)}(x) \right| \min\{(1-x^2)^{s/2}/n^s, (1-x^2)^s\}\, dx$$

where the constant C_s is independent of n and f.

Several generalizations of the Gauss quadrature have been considered. For instance, requiring the weaker polynomial precision condition $r_n(f) = 0$ for all f in π_{2n-m-1} gives a so-called $[2n-m-1, n, w(x)]$ quadrature. The following result gives a characterization of this quadrature. We denote by P_n the polynomials in π_n the polynomials in π_n with leading coefficient equal to 1 that are orthogonal on the unit circle with respect to the weight function

$$w(\cos\theta)|\sin\theta|$$

Theorem 58. Let $0 < m \le n$. Then $a[2n-m-1, n, w(x)]$ quadrature with nodes x_i in $[-1, 1]$ has $n-l$ positive coefficients and l negative coefficients if and only if there exists a polynomial q_m in π_m with leading coefficient equal to 1 and real coefficients such that

$$2^{-n+1} \mathrm{Re}\left\{ z^{-n+1} q_m(z) P_{2n-m-1}(z) \right\} = \prod_{j=1}^{n} (x - x_j)$$

where $x = \cos\theta$ and $z = e^{i\theta}$, $\theta \in [0, \pi]$, and q_m has $m-2l$ zeros inside and $2l$ zeros outside the unit circle.

Another generalization is the quadrature

$$\int_{-1}^{1} f(x) w(x)\, dx = \sum_{j=1}^{n} \sum_{i=1}^{n_j} c_{ji} f^{(i-1)}(x_j) + \sum_{j=1}^{m} \sum_{i=1}^{m_j} d_{ji} f^{(i-1)}(y_j) + r_N(f)$$

where $N = m_1 + \cdots + m_m + n_1 + \cdots + n_n$ and $y_1, \ldots, y_m$ are prescribed. It is well known that the polynomial precision requirement $r_N(f) = 0$ for all f in π_{N+n} completely characterizes the nodes $x_1, \ldots, x_n$. In fact, they are determined by the n equations

$$\int_{-1}^{1} w(x) x^k \prod_{j=1}^{m} (x - y_j)^{m_j} \prod_{j=1}^{m} (x - x_j)^{n_j}\, dx = 0 \qquad k = 0, \ldots, n-1.$$

There are methods for computing the nodes x_j and coefficients c_{ji} and d_{ji}. In fact, by establishing a relation between these coefficients and the eigenvectors of the Jacobi matrix corresponding to the weight function $w(\cdot)$, one can obtain a stable and efficient algorithm to evaluate these nodes and coefficients.

C. Chebyshev Quadratures

It is easy to see that all the Christoffel numbers of the Gauss–Chebyshev quadrature are the same. This leads to the problem of determining the weight functions for which the constants in the corresponding quadratures are independent of the summation index. More precisely, a quadrature formula

$$\int_{-1}^{1} f(x) w(x)\, dx = A_n \sum_{k=1}^{n} f(x_{nk}) + r_n(f)$$

$-1 \le x_{n1} < \cdots < x_{nn} \le 1$, that satisfies $r_n(f) = 0$ for all f in π_{n-1} is called a Chebyshev quadrature formula. Bernstein proved that $w(x) = 1$ does not give a Chebyshev quadrature for $n = 8$ or $n \ge 10$. It has been shown that the weights

$$w(x, p) = (2p + 1 + x)^{-1} (1 - x^2)^{-1/2}$$

where $p \ge 1$, give Chebyshev quadratures for all $n \ge 1$. This means, of course, that $A_n = A_n(p)$ and $x_{nk} = x_{nk}(p)$ exist such that the quadrature formulas are $(n-1)$st degree polynomial precise.

Finally, we discuss multivariable quadrature (or cubature) formulas very briefly. With the exception of very special regions (balls, spheres, cones, etc.), nontensor-product

cubatures are very difficult to obtain. Some of the reasons are that multivariate interpolation may not be unisolvent and that orthogonal polynomials are difficult to find on an arbitrary region D and R^s. There are, however, other methods, including, Monte Carlo, quasi Monte Carlo, and number-theoretic methods. On the other hand, many of the one-variable results can be generalized to Cartesian product regions. For instance, let us study optimal quadratures.

Let $H_p^r(I)$ denote the Sobolev spaces on $I=[-1, 1]$, and $W_{r,p}$ the unit ball of f in $H_p^r(I)$; that is

$$\left\| f^{(r)} \right\|_p \le 1$$

To extend to two variables, for instance, let $V_{r,s;p}(M)$ be the set of all functions $g(\cdot, \cdot)$ in the Sobolev space $H_p^{(r,s)}(I^2)$ such that

$$\left\| g^{(r,s)}(\cdot, \cdot) \right\|_p \le M$$

where M is a fixed positive constant. Now, in the one-variable setting, if

$$r_m(f) = \int_{-1}^{1} f(x)\,dx - \sum_{k=1}^{m} A_k^m f(x_{mk})$$

and

$$r_m(W_{r,p}) = \sup\{|r_m(f)| : f \in W_{r,p}\}$$

it is known that $r_m(W_{r,p})$ attains its minimum at some set of nodes $\bar{x}_{m1}, \ldots, \bar{x}_{mm}$ in I and some coefficients $\bar{A}_1^m, \ldots, \bar{A}_m^m$. The resulting quadrature formula,

$$\int_{-1}^{1} f(x)\,dx = \sum_{k=1}^{m} \bar{A}_k^m f(\bar{x}_{mk}) + \bar{r}_{m,r}(f)$$

is called an optimal quadrature formula for $W_{r,p}$. Let

$$\int_{-1}^{1} f(x)\,dx = \sum_{k=1}^{n} \bar{B}_k^n f(\bar{y}_{nk}) + \bar{r}_{n,s}(f)$$

be an optimal quadrature formula for $W_{s,p}$. The following result has been obtained:

Theorem 59. Let $1 < p \le \infty$. Then the formula

$$\int_{I^2} g(x, y)\,dx\,dy = \sum_{k=1}^{m} \bar{A}_k^m \int_I g(\bar{x}_{mk}, y)\,dy + \sum_{j=1}^{n} \bar{B}_j^n \int_I g(x, \bar{y}_{nj})\,dx - \sum_{k=1}^{m}\sum_{j=1}^{n} \bar{A}_k^m \bar{B}_j^n g(\bar{x}_{mk}, \bar{y}_{nj}) + r_{mn}^*(g)$$

is optimal in the sense that $r_{mn}^*[V_{r,s;p}(M)]$, defined to be

$$\sup\left\{\left|r_{mn}^*(g)\right| : g \in V_{r,s;p}(M)\right\}$$

does not exceed $r_{m,n}[V_{r,s;p}(M)]$, defined analogously, where $r_{mn}(g)$ has the same form as $r_{mn}^*(g)$ with arbitrary nodes x_k and y_j in I and arbitrary coefficients A_j, B_k, and C_{kj}. Furthermore

$$r_{mn}^*[V_{r,s;p}(M)] = M\bar{r}_{r,p}(W_{r,p})\bar{r}_{n,s}(W_{s,p})$$

X. SMOOTHING BY SPLINE FUNCTIONS

Not only are spline functions useful in interpolation and approximation; they are natural data-smoothing functions, especially when the data are contaminated with noise. We first discuss the one-variable setting.

Let $0 = t_0 < \cdots < t_{n+1} = 1$ and $H_2^k = H_2^k[0, 1]$ be the Sobolev space of functions f such that $f, \ldots, f^{(k)}$ are in $L_2 = L_2[0, 1]$, where $k \le n-1$. For any given data $Z = [z_1, \ldots, z_n]^T$, it can be shown that there is a unique function $s_{k,p} = s_{k,p}(\cdot;\ Z)$, which minimizes the functional

$$J_{k,p}(u; Z) = p\int_0^1 [D^k u]^2 + \frac{1}{n}\sum_{i=1}^{n}[u(t_i) - z_i]^2$$

over all functions u in H_2^k. In fact, the extremal function $s_{k,p}$ is a natural spline of order $2k$ having its knots at $t_1, \ldots, t_n$. The parameter $p > 0$, called the smoothing parameter, controls the "trade-off" between the smoothness of $s_{k,p}$ and its approximation to the data Z. Let $s_{k,p}(t_i) = y_i$ and $Y = [y_1, \ldots, y_n]^T$. Then $Y = A_k(p)Z$, where the $n \times n$ matrix $A_k(p)$ is called the influence matrix. Determining $s_{k,p}$ is equivalent to determining Y. Hence, the influence matrix is importtant. It can be shown that $A_k(p) = (I_n + npB_k)^{-1}$, where B_k is an $n \times n$ nonnegative definite symmetric matrix with exactly k zero eigen-values. The other $n-k$ eigenvalues of B_n are positive and simple, regardless the location of the t_i. Let these eigenvalues be $d_i = d_{ni}$ arranged in nondecreasing order as follows: $0 = d_1 = \cdots = d_k < d_{k+1} < \cdots < d_n$. It is clear that if Q is the unitary matrix that diagonalizes B_k in the sense that $B_k = Q^* D_k Q$, where $D_k = \mathrm{diag}[d_1 \cdots d_n]$, then setting $\hat{Y} = Q^*Y$ and $\hat{Z} = Q^*Z$, we have

$$\hat{y}_i = \frac{1}{1 + npd_i}\hat{z}_i$$

where $\hat{Y} = [\hat{y}_1 \cdots \hat{y}_n]^T$ and $\hat{Z} = [\hat{z}_1 \cdots \hat{z}_n]^T$ could be viewed as the Fourier transforms of Y and Z, respectively.

Theorem 60. Let $\{t_i\}$ be quasi-uniformly distributed; that is

$$\max(t_{i+1} - t_i) \le c \min(t_{t+1} - t_i)$$

where c is independent of n as $n \to \infty$. Then there exist positive constants c_1 and c_2 such that

$$c_1 i^{2k} \le n d_{ni} \le c_2 i^{2k}$$

for $i = k+1, \ldots, n$ and $n = 1, 2, \ldots$.

Hence, for quasi-uniformly distributed $\{t_i\}$, we have the following asymptotic behavior of $\hat{y}_i$, $i = k+1, \ldots, n$:

$$\hat{y}_1 \sim \frac{1}{1 + (p^{1/2k} r i)^{2k}} \hat{z}_i \sim b_k(p^{1/2k} r i)\hat{z}_i$$

as $n \to \infty$, where r is a positive constant and b_k a Butterworth-type low-pass filter with the inflection point as its cutoff frequency $i_c = (r p^{1/2k})^{-1}$. The size of the window of this filter decreases monotonically as the smoothing parameter p increases. In the limiting cases as $p \to \infty$, we have

$$\lim_{p\to\infty} \hat{y}_i = \begin{cases} \hat{z}_i & \text{if } i = 1, \ldots, k \\ 0 & \text{if } i = k+1, \ldots, n \end{cases}$$

In this case, the smoothing spline $s_{k,\infty}(\cdot;\ Z)$ is simply the least-squares polynomial of degree $k-1$. On the other hand, as $p \to 0$, we have

$$\lim_{p\to 0} \hat{y}_i = \hat{z}_i, \qquad i = 1, \ldots, n$$

so that $s_{k,0}(\cdot;\ Z)$ is the natural spline of order $2k$ interpolating the data Z.

Suppose that the data are contaminated with noise, say $z_i = g(t_i) + e_i$, where $g \in H_2^k$ and $e_1, \ldots, e_n$ are independent random variables with zero mean and positive standard deviation. A popular procedure for controlling the smoothing parameter p is the method of generalized cross-validation (GCV). We will not go into details but will remark that it is a predictor–corrector procedure.

The idea of a smoothing spline in one dimension has a natural generalization to the multivatiate setting. Unfortunately not much is known for an arbitrary domain D in R^s. In the following, the L_2 norm $\| \ \|_2$ will be taken over all of R^s, and this accounts for the terminology of "thin-plate splines." Again let $p > 0$ be the smoothing parameter and $k > s/2$. Then for every set of data $\{Z_1, \ldots, Z_n\}$ in R and a scattered set of sample points $\{t_1, \ldots, t_n\}$ in R^s, the thin-plate smoothing spline of order $2k$ and with smoothing parameter p is the (unique) solution of the minimization problem,

$$\inf_{u \in D^{-k}L^2(R^s)} \left\{ p|u|_k^2 + \frac{1}{n}\sum_{i=1}^{n} [u(t_i) - z_i]^2 \right\}$$

where

$$|u|_k^2 = \sum_{i_1,\ldots,i_k=1}^{s} \left\| \frac{\partial^k}{\partial x_{i_1} \cdots \partial x_{i_k}} u \right\|_2^2$$

and

$$D^{-k}L^2(R^s) = \{u \in D'(R^s) : D^a u \in L^2(R^s),\ |a| = k\}$$

with $D'(R^s)$ denoting the space of Schwartz distributions on R^s and

$$D^a = \frac{\partial^{|a|}}{\partial x_1^{a_1} \cdots \partial x_s^{a_s}}$$

Of course, all derivatives are taken in the distributional sense. Since $K > s/2$, the elements in $D^{-k}L^2(R^s)$ are continuous functions.

Theorem 61. The smoothing spline $s_{k,p} = s_{k,p}(\cdot;\ Z)$ that minimizes the above optimization problem exists and is unique and satisfies the distributional equation

$$(-1)^k \Delta^k S_{k,p} = \frac{1}{np} \sum_{i=1}^{n} c_i \delta_{t_i}$$

where δ_t is the Dirac distribution and $c_i = z_i - s_{k,p}(t_i; Z)$, $i = 1, \ldots, n$. Furthermore, $\Delta^k s_{k,p}$ is orthogonal to the collection π_{k-1} of polynomials of total degree $k-1$ in the sense that

$$\langle \Delta^k s_{k,p},\ p \rangle = \sum_{i=1}^{n} c_i P(t_i) = 0$$

for all P in π_{k-1}.

Let K_k be the elementary solution of the k-times-iterated Laplacean:

$$\Delta^k K_k = \delta_0$$

Then it is well known that

$$K_k(t) = \begin{cases} c_k |t|^{2k-s} & \text{if } s \text{ is odd} \\ c_k |t|^{2k-s} \log |t| & \text{if } s \text{ is even} \end{cases}$$

where $c_k = (-1)^k 2\pi [2^{k-1}(k-1)!]^2$. Also, if m is a measure with compact support orthogonal to π_{k-1}, then $m * K_k$ is an element of $D^{-k}L^2(R^s)$. Applying this result to $\Delta^k s_{k,p}$, we have

$$(-1)^k \Delta^k s_{k,p} * K_k = \frac{1}{np} \sum_{i=1}^{n} c_i K_k(t - t_i)$$

and

$$\begin{aligned} \Delta^k \big((-1)^k \Delta^k s_{k,p} * K_k \big) &= (-1)^k \big(\Delta^k s_{k,p} \big) * \big(\Delta^k K_k \big) \\ &= (-1)^k \big(\Delta^k s_{k,p} \big) * \delta_0 \\ &= (-1)^k \Delta^k s_{k,p} \end{aligned}$$

This yields

$$s_{k,p} - (-1)^k \Delta^k s_{k,p} * K_k \in \pi_{k-1}$$

Theorem 62. The smoothing spline $s_{k,p}$ satisfies

$$s_{k,p}(t) = (-1)^k \sum_{i=1}^{n} d_i K_k(t - t_i) + P(t)$$

where $P \in \pi_{k-1}$. Furthermore, the polynomial P and the coefficients $d_1, \ldots, d_n$ are uniquely determined by the equations:

$$(-1)^k \frac{1}{np} d_i + s_{k,p}(t_i) - z_i = 0, \qquad i = 1, \ldots, n$$

and

$$d_1 Q(t_1) + \cdots + d_n Q(t_n) = 0, \qquad Q \in \pi_{k-1}$$

Hence, if the values $y_i = s_{k,p}(t_i; Z)$, $i = 1, \ldots, n$, are known, the smoothing spline $s_{k,p}$ is also determined. Writing $Y = A_k(p)Z$, where $Y = [y_1, \ldots, y_n]^T$, it is important to study the influence matrix $A_k(p)$ as in the one-variable setting. Again it is possible to write

$$A_k(p) = (I + npB_k)^{-1}$$

Let $\{t_1, \ldots, t_n\}$ lie in a bounded domain D and R^s. We say that $\{t_i\}$ is asymptotically quasi-uniformly distributed in D if for all large n,

$$\sup_{t \in D} \min_{1 \le i \le n} |t - t_i| \le c \min_{1 \le i < j \le n} |t_1 - t_j|$$

where c is some positive constant, depending only on D.

Theorem 63. Let D satisfy a uniform cone condition and have Lipschitz boundary and $\{t_1, \ldots, t_n\}$ be asymptotically quasi-uniformly distributed in D. Then there exists a positive constant C such that the eigenvalues $d_{n1}, \ldots, d_{nn}$ of B_k satisfy $0 = d_{n1} = \cdots = d_{nk} < d_{n,k+1} < \cdots < d_{nn}$ and

$$i^{2k/s} \le n d_{ni} \le C i^{2k/s}, \qquad i = N+1, \ldots, n$$

for all n, where

$$N = \dim \pi_{k-1} = \binom{k+s-1}{s}$$

Thus, if we study the effect of thin-plate spline smoothing in the "frequency domain" as before by writing

$$\hat{Y} = Q^* Y; \qquad \hat{Z} = Q^* Z$$

where Q is unitary with $QB_kQ^* = \operatorname{diag}[d_{n1} \cdots d_{nn}]$, we obtain

$$\hat{y}_i = \frac{1}{1 + p d_{ni}} \hat{z}_i \sim b_k(p^{1/2k} r i) \hat{z}_i$$

where, again, b_k is a Butterworth-type low-pass filter with cutoff frequency at $(rp^{s/2k})^{-1}$. Similar conclusions can be drawn on the limiting cases as $p \to \infty$ and $p \to 0$, and again the method of GCV is commonly used to choose the smoothing parameter p. It is a predictor–corrector procedure and amounts to choosing p as the minimizer of the GCV function:

$$\frac{1}{n} \sum_{i=1}^{n} [s_{k,p}(t_i; Z) - z_i]^2 \Bigg/ \left[1 - \frac{1}{n} tr A_k(p)\right]^2$$

Of course, $s_{k,p}(t_i; Z) = y_i$ and determining $\{y_i\}$ is equivalent to determining $s_{k,p}$ itself.

In a more general setting, let X, H, K be three Hilbert spaces and T, A be linear operators mapping X onto H and K and having null spaces $N(T)$, $N(A)$, respectively. Let $p > 0$ be the smoothing parameter in the functional

$$J_{k,p}(u, Z) = p\|Tu\|_H^2 + \|Au - Z\|_K^2$$

where Z is the given "data" vector in K and $u \in X$.

Theorem 64. If $N(T) + N(A)$ is closed in X with $N(T) \cap N(A) = \{0\}$, there exists a unique function s_p in X satisfying

$$J_{k,p}(s_p, Z) = \inf\{J_{k,p}(u, Z) : u \in X\}$$

Moreover, if S denotes the space of all "spline" functions defined by

$$S = \{s \in X : T^*Ts \text{ is in the range of } A^*\}$$

where T^*, A^* denote the adjoints of T and A, respectively, then s_p satisfies $s_p \in S$ and

$$T^*Ts_p = p^{-1} A^*(Z - As_p)$$

As a corollary, it is clear that $s_p \in X$ is the smoothing spline function if and only if

$$\langle Ts_p, Tx \rangle_H = p^{-1} \langle Z - As_p, Ax \rangle_K$$

for all $x \in X$.

An interesting special case is $X = H_2^k(D)$, where D is a bounded open set in R^s with Lipschitz boundary and satisfying a uniform cone condition. If $k > s/2$, then the evaluation functional δ_t is continuous. Let $t_i, \ldots, t_n \in D$ and set $K = R^n$ and $H = [L_2(D)]^N$, where $N = s^k$. Also, define

$$Au = \left[\delta_{t_1} u \cdots \delta_{t_n} u\right]^T$$

and

$$T = \frac{\partial^k}{\partial x_{i_1} \cdots \partial x_{i_k}}$$

where $i_1 = 1, \ldots, s$ and $i_k = 1, \ldots, s$, and equip H with the norm $\| \ \|_H$, where

$$\|v\|_H^2 = \sum_{t_1, \ldots, i_k = 1}^{d} \left\| v_{i_{i_1 \cdots i_k}} \right\|_{L_2}^2 (D)$$

Then using Theorem 64 and the fact that $\{t_1, \ldots, t_n\}$ is a π_{k-1} unisolvent set, the following result can be proved:

Theorem 65. There exists a unique s_p in $H_2^k(D)$ that minimizes $J_{k,p}(u, Z)$ for all u in $H_2^k(D)$. Furthermore, s_p satisfies the distributional partial differential equation.

$$(-1)^k \Delta^k s_p = \frac{1}{p} \sum_{i=1}^{n} [z_i - s_p(t_i)] \delta_{t_i}$$

In fact, the following estimate has also been obtained for noisy data:

Theorem 66. Let $z_i = g(t_i) + e_i, \quad i = 1, \ldots, n$, where g is in $H_2^k(D)$, and let $e_1, \ldots, e_n$ be independent identically distributed random variables with variance V. If the points t_i satisfy a quasiuniform distribution condition in D, then there exist positive constants c_1 and c_2, independent of n, such that for $0 \le p \le 1$,

$$E\left[|s_p(\cdot\,; Z) - g|_j^2\right] \le c_1 p^{(k-j)/k} + \frac{1}{n} c_2 V p^{-(2j+s)/2k}$$

for $n^{-2nk/s} \le p \le 1$, where $1 \le j \le k$.

However, no general result on asymptotic optimality of cross-validation is available at the time of this writing.

XI. WAVELETS

The Fourier transform $\hat{f}(w)$ of a function $f(t)$ represents the spectral behavior of $f(t)$ in the frequency domain. This representation, however, does not reflect time-evolution of frequencies. A classical method, known as short-time Fourier transform (STFT), is to window the Fourier integral, so that, by the Plancherel identity, the inverse Fourier integral is also windowed. This procedure is called time-frequency localization. This method is not very efficient, however, because a window of the same size must be used for both low and high frequencies. The integral wavelet transform (IWT), defined by

$$(W_\psi f)(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \overline{\psi\left(\frac{t-b}{a}\right)}\, dt,$$

on the other hand, has the property that the time-window narrows at high frequency and widens at low frequency. This is seen by observing that, for real f and ψ,

$$(W_\psi f)(b, a) = Re \frac{\sqrt{a}}{\pi} \int_0^{\infty} e^{-ib\omega} \hat{f}(\omega) \overline{g\left(a\left(\omega - \frac{\omega_0}{a}\right)\right)}\, d\omega;$$

where $g(\omega) := \hat{\psi}(\omega + \omega_0)$, and ω_0 is the center of the frequency window function $\hat{\psi}(\omega)$. More precisely, the window in the time-frequency domain may be defined by

$$[b - a\Delta_\psi, b + a\Delta_\psi] \times \left[\frac{\omega_0}{a} - \frac{1}{a}\Delta_{\hat{\psi}}, \frac{\omega_0}{a} + \frac{1}{a}\Delta_{\hat{\psi}}\right],$$

where Δ_ψ and $\Delta_{\hat{\psi}}$ denote the standard deviations of ψ and $\hat{\psi}$, respectively, and we have identified the frequency by a constant multiple of a^{-1}. Hence, the IWT has the zoom-in and zoom-out capability.

This localization of time and frequency of a signal function $f(t)$, say, not only makes it possible for filtering, detection, enhancement, etc., but also facilitates tremendously the procedure of data reduction for the purpose of storage or transmittance. The modified signal, however, must be reconstructed, and the best method for this reconstruction is by means of a wavelet series. To explain what a wavelet series is, we rely on the notion of multiresolution analysis as follows: Let $\{V_n\}$ be a nested sequence of closed subspaces of $L^2 = L^2(\mathbb{R})$, such that the intersection of all V_n is the zero function and the union of all V_n is dense in L^2. Then we say that $\{V_n\}$ constitutes a multiresolution analysis of L^2 if for each $n \in \mathbb{Z}$, we have $f \in V_n \Leftrightarrow f(2\cdot) \in V_{n+1}$ and if there exists some $\phi \in V_0$ such that the integer translates of ϕ yields an unconditional basis of V_0. We will also say that ϕ generates a multiresolution analysis of L^2. Examples of ϕ are B-splines of arbitrary order and with $\mathbb{Z}$ as the knot sequence. Next, for each $k \in \mathbb{Z}$, let W_k be the orthogonal complementary subspace of V_{k+1} relative to V_k. Then it is clear that the sequence of subspaces $\{W_k\}$ is mutually orthogonal, and the orthogonal sum is all of L^2. Consequently, every $f \in L^2$ can be decomposed as an orthogonal sum of functions $g_k \in W_k, k \in \mathbb{Z}$. It can be proved that there exists some function ψ whose integer translates form an unconditional basis of L^2. If this ψ, called a *wavelet*, is used as the window function in the IWT, then the "modified" signal $f(t)$ can be reconstructed from IWT at dyadic values by means of a wavelet series.

Let us first introduce the dual $\tilde{\psi} \in W_0$ of ψ defined (uniquely) by

$$\langle \tilde{\psi}, \psi(\cdot - n) \rangle = \delta_{n,0}, \qquad n \in \mathbb{Z}.$$

Then, by using the notation

$$\psi_{k,j}(t) = \psi(2^k t - j)$$

where $j, k \in \mathbb{Z}$, and the same notation for $\tilde{\psi}_{k,j}$, we have the following.

Theorem 67. For every $f \in L^2$, then

$$f(t) = \sum_{k,j \in Z} 2^{k/2} (W_\psi f)(j2^{-k}, 2^{-k}) \tilde{\psi}_{k,j}(t) = \sum_{k,j \in Z} 2^{k/2} (W_{\hat{\psi}} f)(j2^{-k}, 2^{-k}) \psi_{k,j}(t).$$

Algorithms are available to find the coefficients of the partial sums of this series and to sum the series using values of the coefficients. Since these coefficients are $2^{k/2}$ multiples of the IWT at $(j2^{-k}, 2^{-k})$ in the time-scale plane,

these algorithms determine the IWT efficiently without integration and reconstruct $f(t)$ efficiently without summing at every t. Based on two pairs of sequences (p_n, q_n) and (a_n, b_n), which are known as reconstruction and decomposition sequences, respectively, a pyramid algorithm is realizable. Here, $\{p_n\}$ defines the ϕ that generates the multiresolution analysis, namely:

$$\phi(t) = \sum_{n\in\mathbb{Z}} p_n \phi(2t - n)$$

and $\{q_n\}$ relates ψ to ϕ by:

$$\psi(t) \sum_{n\in\mathbb{Z}} q_n \phi(2t - n).$$

In addition, $\{a_n\}$ and $\{b_n\}$ determine the decomposition $V_1 = V_0 \oplus W_0$ by

$$\phi(2t - \ell) = \sum_{n\in\mathbb{Z}} a_{\ell-2n}\phi(t - n) + \sum_{n\in\mathbb{Z}} b_{\ell-2n}\psi(t - n),$$

for all $\ell \in \mathbb{Z}$.

If $\phi(t) = N_m(t)$ is the mth order B-spline with integer knots and supp $N_m = [0, m]$, then we have the following result.

Theorem 68. For each positive integer m, the minimally supported wavelet ψ_m corresponding to $N_m(t)$ is given by

$$\psi_m(t) = \frac{1}{2^{m-1}} \sum_{j=0}^{2m-2} (-1)^j N_{2m}(j + 1) \times \sum_{\ell=0}^{m} (-1)^\ell \binom{m}{\ell} N_m(2t - j - \ell).$$

Furthermore, the reconstruction sequence pair is given by:

$$p_n = s^{-m+1}\binom{m}{n}$$

and

$$q_n = \frac{(-1)^n}{2^{m-1}} \sum_{j=0}^{m} \binom{m}{j} N_{2m}(n - j + 1),$$

with supp$\{p_n\} = [0, m]$ and supp$\{q_n\} = [0, 3m - 2]$.

To describe the pair of decomposition sequences, we need the "Euler–Forbenius" polynomial

$$\prod\nolimits_m(z) := \sum_{n=0}^{2m-2} N_{2m}(n + 1)z^n,$$

Where we have omitted the multiple of $[(2m - 1)!]^{-1}$ for convenience. Then $\{a_n\}$ and $\{b_n\}$ are determined by the Laurent series:

$$G(z) = \frac{1}{2^m}(1 + z)^m \frac{\prod_m(z)}{z\prod_m(z^2)} = \sum_{n\in\mathbb{Z}} a_n z^{-n}$$

$$H(z) = -\frac{1}{2^m}(1 - z)^m \frac{1}{z\prod(z^2)} = \sum_{n\in\mathbb{Z}} b_n z^{-n}.$$

The duals $\tilde{N}_m$ of N_m and $\tilde{\psi}_m$ of ψ can be computed by using the following formulas:

$$\tilde{N}_m(t) = \sum_{n\in\mathbb{Z}} 2a_m \tilde{N}_m(2t - n)$$

$$\tilde{\psi}_m(t) = \sum_{n\in\mathbb{Z}} 2b_n \tilde{N}_m(2t - n).$$

SEE ALSO THE FOLLOWING ARTICLES

FRACTALS • KALMAN FILTERS AND NONLINEAR FILTERS • NUMERICAL ANALYSIS • PERCOLATION • WAVELETS, ADVANCED

BIBLIOGRAPHY

Baker, G. A., Jr., and Graves-Morris, P. R. (1981). "Encyclopedia of Mathematics and Its Applications," Padé Approximants, Parts I and II, Addison-Wesley, Reading, Massachusetts.

Braess, D. (1986). "Nonlinear Approximation Theory," Springer-Verlag, New York.

Chui, C. K. (1988). "Multivariate Splines," CBMS-NSF Series in Applied Math. No. 54, SIAM, Philadelphia.

Chui, C. K. (1992). "An Introduction to Wavelets," Academic Press, Boston.

Chui, C. K. (1997). "Wavelets : A Mathematica Tool for Signal Analysis," SIAM, Philadelphia.

Daubechies (1992). "Ten Lectures on Wavelets," CBMS-NSF Series in Applied Math. No. 61, SIAM, Philadelphia.

Petrushev, P. P., and Popov, V. A. (1987). "Rational Approximation of Real Functions," Cambridge University Press, Cambridge, England.

Pinkus, A. (1985). "n-Widths in Approximation Theory," Springer-Verlag, New York.

Pinkus, A. (1989). "On L^1-Approximation," Cambridge University Press, Cambridge, England.

Wahba, G. (1990). "Spline Models for Observational Data," CBMS-NSF Series in Applied Math. No. 59, SIAM, Philadelphia.



Boolean Algebra

Raymond Balbes
University of Missouri-St. Louis

I. Basic Definitions and Properties
II. Representation Theory
III. Free Boolean Algebras and Related Properties
IV. Completeness and Higher Distributivity
V. Applications

GLOSSARY

Atom Minimal nonzero element of a Boolean algebra.
Boolean homomorphism Function from one Boolean algebra to another that preserves the operations.
Boolean subalgebra Subset of a Boolean algebra that is itself a Boolean algebra under the original operations.
Complement In a lattice with least and greatest elements, two elements are complemented if their least upper bound is the greatest element and their greatest lower bound is the least element of the lattice.
Complete Pertaining to a Boolean algebra in which every subset has a greatest lower bound and a least upper bound.
Field of sets Family of sets that is closed under finite unions, intersections, and complements and also contains the empty set.
Greatest lower bound In a partially ordered set, the largest of all of the elements that are less than or equal to all of the elements in a given set.
Ideal In a lattice, a set that is closed under the formation of finite least upper bounds and contains all elements of the lattice that are less than elements in the ideal.
Lattice [distributive] Partially ordered set for which the least upper bound $x + y$ and the greatest lower bound xy always exist [a lattice in which the identity $x(y + z) = xy + xz$ holds].
Least upper bound In a partially ordered set, the smallest of all of the elements that are greater than or equal to all of the elements of a given set.
Partially ordered set Set with a relation "$\leq$" that satisfies certain conditions.

A BOOLEAN ALGEBRA is an algebraic system consisting of a set of elements and rules by which the elements of this set are related. Systematic study in this area was initiated by G. Boole in the mid-1800s as a method for describing "the laws of thought" or, in present-day terminology, mathematical logic. The subject's applications are much wider, however, and today Boolean algebras not only are used in mathematical logic, but since they form the basis for the logic circuits of computers, are of fundamental importance in computer science.

I. BASIC DEFINITIONS AND PROPERTIES

A. Definition of a Boolean Algebra

Boolean algebras can be considered from two seemingly unrelated points of view. We shall first give a definition in terms of partially ordered sets. In Section I.C, we shall give an algebraic definition and show that the two definitions are equivalent.

Definition. A partially ordered set (briefly, a poset) is a nonempty set P together with a relation $\leq$ that satisfies

1. $x \leq x$ for all x.
2. If $x \leq y$ and $y \leq z$, then $x \leq z$.
3. If $x \leq y$ and $y \leq x$, then $x = y$.

For example, the set of integers with the usual $\leq$ relation is a poset. The set of subsets of any set together with the inclusion relation $\subseteq$ is a poset. The positive integers with the $|$ relation (recall that $x \mid y$ means that y is divisible by x) are a poset. A chain C is a subset of a poset P if $x \leq y$ or $y \leq x$ for any $x, y \in C$. The integers themselves form a chain under $\leq$ but not under $|$ since $3 \leq 5$ by neither 3 nor 5 is divisible by the other.

In any poset, we write $x < y$ when $x \leq y$ and $x \neq y$. We also use $y \geq x$ synonymously with $x \leq y$.

A subset S of a poset P is said to have an upper bound if there exists an element $p \in S$ such that $s \leq p$ for all $s \in S$. An upper bound p for S is called a least upper bound (lub) provided that if q is any other upper bound for S then $p \leq q$. Similarly l is a greatest lower bound for S if $l \leq s$ for all $s \in S$, and if $m \leq s$ for all $s \in S$ then S then $m \leq l$.

Consider the poset $\mathcal{P}(X)$ of all subsets of a set X under inclusion $\subseteq$. For $S, T \in \mathcal{P}(X)$, the set $S \cup T$ is an upper bound for S and T since $S \subseteq S \cup T$ and $T \subseteq S \cup T$; clearly, any other set that contains both S and T must also contain $S \cup T$, so $S \cup T$ is the least upper bound for S and T.

Least upper bounds (when they exist) are unique. Indeed, if z and z' are both least upper bounds for x and y, then since z is an upper bound and z' is least, we have $z' \leq z$. Reversing the roles of z and z', we have $z \leq z'$, so $z = z'$.

Denote the (unique) least upper bound, if it exists, by $x + y$. Thus, $x \leq x + y$, $y \leq x + y$, and if $x \leq z$, $y \leq z$ then $x + y \leq z$. Similarly $x \cdot y$ (or simply xy) denotes the greatest lower bound of x and y if it exists.

A nonempty poset P in which $x + y$ and xy exist for every pair of elements $x, y \in P$, is called a lattice (the symbols $\vee$ and $\wedge$ are often used as notation instead of $+$ and $\cdot$). A lattice is called distributive if $x(y + z) = xy + xz$ holds for all x, y, z. A poset P has a least element (denoted by 0) if $0 \leq x$ for all $x \in P$ and a greatest element (denoted by 1) if $x \leq 1$ for all $x \in P$. It is evident that, when they exist, 0 and 1 are unique. A lattice L with 0 and 1 is complemented if for each $x \in L$ there exists an element $y \in L$ such that $x + y = 1$ and $xy = 0$. The elements x and y are called complements of one another.

Definition. A Boolean algebra is a poset B that is a complemented distributive lattice. That is, B is a poset that satisfies

1. For each $x, y \in B$, $x + y$ and xy both exist.
2. 0 and 1 exist.
3. $x(y + z) = xy + xz$ is an identity in B.
4. For each $x \in B$ there exists $y \in B$ such that $x + y = 1$ and $xy = 0$.

We shall soon see that, in a Boolean algebra, complements are unique. Thus, for each $x \in B$, $\bar{x}$ will denote the complement of x.

B. Examples

Example. The two-element poset **2**, which consists of $\{0, 1\}$ together with the relation $0 < 1$, is a Boolean algebra.

Example. Let X be a nonempty set and $\mathcal{P}(X)$ the set of all subsets of X. Then $\mathcal{P}(X)$, with $\subseteq$ as the relation, is a Boolean algebra in which $S + T = S \cup T$, $ST = S \cap T$, $0 = \emptyset$, $1 = X$, and $\bar{S} = X - S$, where $X - S = \{x \in X \mid x \in S\}$. Note that if X has only one member then $\mathcal{P}(X)$ has two elements, $\emptyset$, X, and is essentially the same as the first example.

Diagrams can sometimes be drawn to show the ordering of the elements in a Boolean algebra. Each node represents an element of the Boolean algebra. A node that is connected to one higher in the diagram is "less than" the higher one (Fig. 1).

Example. Let X be a set that is possibly infinite. A subset S of X is cofinite if its complement $X - S$ is finite. The set of all subsets of X that are either finite or cofinite form a Boolean algebra under inclusion.

The preceding example is a special case of the following.

Example. A field of subsets is a nonempty collection $\mathcal{S}$ of subsets of a set $X \neq \emptyset$ that satisfies

1. If $S, T \in \mathcal{S}$, then $S \cap T$ and $S \cup T$ are also in $\mathcal{S}$.
2. If $S \in \mathcal{S}$, then so is $X - S$.

A field of subsets is a Boolean algebra under the inclusion relation, and it is easy to see that

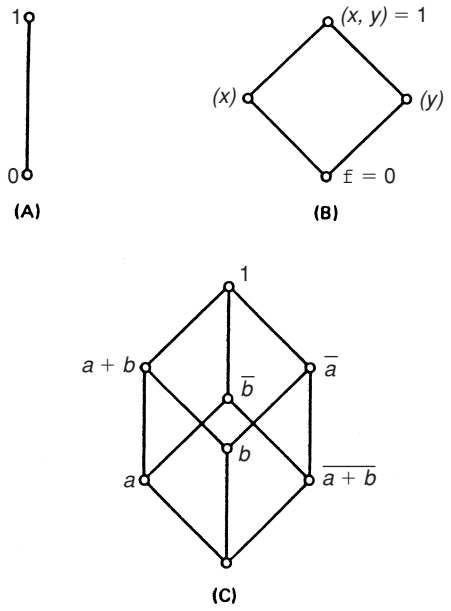


FIGURE 1 Ordering of elements in a Boolean algebra. (A), **2**; (B), $\mathcal{P}(\{x, y, \})$; (C), $\mathcal{P}(\{x, y, z\})$, $a = \{x\}$, $b = \{y\}$.

$$\begin{aligned} S + T &= S \cup T \\ ST &= S \cap T \\ 0 &= \emptyset \\ 1 &= X \\ \bar{S} &= X - S \end{aligned}$$

Note that $\emptyset$ and X are in $\mathcal{S}$ because $\mathcal{S} \neq 0$, so there exists $S_0 \in \mathcal{S}$. Hence, $\bar{S}_0 \in \mathcal{S}$ and so $\emptyset = S_0 \cap S_0 \in \mathcal{S}$. Also $X = X - \emptyset \in \mathcal{S}$.

This is an important example because the elements of the Boolean algebra are themselves sets rather than abstract symbols.

Example. Let $X \neq \emptyset$ be a topological space. A set S is regular open if $S = \text{INT}(\text{CL}(S))$. The set $\mathcal{S}$ of all regular open subsets of X forms a Boolean algebra under inclusion and

$$\begin{aligned} S + T &= \text{INT}(\text{CL}(S \cup T)) \\ ST &= S \cap T \\ S &= X - \text{CL(S)} \\ 0 &= \emptyset \\ 1 &= X \end{aligned}$$

Note that $\mathcal{S}$ is *not* a field of sets since $S \cup T$ may not be a regular open set.

C. Elementary Properties

The following theorem contains several elementary but useful formulas.

Theorem. In a Boolean algebra B,

1. The following are equivalent:
   (i) $x \leq y$; (ii) $x + y = y$; (iii) $xy = x$.
2. $(x + y) + z = x + (y + z)$; $(xy)z = x(yz)$.
3. $x + y = y + x$; $yx = xy$.
4. $x + x = x$; $xx = x$.
5. $x + xy = x$; $x(x + y) = x$.
6. If $x \leq y$, then $x + z \leq y + z$ and $xz \leq yz$.
7. $xy \leq u + v$ if and only if $x\bar{u} \leq \bar{y} + v$; in particular, $xy = 0$ if and only if $x \leq \bar{y}$, and $x \leq y$ if and only if $\bar{y} \leq \bar{x}$.
8. $\overline{x + y} = \bar{x}\bar{y}$; $\overline{xy} = \bar{x} + \bar{y}$ (de Morgan's laws).
9. $x + yz = (x + y)(x + z)$.
10. Complements are unique.
11. $\bar{\bar{x}} = x$.

Proof. We select some of these properties to prove. First, in 1, we prove that (i) implies (ii). Given $x \leq y$, we want to show that y is the least upper bound of x and y. But this is immediate from the definition of least upper bound. Indeed, $x \leq y$ and $y \leq y$ show that y is an upper bound for x, y. Also if $x \leq z$ and $y \leq z$ for some z, then $y \leq z$, so y is the *least* upper bound.

For the first part of 6, simply note that $y + z$ is an upper bound for x and z (since $x \leq y$) so the *least* upper bound $x + z$ of x and z must satisfy $x + z \leq y + z$.

For 7, first observe from the definition of the greatest lower bound and the least upper bound that $x + 1 = 1$ and $x1 = x$ for all x. Now if $xy \leq u + v$, then

$$\begin{aligned} x\bar{u} &= (x\bar{u})1 = (x\bar{u})(y + \bar{y}) \\ &= (x\bar{u})y + (x\bar{u})\bar{y} \leq x(\bar{u}y) + \bar{y} = x(y\bar{u}) + \bar{y} \\ &= (xy)\bar{u} + \bar{y} \leq (u + v)\bar{u} + \bar{y} \\ &= (u\bar{u} + v\bar{u}) + \bar{y} \leq (0 + v) + \bar{y} = v + \bar{y} \end{aligned}$$

To prove the first part of 8, we show that $\overline{x + y}$ is the greatest lower bound for $\bar{x}$ and $\bar{y}$. Now since $x, y \leq x + y$, we have $\overline{x + y} \leq \bar{x}$ and $\overline{x + y} \leq \bar{y}$, by 7, so $\overline{x + y}$ is a lower bound for $\bar{x}$ and $\bar{y}$. But if $z \leq \bar{x}$ and $z \leq \bar{y}$, then $x \leq \bar{z}$ and $y \leq \bar{z}$, so $x + y \leq \bar{z}$ and thus $z \leq \overline{x + y}$.

Formula 9 is just the "distributive law" with the $+\cdot$ interchanged, and it indeed holds since $(x + y)(x + z) = (x + y)x + (x + y)z = xx + yx + xz + yz = x + xy + xz + yz = x + yz$.

Finally, statement 10, that complements are unique, means that if there exist z_1 and z_2 satisfying $x + z_1 = 1$,

$xz_1 = 0$ and $x + z_2 = 1$, $xz_2 = 0$, then $z_1 = z_2$. Now to prove this, $z_1 = z_1 1 = z_1(x + z_2) = z_1 x + z_1 z_2 = 0 + z_1 z_2 \leq z_2$. Similarly, $z_2 \leq z_1$ so $z_1 = z_2$.

In a Boolean algebra B, any finite set $S = \{x_1, \ldots, x_n\}$ has a least upper bound $(\ldots((x_1 + x_2) + x_3) + \cdots) + x_n$. This can be proved by mathematical induction, and in fact the order and parentheses do not matter. That is, $x_1 + (x_2 + x_3)$, $(x_2 + x_1) + x_3$, etc., all represent the same element, namely, the least member of B that is greater than or equal to x_1, x_2, and x_3. Thus, we use the notation $x_1 + \cdots + x_n$ or $\sum S$ or

$$\sum_{i \in (1,\ldots,n)} x_i$$

for the least upper bound of a nonempty set $S = \{x_1, \ldots, x_n\}$. Similarly, $x_1 \ldots x_n$, $\prod S$, and

$$\prod_{i \in (1,\ldots,n)} x_i$$

represent the greatest lower bound of S. All of the properties listed above generalize in the obvious way. For example,

1. $\prod S \leq s \leq \sum S$ for all $s \in S$.
2. $x + \prod S = \prod\{x + s \mid s \in S\}$.
3. $\overline{\sum S} = \prod\{\bar{s} \mid s \in S\}$, $\overline{\prod S} = \sum\{\bar{s} \mid s \in S\}$.

Note that, if $S = \emptyset$, then 0 satisfies the criteria for being a least upper bound for the set S. So we extend the difinitions of $\sum$ and $\prod$ to $\sum \emptyset = 0$, $\prod \emptyset = 1$.

As mentioned above, Boolean algebras can be defined as algebraic systems

Definition. A Boolean algebra is a set B with two binary operations $+$, $\cdot$, a unary operation $-$, and two distinguished, but not necessarily distinct elements 0, 1 that satisfies the following identities:

1. $(x + y) + z = x + (y + z)$; $(xy)z = x(yz)$.
2. $x + y = y + x$; $xy = yx$.
3. $x + x = x$; $xx = x$.
4. $x + xy = x$; $x(x + y) = x$.
5. $x(y + z) = xy + xz$; $x + yz = (x + y)(x + z)$.
6. $0 + x = x$; $1x = x$.
7. $x + \bar{x} = 1$; $x\bar{x} = 0$.

A more concise definition of a Boolean algebra is as an idempotent ring $(R, +, \cdot, 0)$—in the classical algebraic sense—with unity. In this case, the least upper bound is $x + y + xy$; the greatest lower bound is xy; 0 and 1 play their usual role; and $\bar{x} = 1 - x$.

Now that we have two definitions of a Boolean algebra, we must show that they are equivalent. Let us start with the poset definition of B. From the theorem in the last section, it is easy to see that by considering the $+$ and $\cdot$ as operations, rather than as least upper and greatest lower bounds, all of the above identities are true. On the other hand, suppose we start with a Boolean algebra B defined algebraically. We can then define a relation $\leq$ on B by $x \leq y$ if and only if $xy = x$. This relation is in fact a partial order on B since (1) $xx = x$ implies $x \leq x$; (2) if $x \leq y$ and $y \leq z$, then $xy = x$, $yz = y$, so $xz = (xy)z = x(yz) = xy = x$, so $x \leq z$; and (3) if $x \leq y$ and $y \leq x$, then $xy = x$, $yx = y$, so $x = xy = yx = y$. If we continue in this way, $x + y$ turns out to be the least upper bound of x and y under the $\leq$ that we have defined, xy is the greatest lower bound, 0 is the least element, 1 is the greatest, and $\bar{x}$ is the complement of x. Thus B, under $\leq$, is a Boolean algebra.

Since both of these points of view have advantages, we shall use whichever is most appropriate. As we shall see, no confusion will arise on this account.

From the algebraic points of view, the definition of **2** and $\mathcal{P}(\{x, y\})$, shown in Fig. 1, could be given by Table I.

D. Subalgebras

A (Boolean) subalgebra is a subset of a Boolean algebra that is itself a Boolean algebra under the original operations.

A subset B_0 of a Boolean algebra that satisfies the conditions 1–3 below also satisfies the four definitions of a Boolean algebra. This justifies the following:

Definition. A subalgebra B_0 of a Boolean algebra B is a subset of B that satisfies

1. If $x, y \in B_0$, then $x + y$ and $xy \in B_0$.
2. If $x \in B_0$, then $\bar{x} \in B_0$.
3. $0, 1 \in B$.

Every subalgebra of a Boolean algebra contains 0 and 1; $\{0, 1\}$ is itself a subalgebra. The collection of finite and cofinite subsets of a set X forms a subalgebra of $\mathcal{P}(X)$.

TABLE I Operator Tables for 2 and $\mathcal{P}(\{x, y\})$

Boolean Algebra **2**

$+$	0	1
0	0	1
1	1	1

$\cdot$	0	1
0	0	0
1	0	1

$-$	0	1
	1	0

Boolean algebra $\mathcal{P}(\{x, y\})$, where $a = \{x\}$, $b = \{y\}$

$+$	0	a	b	1
0	0	a	b	1
a	a	a	1	1
b	b	1	b	1
1	1	1	1	1

$\cdot$	0	a	b	1
0	0	0	0	0
a	0	a	0	1
b	0	0	b	b
1	0	a	b	1

$-$	0	a	b	1
	1	b	a	0

Now suppose that S is a subset of a Boolean algebra B. Let $S^- = \{\bar{s} \mid s \in S\}$. Then the set of all elements of the form $x = \sum_{i=1}^{n} \prod S_i$, where the S_i are finite nonempty subsets of $S \cup S^-$, form a subalgebra of B. In fact, this set is the smallest subalgebra of B that contains S. It is denoted by $[S]$ and is called the subalgebra of B generated by S. For $x \in B$, $[\{x\}] = \{0, x, \bar{x}, 1\}$; $[\{0, 1\}] = \{0, 1\}$ and $[B] = B$.

E. Ideals and Congruence Relations

Definition. Let B be a Boolean algebra. An ideal I is a nonempty subset of B such that

1. If $x \leq y$, $y \in I$, then $x \in I$.
2. If $x, y \in I$, then $x + y \in I$.

If $I \neq B$, then I is called a proper ideal.

Example. The set I of finite subsets of $\mathcal{P}(X)$ forms an ideal.

Example. $\{0\}$ and B are both ideals in any Boolean algebra B.

Let B be a Boolean algebra. For $b \in B$, the set $\{x \in B \mid x \leq b\}$ is called a principal ideal and is denoted by (b). Now let S be any subset of B. The set $\{x \in B \mid x \leq s_1 + \cdots + s_n, s_i \in S\}$ [denoted by (S)] is the smallest ideal in B that contains S. It is called the ideal generated by S. Of course, if $S = \{b\}$, then $(b) = (\{b\})$.

Ideals are closely related to the standard algebaric notion of a congruence relation. A congruence relation on a Boolean algebra B is a relation Θ that satisfies

1. $(x, x) \in \Theta$ for all $x \in B$.
2. $(x, y) \in \Theta$ implies $(y, x) \in \Theta$.
3. $(x, y)\,(y, z) \in \Theta$ implies$(x, z) \in \Theta$.
4. $(x, y), (x_1, y_1) \in \Theta$ implies $(x + x_1, y + y_1)$, $(xx_1, yy_1) \in \Theta$.
5. $(x, y) \in \Theta$implies $(\bar{x}, \bar{y}) \in \Theta$.

A congruence ralation Θ on B determines a Boolean algebra $B/\Theta = \{[x]_\Theta \mid x \in B\}$, where $[x]_\Theta$ is the congruence class determined by Θ. There is a one-to-one correspondence between the set of all congruences on B and set of all ideals in B. The correspondence is given by $\Theta \to I$, where $x \in I$ if $(x, 0) \in \Theta$. The inverse function $I \to \Theta$ is defined by $(x, y) \in \Theta$ if $x \oplus y \in I$, where $x \oplus y = x\bar{y} + \bar{x}y$ Thus, B/Θ is sometimes written as B/I, where I is the ideal corresponding to Θ. In this notation, B/I is a Boolean algebra in which

$$x/I + y/I = (x + y)/I$$
$$x/I \cdot y/I = (xy)/I$$
$$(\overline{x/I}) = \bar{x}/I$$

and

$$x/I = y/I$$

if there exists $i_0 \in I$ such that

$$x + i_0 = y + i_0$$

F. Homomorphisms

Definition. Let B and B_1 be Boolean algebras. A (Boolean) homomorphism is a function $f : B \to B_1$ that preserves the operations; that is,

1. $f(x + y) = f(x) + f(y)$.
2. $f(xy) = f(x)f(y)$.
3. $f(\bar{x}) = \overline{f(x)}$.
4. $f(0) = 0$, $f(1) = 1$.

Homomorphisms preserve order; that is, if $x \leq y$, then $f(x) \leq f(y)$. But a function that preserves order need not be a homomorphism. Condition 3 is actually a consequence of 1, 2, and 4 since $f(\bar{x})$ is the complement of $f(x)$ and complements are unique.

Example. Let $\mathcal{F}$ be a field of subsets of a set X (see Section I.B) and select an element $p \in X$. Define a function $f : \mathcal{F} \to 2$ by

$$f(S) = \begin{cases} 0 & \text{if} \quad p \notin S \\ 1 & \text{if} \quad p \in S \end{cases}$$

To see that f is a homomorphism, we first show that f preserves order. If $S_1 \subseteq S_2$ and $f(S_2) = 1$, then $f(S_1) \leq 1 = f(S_2)$. But if $f(S_2) = 0$, then $p \in S_2$, so $p \in S_1$, since $S_1 \subseteq S_2$. Therefore, $f(S_1) = 0 = f(S_2)$. Since f preserves order, we have $f(S_1) \leq f(S_1 \cup S_2)$ and $f(S_2) \leq f(S_1 \cup S_2)$, so $f(S_1) + f(S_2) \leq f(S_1 \cup S_2)$. To prove $f(S_1 \cup S_2) \leq f(S_1) + f(S_2)$, suppose that $f(S_1) + f(S_2) = 0$. Then since $\underline{2}$ consists only of 0 and 1, we have $f(S_1) = f(S_2) = 0$, so $p \notin S_1$ and $p \notin S_2$. Hence, $p \notin S_1 \cup S_2$, so $f(S_1 \cup S_2) = 0$. Next, we show $f(S_1)f(S_2) \leq f(S_1 \cap S_2)$; the reserve inclusion follows since f preserves order. If $f(S_1 \cap S_2) = 0$, then $p \notin S_1 \cap S_2$, so either $p \notin S_1$ or $P \notin S_2$. Hence, $f(S_1) = 0$ or $f(S_2) = 0$. In any case, $f(S_1)f(S_2) = 0$. Since 3 is a consequence of 1, 2, and 4 we turn to 4. Now $p \notin \emptyset$ and $p \in X$, so $f(\emptyset) = 0$ and $f(X) = 1$.

Example. Let B be a Boolean algebra and $a \in B$, $a \neq 0$. The ideal (a) is a Boolean algebra (not a subalgebra!) in which, if we denote the least upper bound by $*$, the greatest lower bound by $\times$, and the complement by $^{-a}$, we have for $x, y \in (a]$,

$$x * y = x + y$$
$$x \times y = xy$$
$$\bar{x}^a = a\bar{x}$$

Also, 0 is the least element of $(a]$ and a the greatest element.

The function $f : B \to (a)]$, $f(x) = ax$ is a homomorphism.

Example. Let I be an ideal in a Boolean algebra B. Then $v : B \to B/I$, $v(x) = x/I$ is a homomorphism.

Definition. A homomorphism that is one to one and onto is called an isomorphism.

If there exists an isomorphism f from B to B^1, then it is easy to see that f^{-1} is also an isomorphism. In this case, B and B^1 are said to be isomorphic. This is an important concept because isomorphic Boolean algebras are identical from the point of view of Boolean algebras. That is, there is a one-to-one correspondence between their elements and any property that can be expressed in terms of $\leq, +, \cdot, 0, 1, -$ that is true about one is also true about the other. For example, the Boolean algebra **2** (see Table I) and the field of all subsets of the set $\{p\}$ are isomorphic under the correspondence $0 \leftrightarrow \emptyset$, $1 \leftrightarrow \{p\}$. In fact, we will see later that, for any positive integer n, there is only one (up to isomorphism) Boolean algebra with 2^n elements. Put another way, every Boolean algebra with n elements is isomorphic with the set of all subsets of $\{1, \ldots, n\}$.

The well-known "homomorphism theorem" of classical ring theory applies to Boolean algebras and can be formulated as follows. If $f : B \to B_1$ is an onto homomorphism, then B_1 is isomorphic with B/K, where K is the ideal $K = \{x \in B \mid f(x) = 0\}$. Also let L_1 be a subalgebra of L and I an ideal in L. Then $L_1^1 = \{a \oplus x \mid a \in L_1, x \in I\}$ is a subalgebra of L, and $L_1/(L_1 \cap I)$ is isomorphic with L_1^1/I. (Here, $a \oplus x = \bar{a}x + a\bar{y}$.)

G. Duality

Let P be a mathematical statement concerning Boolean algebras. The dual of P is the statement obtained from P by interchanging the $+$ and $\cdot$ operations and also interchanging the 0's and 1's, where they occur. There are two duality principles that play a role in the study of Boolean algebras. The first—weak duality—is a consequence of the fact that, in the algebraic definition of a Boolean algebra (Section I.C), all of the identities appear along with their dual statements. This implies that, if P is a statement that is true in all Boolean algebras, the dual of P is also true in all Boolean algebras. For example, the statement "If $x + y = \bar{z}$, then $xz = 0$" is always true in any Boolean algebra. Hence, its dual "$xy = \bar{z}$ implies $x + x = 1$" is also true in all Boolean algebras.

The strong-duality principle is that, if a statement is true in a particular Boolean algebra B, its dual is also true in B. The reason for this is that B is "anti-isomorphic" with itself; that is, $x \leq y$ if and only if $\bar{y} \leq \bar{x}$. For example, if there exists an element $b \neq 0$ in a Boolean algebra B such that $0 \leq x < b$ implies $x = 0$, then by strong duality there must also exist an element $c \neq 1$ in B such that $c < x \leq 1$ implies $x = 1$.

II. REPRESENTATION THEORY

A. The Finite Case

In this section we classify all of the finite Boolean algebras. A useful concept for this purpose is that of an atom.

Definition. Let B be a Boolean algebra. An element $a \in B$ is called an atom if $a \neq 0$ and if $x \leq a$, then $x = 0$ or $x = a$.

Theorem. In a Boolean algebra B, the following are equivalent for an element $a \neq 0$.

1. a is an atom.
2. If $a = x + y$, then $a = x$ or $a = y$.
3. If $a \leq x + y$, then $a \leq x$ or $a \leq y$.

Proof. Suppose first that a is an atom and $a = x + y$. Then since $x \leq a$ and $y \leq a$, we have $\{x, y\} \subseteq \{0, a\}$. But x and y cannot both be 0 since $a \neq 0$, so $x = a$ or $y = a$. Next suppose that 2 holds and $a \leq x + y$. Then $a = a(x + y) = ax + ay$, so $a = ax$ or $a = ay$. Hence, $a \leq x$ or $a \leq y$. Finally, if 3 holds and $x \leq a$, then $a = a(x + \bar{x}) = ax + a\bar{x}$; thus, $a = ax$ or $a = a\bar{x}$. So $a \leq x$ or $a \leq \bar{x}$. But $x \leq a$, so if $a \leq x$ then $x = a$, and if $a \leq \bar{x}$ then $x = xa = 0$.

Theorem. Let B be a finite Boolean algebra. Then B is isomorphic with $\mathcal{P}(A)$, where A is the set of atoms of B.

Proof. For each $b \in B$, let $\varphi(b) = \{a \in A \mid a \leq b\}$. Now $b \to \varphi(b)$ preserves order for if $b_1 \leq b_2$ and $a \in \varphi(b_1)$, then $a \leq b_1 \leq b_2$ so $a \in \varphi(b_2)$. Thus, $\varphi(b_1 \cup \varphi(b_2) \subseteq \varphi(b_1 + b_2)$. For the reverse inclusion, let $a \in \varphi(b_1 + b_2)$. Then $a \leq b_1 + b_2$, but by the previous theorem, $a \leq b_1$ or $a \leq b_2$, so $a \in \varphi(b_1)$ or $a \in \varphi(b_2)$. Hence, $a \in \varphi(b_1) \cup \varphi(b_2)$. It is also easily seen that $\varphi(b_1) \cap \varphi(b_2) = \varphi(b_1 b_2)$, $\varphi(0) = \emptyset$, $\varphi(1) = A$, and $\overline{\varphi(b)} = \varphi(\bar{b})$. So φ is a homomorphism. Now for each b, let $\text{len}(b)$ be the number of elements in the longest chain in B that has b as its largest element. This definition is possible since B is finite. We shall

prove, by induction, on the length of b that $b = \sum \varphi(b)$ for all $b \in B$. If $\text{len}(b) = 1$, then $b = 0$ so $\varphi(0) = \emptyset$. Hence $0 = \sum\{0\} = \sum \varphi(0)$. Suppose that $\sum \varphi(x) = x$ for all x such that $\text{len}(x) < n$. Now suppose $\text{len}(b) = n$. If b is an atom, then $b \in \varphi(b)$, so $b = \sum\{a \in A \mid a \leq b\} = \varphi(b)$, so we can assume that b is not an atom. Then there exists $c \neq b$ and $d \neq b$ such that $b = c + d$. But since $c < b$ and $d < b$, we can conclude that $\text{len}(c) < n$ and $\text{len}(d) < n$. Thus, $c = \sum \varphi(c)$ and $d = \sum \varphi(d)$. It follows that

$$b = c + d = \sum \varphi(c) + \sum \varphi(d)$$
$$= \sum[\varphi(c) + \varphi(d)] = \sum \varphi(c + d) = \sum \varphi(b)$$

To complete the proof first note that, by distributivity, if $S \subseteq A, a \in A$, and $a \leq \sum S$, then $a = a \sum S = \sum\{as \mid s \in S\}$ and each as is either 0 or an atom. Now let $S \in \mathcal{P}(A)$. Then $\sum S \in B$ and $\varphi(\sum S) = \{a \in A \mid a \leq \sum S\} = S$, so φ is onto. Finally, to show that φ is one to one, let $\varphi(b_1) = \varphi(b_2)$; then $b_1 = \sum \varphi(b_1) = \sum \varphi(b_2) = b_2$.

B. General Representation Theory

The representation of finite Boolean algebras as fields of sets can be generalized to arbitrary Boolean algebras. In this section we show how this is done and go farther by showing that, as a category, Boolean algebras and homomorphisms are equivalent to a certain category of topological spaces and continuous functions.

Definition. An ideal I in a Boolean algebra B is maximal provided that

1. $I \neq B$.
2. If J is an ideal and $I \subseteq J \subseteq B$, then $J = I$ or $J = B$.

It is easy to see that a principal ideal $(b]$ is maximal if and only if $\bar{b}$ is an atom. A more general theorem is the following:

Theorem. An ideal I is maximal if and only if $I \neq B$ and $xy \in I$ implies $x \in I$ or $y \in I$.

Proof. First suppose that I is maximal, $xy \in I$, but $x \notin I$ and $y \notin I$. Then the ideal generated by $I \cup \{x\}$ properly contains I and, by hypothesis, must be B. So $1 \in B = (I \cup \{x\}]$ and $1 = i + x$ for some $i \in I$. Thus, $\bar{x} \leq i$ so $\bar{x} \in I$. Similarly, $\bar{y} \in I$ and hence $\overline{xy} = \bar{x} + \bar{y} \in I$. But then $1 = (xy) + (\overline{xy}) \in I$, which implies that $I = B$, a contradiction. So $x \in I$ or $y \in I$.

Now assume the hypothesis of the converse and that there is an ideal J such that $I \subset J \subseteq B$. Let $x \in J - I$. To prove $\bar{B} = J$, let $z \in B$. Then $x(\bar{x}z) = 0 \in I$ and $x \notin I$ so $\bar{x}z \in I$. Hence, $\bar{x}z \in J$. So $x + z = x + \bar{x}z \in J$ and therefore $z \in J$. Thus, $B = J$.

It can also be shown that an ideal I is maximal if and only if B/I is isomorphic with **2** and that the set of homomorphisms of B onto **2** are in one-to-one correspondence with the set of maximal ideals. However, the crucial result about maximal ideals concerns their existence.

Theorem. In a Boolean algebra B every element $b \neq 1$ is contained in a maximal ideal.

Proof. (Note: This proof requires a knowledge of Zorn's lemma.) Let $\mathcal{P}$ be the poset of proper ideals that contain b; $(b]$ is one such ideal. Now if $\mathcal{S}$ is a chain in $\mathcal{P}$, then $\cup \mathcal{S}$ is an upper bound for $\mathcal{S}$ since it is an ideal and, being a union of ideals that do not contain 1, cannot itself contain 1. So by Zorn's lemma, $\mathcal{P}$ has a maximal element, say I. Clearly, I is an ideal that contains b. Now if J is an ideal that contains I, then $b \in J$, so $J \in \mathcal{P}$. But I is a maximal element of $\mathcal{P}$, so $J = I$; hence, I is a maximal ideal.

Let B be a Boolean algebra and $\mathcal{I}$ the set of maximal ideals in B. For each $b \in B$, let $\hat{b} = \{I \in \mathcal{I} \mid b \notin I\}$. It is easy to verify that $\hat{B} = \{\hat{b} \mid b \in B\}$ is a field of sets. Specifically,

$$\hat{b}_1 \cup \hat{b}_2 = b_1 + b_2$$
$$\hat{b}_1 \cap \hat{b}_2 = b_1 b_2$$
$$\hat{0} = \emptyset$$
$$\hat{1} = \mathcal{I}$$
$$\hat{\bar{b}} = \mathcal{I} - \hat{b}$$

Moreover, the assignment $\varphi : b \to \hat{b}$ is an isomorphism. To see that φ is one to one, suppose $b_1 \neq b_2$ and without loss of generality say $b_1 \not\leq b_2$. Then $b_1 + b_2 \neq 1$. So there is a maximal ideal I containing $b_1 + b_2$; thus, $b_2 \in I$ and $b_1 \in I$, so $b_1 \in I$. Thus, $I \in \hat{b}_1$, but $I \notin \hat{b}_2$. Hence, $\hat{b}_1 \neq \hat{b}_2$.

For the topological representation theory, we start with a Boolean algebra B and form a topological space $\mathcal{S}(B)$ as follows. The points of $\mathcal{S}(B)$ are the maximal ideals $\mathcal{I}$ of B, and $\{\hat{x} \mid x \in B\}$ is taken as a basis for the topology. This space is a compact Hausdorff space in which the sets that are both open and closed (these turn out to be exactly $\{\hat{x} \mid x \in B\}$ form a basis. Such a space X is called a Boolean (also Stone) space. Let $\mathcal{B}(X)$ denote the field of open–closed sets of a Boolean space.

Theorem. For a Boolean algebra B, $\mathcal{B}(\mathcal{S}(B))$ is isomorphic with B and for each Boolean space X, $\mathcal{S}(\mathcal{B}(X))$ is homeomorphic with X.

This identification can be extended to a coequivalence $\mathcal{C}$ (in the sense of category theory) that assigns to each Boolean algebra B its corresponding Boolean space $\mathcal{S}(B)$ and for each homomorphism $f : B \to B_1$ the continuous function $f : \mathcal{S}(B_1) \to \mathcal{S}(B)$ defined by $\hat{f}(I) = f^{-1}(I)$,

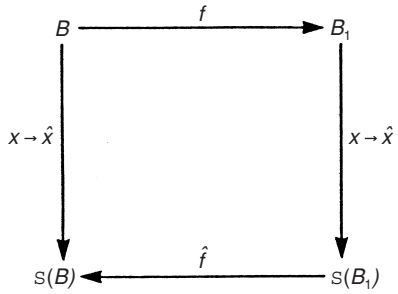


FIGURE 2 Coequivalence between homomorphisms and continuous functions.

$I \in \mathcal{S}(B_1)$. Furthermore, $\hat{f}^{-1}[\hat{x}] = f(x)$ for each $x \in B$ (Fig. 2).

This coequivalence can be very useful since problems in Boolean algebra can be transformed into questions or problems about Boolean spaces that may be more amenable to solution.

III. FREE BOOLEAN ALGEBRAS AND RELATED PROPERTIES

The basic fact about extending functions to homomorphisms is given by the following:

Theorem. Let B, B_1 be Boolean algebras, $\emptyset \neq S \subseteq B$, and $f : S \to B_1$ a function; f can be extended to a homomorphism $g : [S] \to B_1$ if and only if: (l) If $\Pi T_1 \leq \sum T_2$ then $\Pi f[T_1] \leq \sum f[T_2]$, where $T_1 \cup T_2$ is any finite nonempty subset of S. (Note that this includes the cases $\Pi T = 0 \Rightarrow \Pi f[T] = 0$ and $\sum T = 1 \Rightarrow \sum f[T] = 1$, where T is a finite nonempty set).

Sketch of Proof. To define $g : [S] \to B$, we set $g(\sum_{i=1}^{n} \Pi T_i = \sum \Pi f[T_i]$. The condition (1) is sufficient to prove that g is a well-defined homomorphism. Clearly, $g \mid S = f$.

A free Boolean algebra with free generators $\{x_i\}_{i \in I}$ is a Boolean algebra generated by $\{x_i\}_{i \in I}$ and such that any function $f : \{x_i \mid i \in I\} \to B_1$ can be extended to homomorphism. Free Boolean algebras with free generating sets of the same cardinality are isomorphic. For the existence of a free Boolean algebra with a free generating set of cardinality α, we start with a set X of cardinality α. For each $x \in X$, let $S_x = \{S \subseteq X \mid x \in S\}$. Now $\{S_x \mid x \in X\}$ freely generates a subalgebra C of the power set of X. The set $\{S_x \mid x \in X\}$ is said to be independent because if satisfies $\bigcap_{x \in A} S_x \subseteq \bigcup_{y \in B} S_y$ and implies $A \cap B \neq \phi$ for any finite nonempty subset $A \cup B$ of X. But this condition clearly implies that (1) holds, so C is free.

Free Boolean algebras have many interesting properties:

1. Every chain in a free Boolean algebra is finite or countably infinite.
2. Every free Boolean algebra has the property that, if S is a set in which $xy = 0$ for all $x \neq y$ in S, then S is finite or countably infinite. (This is called the countable chain condition).
3. Countable atomless Boolean algebras are free.
4. The finite free Boolean algebras are all of the form $\mathcal{P}(\mathcal{P}(S))$ for finite S.

The Boolean space corresponding to the free Boolean algebra on countably many free generators is the Cantor discontinuum. So for example, the topological version of "Every countable Boolean algebra is projective" is "Every closed subset of the Cantor discontinuum is a retract of itself."

IV. COMPLETENESS AND HIGHER DISTRIBUTIVITY

The canonical work on this topic is "Boolean algebras" by R. Sikorski, and there are many more recent results as well. We present here only a brief introduction.

Definition. A Boolean algebra is complete if $\sum S$ and ΠS exist for all $S \subseteq B$.

For example, $\mathcal{P}(S)$ is complete, as is the Boolean algebra of regular open sets of a topological space. Here one must be careful, for if S is a family of regular open sets, $\cap S$ may not be regular open. It can be shown that $\Pi S = \mathrm{INT}(\cap S)$ and $\sum S = \mathrm{INT}(CL(\cup S))$.

There are many useful ways to generalize the distributive law $x(y + z) = xy = xz$; we will restrict our attention to the following:

Definition. A Boolean algebra is completely distributive provided that, if $\{a_{ij}\}_{i \in I, j \in J}$ is a set of elements such that $\sum_{j \in J} a_{ij}$ exists for each $i \in I$, $\Pi_{i \in I} \sum_{j \in J} a_{ij}$ exists, and $\Pi_{i \in I} a_{i\varphi(i)}$ exists for each $\varphi \in J^I$, then $\sum_{\varphi \in J^I} \Pi_{i \in I} a_{i_\varphi}(i)$ exists and

$$\prod_{i \in I} \sum_{J \in J} a_{ij} = \sum_{\varphi \in J^I} \prod_{i \in I} a_{i\varphi(i)}.$$

A typical result involving these notions is the following:

Theorem. Let B be a Boolean algebra. Then the following are equivalent:

1. B is complete and completely distributive.
2. B is complete and every element is a sum of atoms.
3. B is isomorphic with the field of all subsets of some set.

A striking theorem of Sikorski, from which it follows that the injectives in the category of Boolean algebras are those that are complete, can be formulated as follows:

Theorem. If B_1 is a subalgebra of a Boolean algebra B, A is a complete Boolean algebra, and $f : B_1 \to A$ is a homomorphism, then there exists a homomorphism $g : B \to A$ such that $g \mid B_1 = f$.

Proof. By Zorn's lemma, there exists a homomorphism $f_1 : B_0 \to A$ that is maximal with respect to being an extension of f. We will, of course, be done if $B_0 = B$, so suppose $B_0 \subset B$. Select $a \in B - B_0$ and set

$$b = \sum f_1[\{x \in B_1 \mid x \leq a\} \cap B_0].$$

Now, it can be shown that $a \leq x, x \in B_0 \Rightarrow b \leq f_1(x)$, and $x \leq a, x \in B_0 \Rightarrow f_1(x) \leq b$, so by our theorem on extending functions to homomorphism, f_1 can be extended to $[\{a\} \cup B_0]$, contradicting the maximality of f_1.

Another interesting series of results concerns the embedding of Boolean algebras into ones that are complete. Specifically, a regular embedding $f : B \to B_1$ is a one-to-one homomorphism with the property that, if $\sum S$ exists in B, then $\sum f[S]$ exists in B_1 and is equal to $f(\sum B)$ and similarly for products.

To outline one of the basis results, we shall call an ideal closed if it is an intersection of principal ideals. The set $\bar{B}$ of all closed ideals forms a complete Boolean algebra under inclusion, and the map $v : B \to \bar{B}$, $v(x) = (x]$ is a regular embedding. The pair (B, v) is called the MacNeille completion of B. B is isomorphic with the Boolean algebra of regular open subsets of the Boolean space of B.

V. APPLICATIONS

A. Logic

One of the most importance areas of application for Boolean algebras is logic. Indeed, many problems in logic can be reformulated in terms of Boolean algebras where a solution can be more easily obtained. For example, the fundamental compactness theorem for the (classical) propositional calculus is equivalent to the theorem that every element $a \neq 1$ in a Boolean algebra is contained in a maximal ideal.

Thus, Boolean algebras can be used as models for logical systems. A striking example of this is the simplification of the independence proofs of P. J. Cohen by means of Boolean models. For example, by studying the appropriate Boolean algebra, we can prove that the axiom of choice is not a consequence of set theory—even with the continuum hypothesis.

To illustrate, we start with the set S of formulas $\alpha, \beta, \gamma, \ldots$ of the classical propositional calculus. An equivalence relation $\equiv$ is defined by identifying α and β, provided the $\alpha \to \beta$ and $\beta \to \alpha$ are both derivable. A Boolean algebra is defined on the classes $[\alpha], [\beta], [\gamma], \ldots$ by the partial ordering $[\alpha] \leq [\beta]$ if and only if $\alpha \to \beta$ is derivable.

The resulting Boolean algebra is called the Lindenbaum–Tarski algebra and is in fact the free Boolean algebras with free generators $[\alpha], [\beta], [\gamma], \ldots$, where α, β, γ are the propositional variables. The construction of the Lindenbaum–Tarski algebra for the restricted prelicate calculus is similar, and it is in this context that the independence proofs can be given.

B. Switching Functions and Electronics

Switching functions are mathematical formulas that enable us to describe and design certain parts of electronic circuits. They are widely used in both communications and computer applications.

First, we shall define some notation. We use 2^n to represent the set of all n-tuples of elements in $\mathbf{2}$. That is, $\mathbf{2}^2 = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, and so on. For $\sigma \in \mathbf{2}^n$, we sometimes write $\sigma = (\sigma(1), \ldots, \sigma(n))$, so for $\sigma = (0, 1) \in \mathbf{2}^2$, we have $\sigma(1) = 0$ and $\sigma(2) = 1$.

A switching function is any function $f : \mathbf{2}^n \to \mathbf{2}$. For example, if $n = 2$, there are 16 switching functions, the values of which are given in Table II.

Explicity, f_7 is defined by $f_7(0, 0) = 0$; $f_7(0, 1) = 1$; $f_7(1, 0) = 1$; $f_7(1, 1) = 0$.

For $x \in \mathbf{2}$ (that is, $x = 0$ or $x = 1$), we write $x^0 = \bar{x}$ and $x^1 = x$. Now for a fixed value of n, a switching function is called a complete product if it has the form $f(x_1, \ldots, x_n) = x_1^{\sigma(1)} \ldots x_n^{\sigma(n)}$, where $\sigma \in 2^n$. For example, if $n = 3$, then $f(x_1, x_2, x_3) = x_1 \bar{x}_2 x_3$ is a complete product, but $f(x_1 x_2 x_3) = \bar{x}_2 x_3$ is not.

TABLE II The 16 Switching Functions for $n = 2$

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
(0, 0)	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
(0, 1)	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
(1, 0)	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
(1, 1)	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1

TABLE III Table for Constructing Disjunctive Normal Forms for $f(0,0)\equiv 1, f(0,1)\equiv 1, f(1,0)\equiv 0, f(1,1)=1$

σ	$f(\sigma)$	Complete product	$f(\sigma)x_1^{\sigma(1)}x_2^{\sigma(2)}$
(0, 0)	1	$\bar{x}_1\bar{x}_2$	$1\cdot\bar{x}_1\bar{x}_2$
(0, 1)	1	$\bar{x}_1 x_2$	$1\cdot\bar{x}_1 x_2$
(1, 0)	0	$x_1\bar{x}_2$	$0\cdot x_1\bar{x}_2$
(1, 1)	1	$x_1 x_2$	$1\cdot x_1 x_2$

It turns out that every switching function can be written as a sum of complete products. Indeed, if $f:\mathbf{2}^n\to\mathbf{2}$ is any switching function, it can be written in the form.

$$f(x_1,\ldots,x_n)=\sum_{\sigma\in 2^n} f(\sigma)x_1^{\sigma(1)}\ldots x_n^{\sigma(n)}.$$

To illustrate how to do this, suppose $f(0,0)=1$, $f(0,1)=1$, $f(1,0)=0$, and $f(1,1)=1$. Now, form the table represented by Table III. The products in the third column are formed by putting the bar over the ith factor when the ith component of σ is 0.

Next add the terms $f(\sigma)x^{\sigma(1)}x^{\sigma(2)}$ to obtain $f(x_1,x_2)=1\bar{x}_1\bar{x}_2+1\bar{x}_1x_2+0x_1\bar{x}_2+1x_1x_2=\bar{x}_1\bar{x}_2+\bar{x}_1x_2+x_1x_2$. This last form is called the disjunctive normal form, but f can be simplified further since $\bar{x}_1\bar{x}_2+\bar{x}_1x_2+x_1x_2=\bar{x}_1(x_2+x_2)+x_1x_2=\bar{x}_1+x_1x_2=(\bar{x}_1+x_1)(\bar{x}_1+x_2)=\bar{x}_1+x_2$. That is, $f(x_1x_2)=\bar{x}_1+x_2$.

Thus, starting with the values of f, we have shown how to represent f as a switching function in disjunctive normal form. On the other hand, starting with an f defined as a function of $x_1,\ldots,x_n$, it is easy to represent it in disjunctive normal form. Simply multiply each term by $x_i+\bar{x}_i$ for each missing variable x_i. An example will clarify the procedure. To put $x_1\bar{x}_2+x_3$ in disjunctive normal form, write

$$\begin{aligned}x_1\bar{x}_2+x_3 &= x_1\bar{x}_2(x_3+\bar{x}_3)+x_3(x_1+\bar{x}_1)(x_2+\bar{x}_2)\\ &= x_1\bar{x}_2x_3+x_1\bar{x}_2\bar{x}_3+x_1x_2x_3+\bar{x}_1\bar{x}_2x_3\\ &\quad+\bar{x}_1x_2x_3+\bar{x}_1\bar{x}_2x_3\\ &= x_1\bar{x}_2x_3+x_1\bar{x}_2\bar{x}_3+x_1x_2x_3\\ &\quad+\bar{x}_1x_2x_3+\bar{x}_1\bar{x}_2x_3\end{aligned}$$

For a given n, each disjunctive normal form is a distinct function. So there are $2^{(2^n)}$ distinct switching functions of n variables, all of which are uniquely given in disjunctive normal form as described above.

The application of switching functions to circuit design can be described as follows. Suppose that an electric current is flowing from A to B and x_1, x_2 represent on–off switches (Fig. 3A). If x_1 and x_2 are both off, no current will be detected at B, whereas if *either* x_1 of x_2 is on, current will be detected at B. The circuit is therefore represented by x_1+x_2. Figure 3B represents x_1x_2. In Fig. 3C,

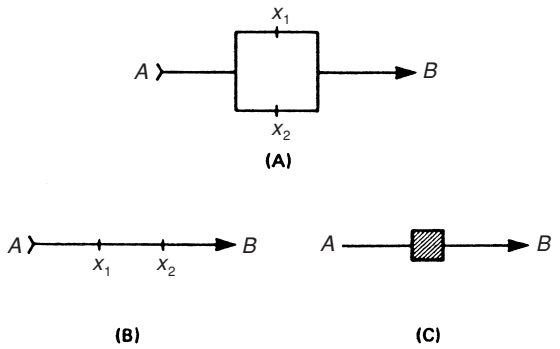


FIGURE 3 Application of switching functions to circuit design. (A) x_1+x_2; (B) x_1x_2; (C)$\bar{x}_1$.

the "black box" inverts the current. That is, if current is flowing at A, no current reaches B, but if current is not flowing from A, current will be detected at B.

To see how switching circuits are used in this context we consider a "half-adder." This is an electronic circuit that performs the following binary arithmetic:

$$\begin{array}{rr} 0 & 0\\ +0 & +1\\ \hline 0\ \text{carry}=0 & 1\ \text{carry}=0\end{array}\qquad\begin{array}{rr} 1 & 1\\ +0 & +1\\ \hline 1\ \text{carry}=0 & 0\ \text{carry}=1\end{array}$$

It is called a half-adder because it does not take into account any carries from previous additions.

Now, if the addends are named A and B, the sum S, and the carry C, Table IV describes the arithmetic. From our previous discussion we can write $S=A\bar{B}+\bar{A}B$ and $C=AB$. The design of the half-adder is shown in Fig. 4.

The expression $A\bar{B}+\bar{A}B$ is quite simple, but in designing more complicated circuits, economies can be made by algebraically simplifying the switching functions. There are several such methods. The Veitch–Karnaugh and Quine methods are examples.

C. Other Applications

Boolean algebras, especially with additional operations, have applications to nonclassical logic, measure theory,

TABLE IV Half-Adder Table

A	B	Sum	Carry
0	0	0	0
0	1	1	0
1	0	1	0
1	1	0	1

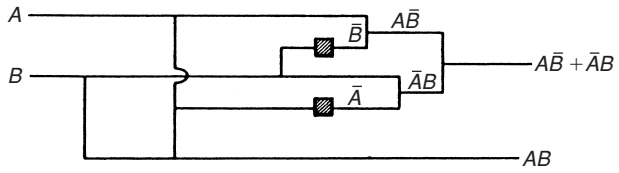


FIGURE 4 Design of the half-adder.

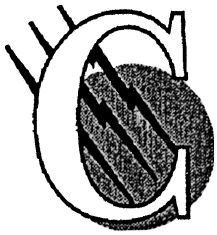
functional analysis, and probability. A detailed summary of these can be found in R. Sikorski's "Boolean Algebras." A concise exposition of nonstandard analysis—via Boolean algebra—can be found in A. Abian's "Boolean Rings."

SEE ALSO THE FOLLOWING ARTICLES

ALGEBRA, ABSTRACT • ALGEBRAIC GEOMETRY • CIRCUIT THEORY • MATHEMATICAL LOGIC • SET THEORY

BIBLIOGRAPHY

Abian, A. (1976). "Boolean Rings," Branden, Boston.
Bell, J. L. (1977). "Boolean Valued Models and Independence Proofs in Set Theory," Clarendon, Oxford.
Birkhoff, G. (1967). "Lattice Theory," *Am. Math. Soc*., Providence.
Brown, F. (1990). "Boolean Reasoning," Kluwer Academic, Boston, Dordrecht, London.
Delvos, F., and Schempp, W. (1989). "Boolean Methods and Interpolation and Approximation," Longman, Harlow, England.
Dwinger, P. (1971). "Inroduction to Boolean Algebras," Physica-Verlag, Wirzburg.
Hailperin, T. (1986). Boole's Logic and Probability. *In* "Studies in Logic and the Foundations of Mathematics" (J. Barwise, D. Kaplan, H. J. Keisler, P., Suppes, and A. S. Troelstra, eds.), North-Holland, Amsterdam.
Halmos, P., and Givant, S. (1998). Logic as Algebra. *In* "The Dolciani Mathematical Expositions," Math. Assoc. of America, Washington, DC.
Johnstone, P. (1983). "Stone Spaces," Cambridge Univ. Press, Cambridge, UK.
Monk, J. D. (1989). "Handbook of Boolean Algebras," Vol. 1–3, Elsevier, New York.
Rosser, J. (1969). "Simplified Independence Proofs," Academic Press, New York.
Schneeweiss, W. (1989). "Boolean Functions with Engineering Applications and Computer Programs," Springer-Verlag, Tokyo.
Sikorski, R. (1969). "Boolean Algebras, 3rd Ed., "Springer-Verlag, Berlin/New York.
Stormer, H. (1990). Binary Functions and Their Applications. *In* "Lecture notes in Economics and Mathematical Systems" (M. Beckmann and W. Krelle, eds.), Springer-Verlag, Berlin, Heidelberg, New York.



Calculus

A. Wayne Roberts

Macalester College

I. Two Basic Problems in Calculus
II. Limits
III. Functions and Their Graphs
IV. Differentiation
V. Integration
VI. Fundamental Theorem of Calculus
VII. Infinite Series
VIII. Some Historical Notes

GLOSSARY

Antiderivative A function F is called the antiderivative of the function f if $F'(x) = f(x)$.

Chain rule Provides a rule for finding the derivative of $f(x)$ if the value of f at x is found by first finding $u = g(x)$, then $h(u)$; that is, $f(x) = h(g(x))$. In this case, if g and h are differentiable, then $f'(x) = h'(g(x))g'(x)$.

Derivative The derivative of a function f is the function f' defined at x by

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

whenever this limits exists.

Differential When the function f has a derivative $f'(x_0)$ at x_0, the differential is the quantity $f'(x_0)h$ where h is a small number; if $y = f(x)$, it is common to represent h by dx and to define the differential to be $dy = f'(x_0)\,dx$.

Exponential function While $f(x) = b^x$ is, for any positive number b, an exponential function, the term is commonly applied to the situation in which $b = e$ where e is the number 2.718 (rounded to three decimals).

Function Rule that assigns to one real number x a unique real number y; one commonly lets a letter f (or g, h, etc.) represent the rule and writes $y = f(x)$. The concept of function is used in more general settings, but always with the idea of assigning to members of one set (the domain of the function) a unique member of another set (the range of the function).

Fundamental theorem Speaking loosely, the two principal ideas of calculus both relate to a graph of a function $y = f(x)$. One is concerned with how to find the slope $f'(x)$ of a line tangent to the graph at any point x; the other asks how to find the area under the graph between $x = a$ and $x = b$. The fundamental theorem relates these seemingly unrelated ideas. A more technical statement of the theorem is to be found below.

Graph If a function f is defined for each real number x in a set, then we may set $y = f(x)$ and plot the points (x, y) in the xy plane to obtain a "picture" corresponding to the function. More formally, the graph of f is the set $G = \{(x, f(x))\colon x \text{ is in the set on which } f \text{ is defined}\}$.

Indefinite integral Another term for the antiderivative.

Infinite series A sequence $\{S_n\}$ in which each term is obtained from the preceding one by adding one term; thus $S_n = S_{n-1} + a_n$. Such a series is commonly represented by $a_1 + a_2 + \cdots$. Terms may be functions $a_n(x)$, so a series may represent a function as well as a constant.

Inflection point If a function f is continuous at x_0, then x_0 is called an inflection point for f if the graph changes from a curve that holds water (convex) to one that sheds water (concave) or conversely at x_0.

Integral Also called the definite integral of f between $x = a$ and $x = b$. It may be described informally as the area under the graph of $y = f(x)$, but such a definition is subject to logical problems. A formal definition of the Riemann integral is given below.

Limit When two variables are related to each other by a clear rule; that is, when $y = f(x)$, we may ask how y is affected by taking values of x closer and closer to some fixed value x_0, or by taking values of x that get larger and larger without bound. If y also gets closer to some fixed y_0, then we say in the two cases described,

$$y_0 = \lim_{x \to x_0} f(x) \qquad \text{or} \qquad y_0 = \lim_{x \to \infty} f(x).$$

Natural logarithm The logarithm of a positive number N to a base $b > 0$ is the exponent r such that $b^r = N$. When the base b is chosen to be the number $e \approx 2.7$, then the logarithm is said to be the natural logarithm of N.

Tangent While there is an intuitive idea of what is meant by a tangent to the graph of $y = f(x)$ at $(x_0, f(x_0))$ (it is the line that just touches the curve at the point), the logically satisfactory way to define it is to say that it is the line through the specified point that has a slope equal to $f'(x_0)$.

CALCULUS is the branch of mathematics that provides computational (or calculating) rules for dealing with problems involving infinite processes. Approximating the circumference of a circle by finding, for larger and larger values of n, the perimeter of a regular inscribed n-sided polygon illustrates an infinite process. Since we cannot hope to complete an infinite process, our only hope is to establish that the longer a certain process is continued, the closer it comes to some limiting position. Thus, calculus is sometimes described as the study of limits, or defined as the discipline that provides algorithms and standard procedures for calculating limits.

I. TWO BASIC PROBLEMS IN CALCULUS

Problem 1. If to each real number x we assign a second number $y = x^2$, the resulting pairs (x, y) may be used as coordinates to locate points relative to a set of perpendicular axes (Fig. 1). The resulting curve is a parabola, a curve studied by the Greeks and frequently encountered in applications of mathematics.

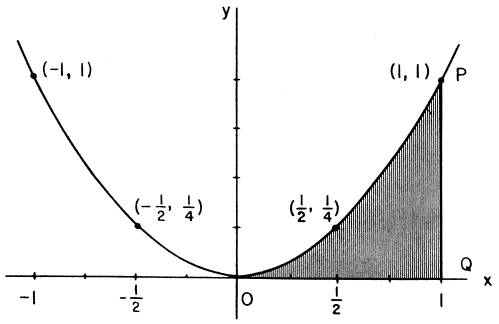


FIGURE 1 The curve is $y = x^2$.

The problem we wish to consider is that of finding the shaded area OPQ under the parabola, above the x axis from 0–1. Archimedes (287–212 B.C.) could have solved this problem using what he called the method of exhaustion. The line of attack we shall use is suggested by Fig. 2, from which we see that the desired area is

(a) less than the area of rectangle A_1 given by

$$R_1 = 1(1)^2 = 1$$

(b) less than the sum of the areas of rectangles B_1 and B_2 given by

$$R_2 = \tfrac{1}{2}\left(\tfrac{1}{2}\right)^2 + \tfrac{1}{2}\left(\tfrac{2}{2}\right)^2$$

(c) less than the sum of the areas of rectangles C_1, C_2, and C_3 given by

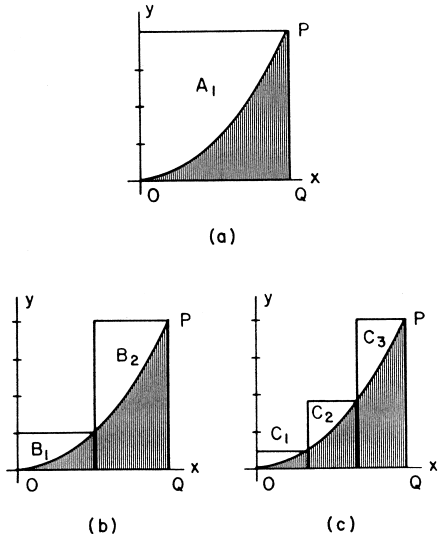


FIGURE 2 The curve is $y = x^2$.

$$R_3 = \tfrac{1}{3}\left(\tfrac{1}{3}\right)^2 + \tfrac{1}{3}\left(\tfrac{2}{3}\right)^2 + \tfrac{1}{3}\left(\tfrac{3}{3}\right)^2$$

Proceeding in this way, we see that after n steps, the desired area will be less than

$$R_n = \frac{1}{n}\left(\frac{1}{n}\right)^2 + \frac{1}{n}\left(\frac{2}{n}\right)^2 + \cdots + \frac{1}{n}\left(\frac{n}{n}\right)^2 \qquad (1)$$

but intuitively, at least, this process should bring us closer and closer to the desired answer.

This is not exactly Archimedes' method of exhaustion, but it is clearly an exhausting process to carry out as n increases. It is, in fact, an infinite process.

Problem 2. Our second problem again makes reference to the graph of $y = x^2$. This time we seek the slope of the line tangent to the graph at $P(1, 1)$. That is, with reference to Fig. 3, we wish to know the ratio of the "rise" BT to the "run" PB (or, stated another way, we wish to know the tangent of the angle that PT makes with the positive x axis). Our method, certainly known to Fermat (1601–1665), is again suggested by a sequence of pictures.

From the graphs in Fig. 4, we see in turn

$$\text{(a)}\quad \text{slope PS}_1 = \frac{B_1S_1}{PB_1} = \frac{(1+1)^2 - 1^2}{1} = 3$$

$$\text{(b)}\quad \text{slope PS}_2 = \frac{B_2S_2}{PB_2} = \frac{(1+1/2)^2 - 1^2}{1/2} = \frac{5}{2}$$

$$\text{(c)}\quad \text{slope PS}_3 = \frac{B_3S_3}{PB_3} = \frac{(1+1/3)^2 - 1^2}{1/3} = \frac{7}{3}$$

And again there is a pattern that enables us to see where we will be after n steps:

$$\text{slope PS}_n = \frac{B_nS_n}{PB_n} = \frac{(1+1/n)^2 - 1^2}{1/n} \qquad (2)$$

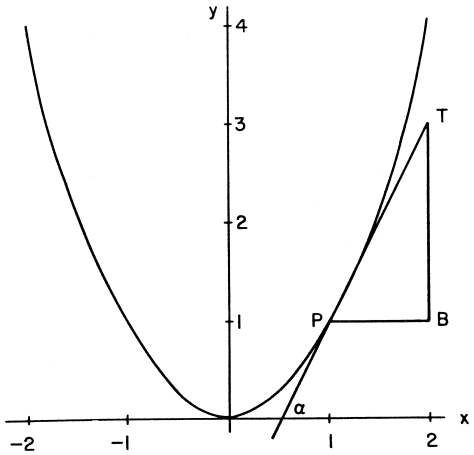


FIGURE 3 The curve is $y = x^2$.

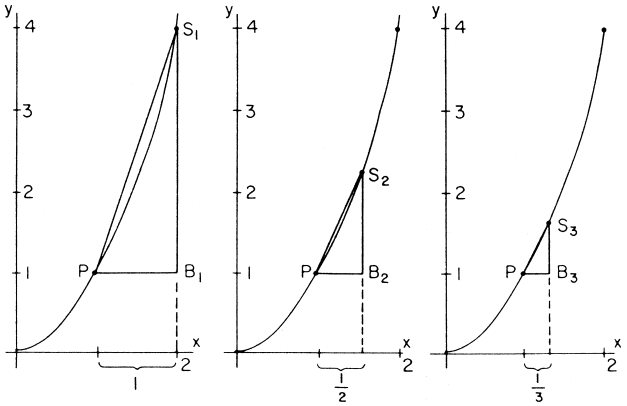


FIGURE 4 The curve is $y = x^2$.

The slope of the desired line is evidently the limiting value of this infinite process.

Our interest in both Eqs. (1) and (2) focuses on what happens as n increases idenfinitely. Using the symbol $\to\infty$ to represent the idea of getting large without bound, mathematicians summarize their interest in Eq. (1) by asking for

$$\lim_{n\to\infty} R_n = \lim_{n\to\infty}\left[\frac{1}{n}\left(\frac{1}{n}\right)^2 + \frac{1}{n}\left(\frac{2}{n}\right)^2 + \cdots + \frac{1}{n}\left(\frac{n}{n}\right)^2\right]$$

and in Eq. (2) by asking for

$$\lim_{n\to\infty} \frac{B_nS_n}{PB_n} = \lim_{n\to\infty} \frac{(1+1/n)^2 - 1^2}{1/n}$$

II. LIMITS

Calculus is sometimes said to be the study of limits. That is because the nature of an infinite process is that we cannot carry it to completion. We must instead make a careful analysis to see if there is some limiting position toward which things are moving.

A. Algebraic Expressions

Limits of certain algebraic expressions can be found by the use of simplifying formulas. Archimedes was greatly aided in his study of the area under a parabola because he had discovered

$$1^2 + 2^2 + \cdots + n^2 = \tfrac{1}{6}n(n+1)(2n+1) \qquad (3)$$

Sometimes the limit of a process is evident from a very minor change of form. The limit of the sum (2) above is easily seen if we write it in the form

$$\lim_{n\to\infty}\frac{B_nS_n}{PB_n}=\lim_{n\to\infty}\frac{(1+1/n)^2-1^2}{1/n}$$
$$=\lim_{n\to\infty}\frac{1+2/n+1/n^2-1}{1/n}$$
$$=\lim_{n\to\infty}(2+1/n)=2 \tag{4}$$

Not every question about limits is concerned with the consequence of a variable growing large without bound. The student of calculus will very soon encounter problems in which a variable approaches zero. A typical question asks what happens to the quotient

$$\frac{\sqrt{x+h}-\sqrt{x}}{h}$$

as h gets closer and closer to zero. Since both the numerator and the denominator approach zero, the effect on the quotient is not obvious. In this case, a little algebraic trickery helps. One notes that

$$\left(\frac{\sqrt{x+h}-\sqrt{x}}{h}\right)\frac{\sqrt{x+h}+\sqrt{x}}{\sqrt{x+h}+\sqrt{x}}=\frac{(x+h)-x}{h(\sqrt{x+h}+\sqrt{x})}$$
$$=\frac{1}{\sqrt{x+h}+\sqrt{x}}$$

In this form, it is clear that as h gets very small, the quotient approaches $1/(2\sqrt{x})$; we write

$$\lim_{h\to0}\frac{\sqrt{x+h}-\sqrt{x}}{h}=\frac{1}{2\sqrt{x}} \tag{5}$$

B. The Natural Base

An investment yielding 8% per year will, for an investment of A dollars, return $A+0.08A=A(1+0.08)$. If the investment is one that adds interest semiannually, then the investment will be worth $A+\frac{1}{2}(0.08)A=A(1+0.08/2)=S$ after 6 months, and by the end of the year it will have grown to $S+\frac{1}{2}(0.08)S=S(1+0.08/2)=A(1+0.08/2)^2=A(1.0816)$.

If interest is added every month (taken for convenience to be $\frac{1}{12}$ of a year), the investment at the end of a year will have grown to

$$A\left(1+\frac{0.08}{12}\right)^{12}=A(1.0830)$$

There is a definite advantage to the investor in having the interest added as frequently as possible.

Jakob Bernoulli (1654–1705) once posed the following problem. An investment of A dollars at $i\%$ that adds interest at the end of each of n periods will yield

$$A(1+i/n)^n$$

Will this sum grow infinitely large if the number n of periods is increased to the point where interest is apparently being added instantaneously?

This apparently frivolous problem is one that involves an infinite process. It is also a problem that, as we shall see below, has profound implications. In order to free ourselves from considerations of a particular choice of i and to focus on the principal question of what happens as n increases, let us agree to set $n=mi$:

$$A(1+i/n)^n=A[(1+1/m)^m]^i$$

Now study what happens as m increases to the expression

$$e_m=\left(1+\frac{1}{m}\right)^m \tag{6}$$

Beginners will be excused if they make the common guess that as m gets increasingly large, e_m gets increasingly close to 1. Even in the face of early evidence provided by the computations

$$e_1=(1+1)=2.00$$
$$e_2=\left(1+\tfrac{1}{2}\right)^2=\tfrac{9}{4}=2.25$$
$$e_3=\left(1=\tfrac{1}{3}\right)^3=\tfrac{64}{27}=2.37$$

beginners find it hard to believe that as m gets larger and larger, e_m will eventually get larger than 2.71, but that it will never get as large as 2.72. They are more incredulous when told that the number that e_m approaches rivals π as one of the most important constants in mathematics, that it has been given the name e as its very own, and that accurate to 10 decimal places, $e=2.7182818285$.

If in Eqs. (6) we replace $1/m$ by h, we get an expression often taken as the definition of e:

$$\lim_{h\to0}(1+h)^{1/h}=e$$

From here it is a short step to a second limit to which we will refer below:

$$\lim_{h\to0}\frac{e^h-1}{h}=1 \tag{7}$$

C. Trigonometric Expressions

A final tricky but important limit must be mentioned. Students in elementary school learn to measure angles in degrees. When the angle x is so measured, then one finds that

$$\frac{\sin5^\circ}{5}=0.017431\qquad\frac{\sin3^\circ}{3}=0.017445$$
$$\frac{\sin1^\circ}{1}=0.017452$$

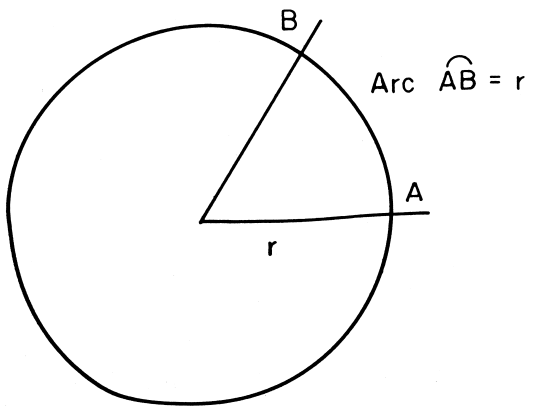


FIGURE 5 Radius of the circle is arbitrary. The angle is 1 radian ≈57°.

and that in general

$$\lim_{x\to 0}\frac{\sin x}{x}=\frac{\pi}{180}=0.017453$$

This awkward number would permeate computations made in calculus, were it not for a uniformly adopted convention. Radians, rather than degrees, are used to measure angles. The central angle of a circle that intercepts an arc of length equal to the radius (Fig. 5) is said to have a measure of one radian. The familiar formula for the circumference of a circle tells us that there will be 2π radians in a full circle of 360°, hence that π radians $= 180°$. Using radians to measure the angle x, it turns out that

$$\lim_{x\to 0}\frac{\sin x}{x}=1 \tag{8}$$

For this reason, radians are always used to measure angles in calculus and in engineering applications that depend on calculus.

III. FUNCTIONS AND THEIR GRAPHS

When the value of one variable y is known to depend on a second variable x, it is common in ordinary discourse as well as in mathematics to say that y is a function of x. An important difference must be noted, however. General usage is commonly imprecise about the exact nature of the relationship (corporate income is a function of interest rates), but mathematicians use the term to imply a precise relationship. For a given value of x, a unique value of y is determined. Mathematicians abbreviate this relationship by writing $y = f(x)$.

Just as a corporation may draw a graph showing how its income has varied with changing interest rates, so a mathematician may draw a graph showing how y varies with changes in x. There are short cuts to drawing such graphs, many of them learned in the study of analytic geometry, which is taught prior to or along with calculus. But even without shortcuts, the method of calculating a table of values and then plotting points can be used to sketch a useful graph. This latter process is illustrated in Table I and Fig. 6 for the function $y = f(x) = 2\sqrt{x} - x^2$.

TABLE I Table of Values

x	$y = 2\sqrt{x} - x^2$
0	0
$\frac{1}{4}$	$2\left(\frac{1}{2}\right) - \frac{1}{16} = \frac{15}{16}$
$\frac{1}{2}$	$2\frac{1}{\sqrt{2}} - \frac{1}{4} \approx 1.2$
1	$2\sqrt{1} - 1 = 1$

IV. DIFFERENTIATION

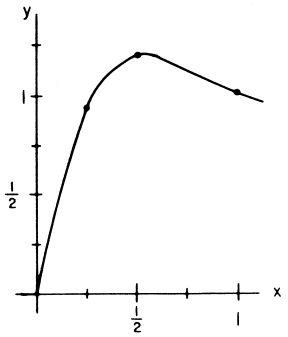
A. The Derivative Defined

Suppose that the distance y that an automobile has travelled in x hours is given by $y = f(x)$. We pose the following question. What does the speedometer read when $x = 2$?

As a first step, we might find the average speed during the third hour by computing

$$\frac{(\text{distance at } x=3)-(\text{distance at } x=2)}{\text{time elapsed}}=\frac{f(3)-f(2)}{3-2}$$

It is clear, of course, that the average speed over the third hour hardly tells us the speedometer reading at $x = 2$. To get even a reasonable approximation to the speedometer reading at $x = 2$, we would want an average over a much smaller time interval near $x = 2$. Let h represent some fraction of an hour after $x = 2$. The average speed over this time interval would be



x	y
0	0
$\frac{1}{4}$	$\frac{15}{16}$
$\frac{1}{2}$	1.2 (approx.)
1	1

FIGURE 6 This is the graph of $y = 2\sqrt{x} - x^2$, drawn by plotting from the table at the right.

$$\frac{(\text{distance at } x = 2 + h) - (\text{distance at } x = 2)}{(2 + h) - 2}$$
$$= \frac{f(2 + h) - f(2)}{h}$$

No information would be gained by setting $h = 0$, but the smaller the positive number h is taken, the better will be our estimate of the speed at $x = 2$. And more generally, the speedometer reading at time x will be approximated by

$$\text{speed at time } x = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h} \tag{9}$$

Consider now a situation in which water runs at 8 in.^{3}/sec into a cone-shaped container that is 12 in. across at the top and 12 in. deep (Fig. 7). The height y of the water rises quickly at first, less quickly as the level rises. In fact, using the formula for the volume of a cone, we can actually determine that the height y at time x will be $y = 2\sqrt[3]{12x/\pi}$, but for our purposes here it suffices to notice that

$$y = f(x)$$

We now ask how fast, in inches per second, y is increasing when $x = 2$. One way to get started is to find the average rate at which y increased during the third minute.

$$\frac{(\text{height at } x = 3) - (\text{height at } x = 2)}{\text{elapsed time}} = \frac{f(3) - f(2)}{3 - 2}$$

Since the rate is slowing down as time passes, however, it is clear that we would get a better approximation to the rate at $x = 2$ if we used a smaller time interval near $x = 2$. Again we let h represent some short interval of time after $x = 2$. Again we see that the desired rate at $x = 2$ will be given by the limit of

$$\frac{(\text{height at } x = 2 + h) - (\text{height at } x = 2)}{h}$$
$$= \frac{f(2 + h) - f(2)}{h}$$

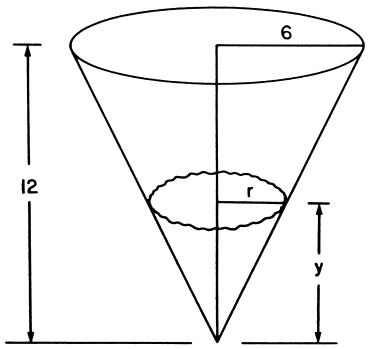


FIGURE 7 Water at height y in cone.

TABLE II Meaning of the Derivative

Given meaning of $f(x)$	Interpretation of $f'(x)$
Distance: $f(x)$ is the distance a moving object travels from some fixed starting point in time x.	Velocity: $f'(x)$ gives the velocity at time x.
Amount: $f(x)$ gives the measure of some changeable quantity (volume of a melting ice cube, length of a shadow) at time x.	Rate of change: $f'(x)$ gives the measure in (units)/(time interval) of the rate at which the quantity is changing.
Graph: The coordinates (x,y) of points on a graph in the plane are related by $y = f(x)$.	Slope of a tangent line: the line tangent to the graph at (x, y) has slope $f'(x)$.

And again for an arbitrary time x, we see that the rate at which height increases is

$$\lim_{h \to 0} \frac{f(x + h) - f(x)}{h} \tag{10}$$

Finally, look again at our introductory Problem 2. Using the notation $f(x) = x^2$ now available to us, we see that Eq. (2) may be written in the form

$$\text{slope PS}_n = \frac{f(1 + 1/n) - f(1)}{1/n}$$

We find the slope of the line tangent to the graph at $x = 1$ by taking the limit of this expression as $n \to 1$. Similar reasoning at an arbitrary point x, making the substitution $h = 1/n$, would lead us to the conclusion that the slope of a line tangent to the graph at x is given by

$$\text{slope} = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h} \tag{11}$$

whenever this limit exists.

The expression that we have encountered in Eqs. (9), (10), and (11) turns up time and again in the applications of mathematics. For a given function f, the derived function or derivative f' is the function defined at x by

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$$

It is important to recognize that the meaning of the derivative is not restricted to just one interpretation. The three important ones that we have discussed are summarized in Table II. Thus, the computational techniques developed by Fermat and other mathematical pioneers have turned out to be of central importance in modern mathematical analysis.

B. Some Derivatives Calculated

We begin with a calculation that can be checked against the answer of 2 obtained in Eq. (4) as the slope of

the line tangent to the graph of $y=x^2$ at $x=1$. For $f(x)=x^2$, $f(x+h)=(x+h)^2=x^2+2xh+h^2$, so

$$\frac{f(x+h)-f(x)}{h}=\frac{x^2+2xh+h^2-x^2}{h}=\frac{h(2x+h)}{h}$$

$$f'(x)=\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} \tag{12}$$

$$=\lim_{h\to 0}(2x+h)=2x$$

The value of $f'(1)=2$, as predicted.

For the function $f(x)=x^3$, we need to use the fact that $f(x+h)=(x+h)^3=x^3+3x^2h+3xh^2+h^3$. Then

$$f'(x)=\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

$$=\lim_{h\to 0}(3x^2+3xh+h^2)=3x^2 \tag{13}$$

For the function $f(x)=\sqrt{x}$, we make use of the previously noted limit in Eq. (5) to write

$$f'(x)=\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

$$=\lim_{h\to 0}\frac{\sqrt{x+h}-\sqrt{x}}{h}=\frac{1}{2\sqrt{x}} \tag{14}$$

Equations (12), (13), and (14) illustrate a general result that Newton discovered by using the binomial theorem (which he also discovered):

$$\text{If}\quad f(x)=x^r\quad \text{where } r \text{ is any real number,}$$

$$\text{then}\quad f'(x)=rx^{r-1}. \tag{15}$$

Two special cases, both easy to prove directly, should be noted here because we need them later. When $r=1$, the formula tells us that the derivative of $f(x)=x$ is $f'(x)=1$; and any constant function $f(x)=c$ may be thought of as $f(x)=cx^\circ$, so that the formula gives the correct result, $f'(x)=0$.

The exponential function $E(x)=e^x$ is important in mathematics and important to an application we shall mention below. Let us use the property of exponents that says $e^{x+h}=e^xe^h$ and Eq. (7) to find

$$E'(x)=\lim_{h\to 0}\frac{e^{x+h}-e^x}{h}$$

$$=\lim_{h\to 0}e^x\frac{e^h-1}{h}=e^x \tag{16}$$

Functions of the form $f(x)=ke^x$ are the only functions that satisfy $f'(x)=f(x)$, and this fact accounts for the great importance of the exponential function in mathematical applications.

Finally, we note that differentiation is a linear operation. This means that if f and g are differentiable functions and r and s are real numbers, then the function h defined by $h(x)=rf(x)+sg(x)$ is differentiable, and $h'(x)=rf'(x)+sg'(x)$.

C. The Derivative and Graphing

The rule just stated, together with our ability to differentiate both $\sqrt{x}$ and x^2, enables us to find the derivative of $f(x)=2\sqrt{x}-x^2$, which is graphed in Fig. 6:

$$f'(x)=2/2\sqrt{x}-2x$$

This means that the tangent line to the graph at $x=\frac{1}{4}$ is $f'(\frac{1}{4})=2-\frac{1}{2}=\frac{3}{2}$; at $x=1$, it is $f'(1)=1-2=-1$. Of more interest is the fact that the continuous function $f'(x)$, in passing from $\frac{3}{2}$ at $x=\frac{1}{4}$ to -1 at $x=1$, must be 0 somewhere between $x=\frac{1}{4}$ and $x=\frac{3}{2}$. Where would $f'(x)=0$? Evidently this occurs at the high point of the graph, and herein lies another great source of applications. The maximum or minimum of a function can often be found by finding the value of x for which $f'(x)=0$. For the example at hand, setting

$$1/\sqrt{x}-2x=0$$

gives

$$1/\sqrt{x}=2x$$

Squaring both sides and multiplying through by x gives $1=4x^3$, or $x=\sqrt[3]{\frac{1}{4}}\approx 0.63$. That is not a value that would be easily guessed by plotting points. The information now in hand enables us to draw a much better graph of $y=2\sqrt{x}-x^2$ (see Fig. 8).

The connection between the values of $f'(x)$ and the slope of lines tangent to the graph of $y=f(x)$ enables us to prove the so-called mean value theorem. This theorem says that if two points on the graph of $y=f(x)$, say $(a, f(a))$ and $(b, f(b))$ are connected by a line segment, then at some point c located between $x=a$ and $x=b$, there will be a tangent to the graph that is parallel to the line (Fig. 9). The formal statement of this theorem is as follows.

Mean value theorem. Let f be differentiable on an interval that includes a and b in its interior. Then there is a point c between a and b such that

$$\frac{f(b)-f(a)}{b-a}=f'(c)$$

Suppose f' is positive on a certain interval. Then if $x_1<x_2$, the mean value theorem tells us that

$$f(x_2)-f(x_1)=f'(c)(x_2-x_1)>0$$

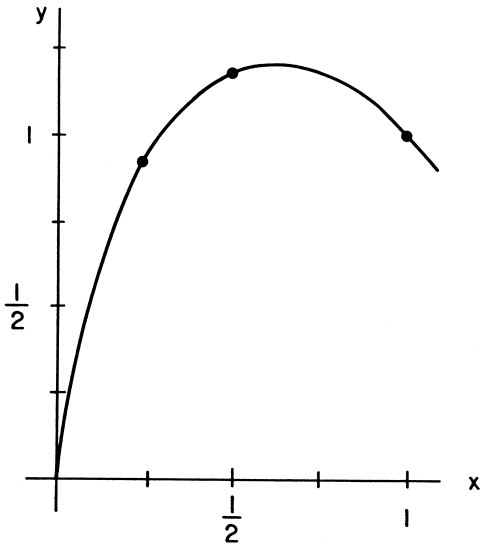


FIGURE 8 The maximum occurs at (0.63, 1.19). This is an improvement on Fig. 6 illustrating that calculus provides information enabling us to improve our graph.

We conclude (Fig. 10) that f must be increasing on the interval. Similar reasoning leads to the conclusion that if $f'(x) < 0$ on an interval, f is decreasing on the interval.

If f is differentiable at a point $x = a$ where its graph achieves a relative high point, then since f swithces from an increasing to a decreasing function at that point, $f'(x)$ switches from positive to negative at $x = a$. We conclude that $f'(a) = 0$ (Fig. 10). It can similarly be shown that a differentiable function will have a derivative of 0 at a relative low point.

An analysis of the graph of $y = S(x) = \sin x$ can help us anticipate what $S'(x)$ might be. In Fig. 11, note the tangent segments sketched at the high point H, low point

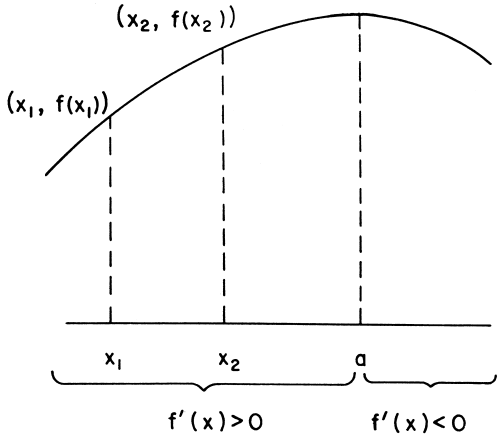


FIGURE 10 The function increases when $f'(x) > 0$, decreases when $f'(x) < 0$.

L, and the points P, Q, and R where the graph of $y = \sin x$ crosses the x axis. The slopes at H and L are clearly 0, meaning that $S'(\pi/2) = S'(3\pi/2) = 0$. We have plotted these points on the second set of axes in anticipation of graphing $y = S'(x)$.

We also note from the symmetry of the graph of $y = \sin x$ that if the slope at P is k, then the slope at Q will be $-k$ and the slope at R will be k again. The appropriate points are again plotted on the second set of axes.

A person familiar with trigonometric functions who thinks about passing a curve through the plotted points might well guess the $S'(x) = \cos x$. Verification of this result depends on use of the addition formulas for the sine function and, as has been mentioned previously, on the use of radian measure, which allows the use of Eq. (8).

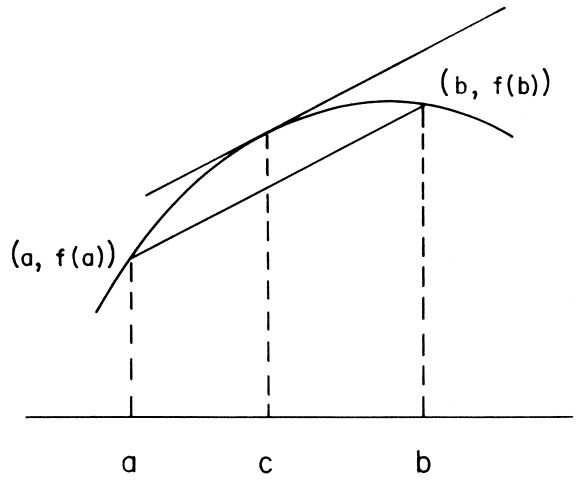


FIGURE 9 The tangent line is parallel to the secant line.

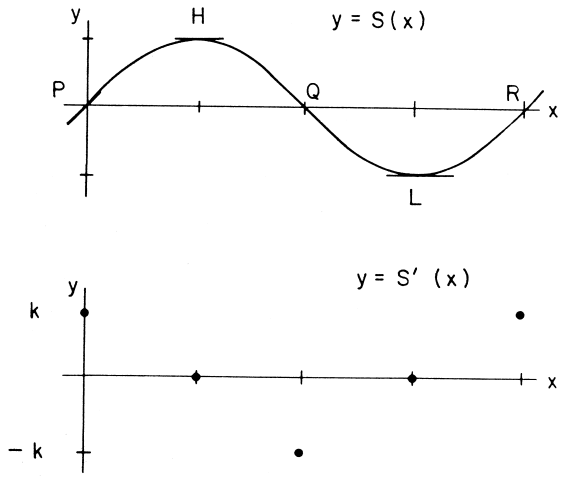


FIGURE 11 The top curve is a sine wave.

D. Four Important Applications

1. Falling Bodies

We have seen that if $y = f(x)$ describes the distance y that a body moves in time x, then $f'(x)$ gives the velocity at time x. Let us take this another step. Suppose the velocity is changing with time. That rate of change, according to the second principle in Table II, will be described by the derivative of $f'(x)$, designated by $f''(x)$. The change of velocity with respect to time is called acceleration; acceleration at time x is $f''(x)$.

When an object is dropped, it picks up speed as it falls. This is called the acceleration due to gravity. There is speculation as to how Galileo (1564–1642), with the measuring instruments available to him, came to believe that the acceleration due to gravity is constant. But he was right; the acceleration is designated by g and is equal to 32.2 ft/sec^2.

Now put the two ideas together. The acceleration is $f''(x)$; the acceleration is constant; $f''(x) = -g$. We use $-g$ because the direction is down. What function, if differentiated, gives the constant $-g$? We conclude that $f'(x) = -gx + c$. The c is included because the derivative of a constant is 0, so we can only be sure of $f'(x)$ up to a constant. Since $f'(x)$ is the velocity at time x, and since $f'(0) = c$, the constant c is usually designated by v_0, which stands for the initial velocity—allowing the possibility that the object was thrown instead of dropped.

The function $f'(x) = -gx + v_0$ is called the antiderivative of $f''(x) = -g$. Now to find the distance y that the object has fallen, we seek the antiderivative of $f'(x)$; what function was differentiated to give $f'(x) = -gx + v_0$? Contemplation of the derivatives computed above leads to the conclusion that

$$y = f(x) = -\tfrac{1}{2}gx^2 + v_0x + y_0$$

this time the constant y_0 represents the initial height from which the object was dropped or thrown.

2. Growth and Decay

Imagine an island nation in which the population p is unaffected by immigration into or out of the country. Let the size of the population at time x be $p = f(x)$. Then according to the second principle of Table II, $f'(x)$ is the rate of growth of the population.

Let us make the reasonable assumption that the rate at which the population grows at time x is proportional to the size of the population at that time. Translated into the language of calculus, this says

$$f'(x) = kf(x) \tag{17}$$

This is a differential equation. We are asked to find a function, the derivative of which is a constant multiple of the function with which we started. If we recall from Eq. (16) that $E'(x) = E(x)$ for $E(x) = e^x$, we can quite easily convince ourselves that a solution to (17) is

$$p = f(x) = e^{kx} \tag{18}$$

This explains why it is said that population grows exponentially.

We need not be talking about people on an island. We might be talking about bacteria growing in a culture, or ice melting in a lake. Anytime the general principle is that growth (or decay) is proportional to the amount present, Eq. (17), perhaps with negative k, describes the action and Eq. (18) provides a solution.

3. Maxima and Minima

The usefulness in practical engineering and scientific problems of our ability to find a high or low point on a graph can be hinted at with the following simple problem. Suppose that a box is to be made from a 3-ft-square piece of sheet metal by cutting squares from each corner, then folding the edges up (see Fig. 12) and welding the seams along the corners. If the edges of the removed squares are x in length, the volume of the box is given by the function

$$y = f(x) = x(3 - 2x)^2 = 4x^3 - 12x^2 + 9x$$

which is graphed in Fig. 13. The derivative is

$$f'(x) = 12x^2 - 24x + 9 = 3(2x - 1)(2x - 3)$$

from which we see that with x restricted by the nature of the problem to be between 0 and $\frac{3}{2}$, the maximum volume is obtained by choosing $x = \frac{1}{2}$.

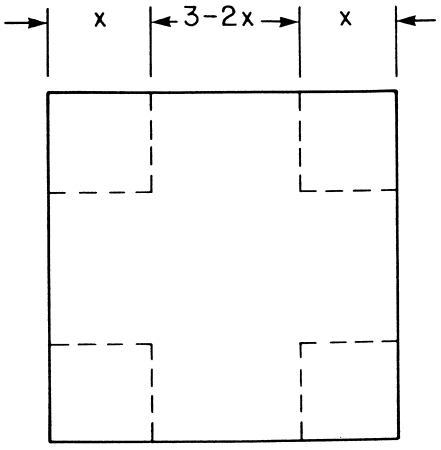


FIGURE 12 Construction of a box.

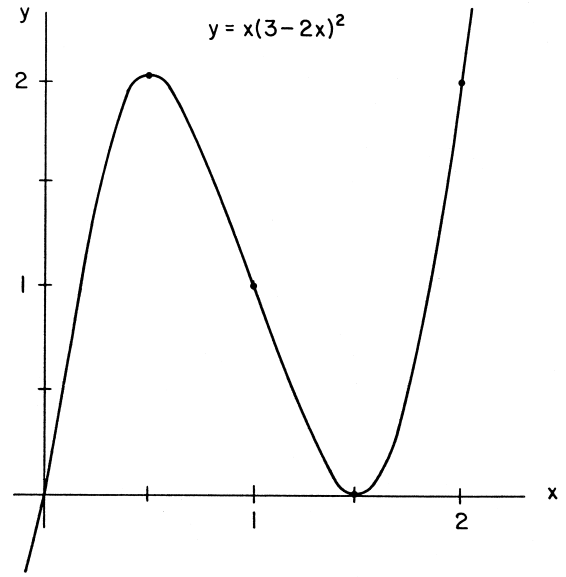


FIGURE 13 Maximum occurs at $(\frac{1}{2}, 2)$.

4. Approximation

Suppose we seek a quick estimate for $\sqrt{9.54}$. We might proceed as follows. The function $f(x) = \sqrt{x}$ has, according to Eq. (14), a derivative of $f'(x) = 1/2\sqrt{x}$. The line tangent to the graph of $y = \sqrt{x}$ at $x = 9$ has a slope of $f'(9) = \frac{1}{6}$ (Fig. 14). The equation of this line is $y - 3 = \frac{1}{6}(x - 9)$.

When x is a little larger than 9, it is clear that the corresponding value of y on the tangent line is going to be larger than the value of y given by $y = \sqrt{x}$. But if x is not too much larger than 9, the two values will be approximately equal. In fact, when $x = 9.54$, the two values are

$$y = 3 + \tfrac{1}{6}(9.54 - 9) = 3.09$$

$$y = \sqrt{9.54} \approx 3.08869$$

This gives us a method of approximating $f(x)$ near a value x_0 where $f(x_0)$ is easily calculated.

We use

$$y = f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

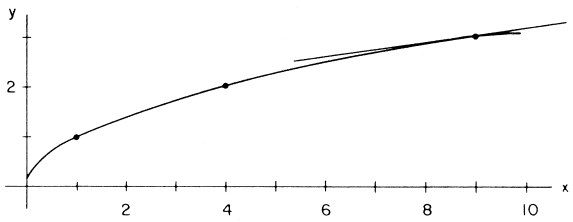


FIGURE 14 The curve is $y = \sqrt{x}$. The line is tangent at (9, 3). Its slope is $\frac{1}{6}$.

There is a school of thought that says that the derivative is best understood as a tool that enables us to approximate $f(x)$ in a neighborhood of x_0 with a linear function.

V. INTEGRATION

A. The Integral Defined

We introduced our discussion of differentiation with a problem about an automobile. We shall do the same in introducing integration.

The acceleration of an automobile is often judged by how fast it can go from a standing position to a speed of 60 mph (88 ft/sec). Consider a car that can do this in 10 sec. Its velocity v is a function of x, the number of seconds since it started: $v = g(x)$, where $g(0) = 0$ and $g(10) = 88$. The question we pose is this. How far will this car have travelled in the 10-sec test?

Distance equals rate times time, but our problem is that the rate keeps changing. We must therefore approximate our answer. One way to approach the problem is to subdivide the time span into n intervals, to pick a time t_i between each x_{i-1} and x_i, and use $g(t_i)$ as the approximate speed during that interval. Then the approximate distance travelled during the time from x_{i-1} to x_i would be $g(t_i)(x_i - x_{i-1})$, and the total distance traveled would be approximated by

$$\text{distance} = g(t_1)(x_1 - x_0) + g(t_2)(x_2 - x_1) + \cdots + g(t_n)(x_n - x_{n-1}) \tag{19}$$

Intuition tells us that smaller subintervals will result in better approximations to the total distance traveled.

We turn our attention now to the subject of work. Physicists define work done in moving an object to be the product of the force exerted in the direction of motion and the distance moved. If a constant force of 15 lb is exerted in pushing an object 20 ft, then it is said that $20 \times 15 = 300$ ft · lb of work has been done.

Anyone who has pushed an object along a floor knows, however, that it takes more force to get the object moving than to keep it moving. In this and other realistic situations (stretching a spring), the force is likely to change as the distance from the starting point changes. We say that the force F is a function of the distance x: $F = f(x)$.

How, when $f = f(x)$, shall we find the work done in moving from $x = a$ to $x = b$? A sensible approximation might be found in this way: Subdivide the segment from a to b into n subintervals. In each subinterval x_{i-1} to x_i, choose a point t_i and use $f(t_i)$ as an approximation of the force exerted over the subinterval (Fig. 15).

The total work done would then be approximated by the sum

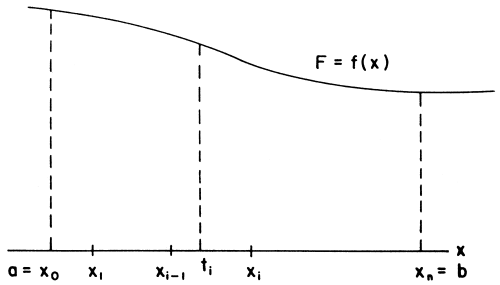


FIGURE 15 Total work done approximation.

$$f(t_1)(x_1 - x_0) + f(t_2)(x_2 - x_1) + \cdots + f(t_n)(x_n - x_{n-1}) \tag{20}$$

Intuition suggests that better approximations can be obtained by taking more intervals of shorter length.

As a final motivating example, we direct attension again to problem 1, the problem of finding the area under the parabolic graph of $y = x^2$ from $x = 0$ to $x = 1$. If we set $f(x) = x^2$, then, corresponding to a partition of the interval into n segments, the area can be approximated (Fig. 16) by the sum

$$f(t_1)(x_1 - x_0) + f(t_2)(x_2 - x_1) + \cdots + f(t_n)(x_n - x_{n-1}) \tag{21}$$

where t_i is chosen to satisfy $x_{i-1} \le t_i < x_i$. Figure 2 shows the areas obtained if we choose $n = 1, 2$, and 3, respectively, if we choose intervals of equal length and if in every case we choose $t_i = x_i$, the right-hand endpoint. If we choose n intervals of equal length so that $x_i - x_{i-1} = 1/n$ for every i, and if we choose $t_i = x_i = i/n$, then Eq. (21) becomes the sum R_n given in Eq. (1).

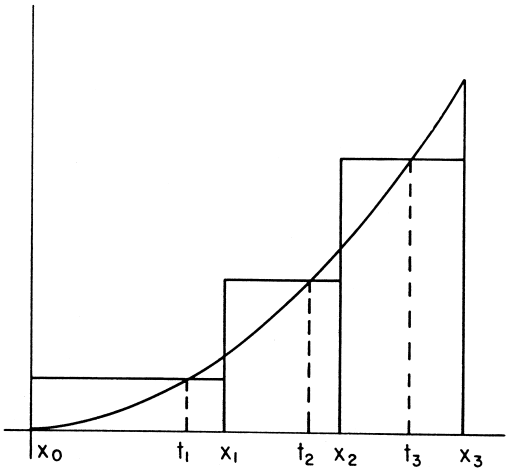


FIGURE 16 The curve is $y = x^2$. The intervals may be unequal, the t_i located differently (not a center, or always the same proportion of the way between x_{i-1} and x_i) within the intervals.

When we write Eq. (1) in the more general form of Eq. (21) and then compare Eq. (21) with Eqs. (19) and (20), we see that the problem Archimedes considered of finding the area under a parabola leads to a computation that comes up in other problems as well. There are, in fact, problems in virtually every area of engineering and science that lead to sums like Eq. (21) above. They are called Riemann sums.

Let us state things formally. A partition of the interval from a to b into n segments is a set P of points:

$$P = \{a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b\}$$

In each subinterval we select a tag t_i: $x_{i-1} \le t_i \le x_i$. Nothing in the theory says that the intervals must be of equal length, or that the t_i must all be chosen in a certain way—a fact we emphasized in Fig. 16. The longest of the subintervals is called the gauge of P, and this number is designated by $|P|$. It is to be noted that when we require $|P|$ to be less than some certain g, then every subinterval of P has length less than g. If g is defined for all x between a and b, then corresponding to a choice of a partition P and a selection of tags $\{t_i\}$, we can form the Riemann sum

$$R(f, P, \{t_i\}) = f(t_1)(x_1 - x_0) + f(t_2)(x_2 - x_1) + \cdots + f(t_n)(x_n - x_{n-1})$$

For a large class of functions, it happens that by specifying a small enough number g, we can guarantee that if $|P| < g$, then the sum $R(f, P, \{t_i\})$ will be very close to some fixed number—and this will be true no matter how the tags $\{t_i\}$ are chosen in the subintervals. Such a function f is said to be Riemann integrable on the interval from a to b. The value around which the sums congregate is designated by an elongated S, stretched between an a and ab, and set before the functional symbol f:

$$\int_a^b f$$

This number is called the integral of f.

Which functions are integrable? Numerous answers can be given. Monotone increasing functions, continuous functions, and bounded functions continuous at all but a finite number of points are all integrable. Suffice to say that most functions that turn up in applied mathematics are integrable. This is also a good place to say that following the work of Lebesque (1875–1941), the integral we have defined has been generalized in numerous ways so as to enlarge the class of integrable functions.

B. Evaluation of Integrals

No matter what the source (distance traveled, work done, etc.) of a Riemann sum, the corresponding integral may be interpreted as an area. Thus, for $f(x) = x$, we easily see (Fig. 17) that

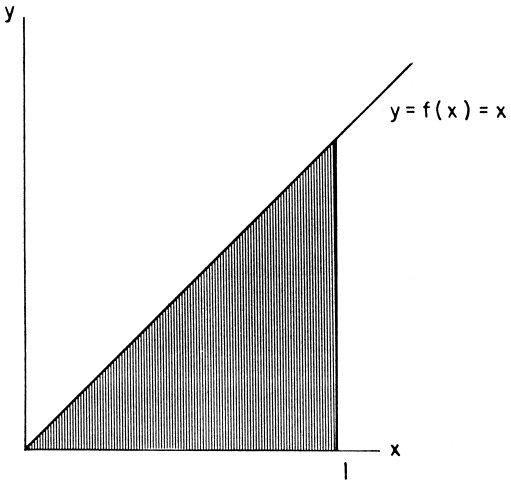


FIGURE 17 Interpreting an integral as an area.

$$\int_0^1 f = \int_0^1 x = \frac{1}{2}$$

When a function is known to be Riemann integrable, then the fact that *all* Riemann sums get close to the correct value for partitions with sufficiently small gauge means that we may use *any* convenient sum. We have seen that for $f(x) = x^2$, one such sum was given by R_n as expressed in Eq. (1):

$$R_n = \frac{1}{n}\left(\frac{1}{n}\right)^2 + \frac{1}{n}\left(\frac{2}{n}\right)^2 + \cdots + \frac{1}{n}\left(\frac{n}{n}\right)^2$$

$$= \frac{1}{n^3}(1^2 + 2^2 + \cdots + n^2)$$

We have already said that Archimedes was helped in his work because he knew Eq. (3), enabling him to write

$$R_n = \frac{1}{n_3}\left[\frac{1}{6}n(n+1)(2n+1)\right]$$

$$= \frac{1}{6}\left(\frac{n}{n}\right)\left(\frac{n+1}{n}\right)\left(\frac{2n+1}{n}\right)$$

$$R_n = \frac{1}{6}\left(1 + \frac{1}{n}\right)\left(2 + \frac{1}{n}\right)$$

Since the partition used here had n equal intervals, the requirement that the gauge get smaller is equivalent to requiring that n get larger. Thus,

$$\int_0^1 x^2 = \lim_{n\to\infty} R_n = \frac{1}{6}(1)(2) = \frac{1}{3}$$

With almost no extra effort, the length of the interval can be extended to an arbitrary b, and it can be determined that

$$\int_0^b x^2 = \frac{b^3}{3}$$

Proceeding in just this way, the mathematician Cavalieri (1598–1647) was able to find formulas similar to Eq. (3) for sums of the form $1^k + 2^k + \cdots + n^k$ for $k = 1, 2, \ldots, 9$. He was thus able to prove that

$$\int_0^b x^k = \frac{b^{k+1}}{k+1} \tag{22}$$

for all positive integers up to 9, and he certainly correctly anticipated that Eq. (22) holds for all positive integers. But each value of k presented a new problem, and as we said, Cavalieri himself stalled on $k = 10$. Reasonable people will be discouraged by the prospect of finding $\int_a^b f$ for more complicated functions. Indeed, the difficulty of such calculations made most such problems intractable until the time of Newton (1642–1727) and Leibniz (1647–1716), when discovery of the fundamental theorem of calculus, discussed below, brought such problems into the realm of feasibility.

The computer age has changed all that. For a given function, it is now possible to evaluate Riemann sums very rapidly. Along with this possibility have come increased emphases on variations of Riemann sums that approximate $\int_a^b f$. These are properly studied in courses on numerical analysis.

Before leaving the topic of evaluation, we note that, like differentiation, integration is a linear operator. For two integrable functions f and g and two real numbers r and s,

$$\int_a^b (rf + sg) = r\int_a^b f + s\int_a^b g$$

Thus, using what we know from our calculations above,

$$\int_0^1 (4x - 2x^2) = 4\int_0^1 x - 2\int_0^1 x^2 = 4\left(\frac{1}{2}\right)$$

$$-2\left(\frac{1}{3}\right) = \frac{4}{3}$$

C. Applications

It seems more difficult to group the diverse applications of integration into major classifications than is the case for differentiation. We shall indicate the breadth of possibilities with several examples typically studied in calculus.

1. Volumes

Suppose the graph of $y = x^2$ is rotated about the x axis to form a so-called solid of revolution that is cut off by the

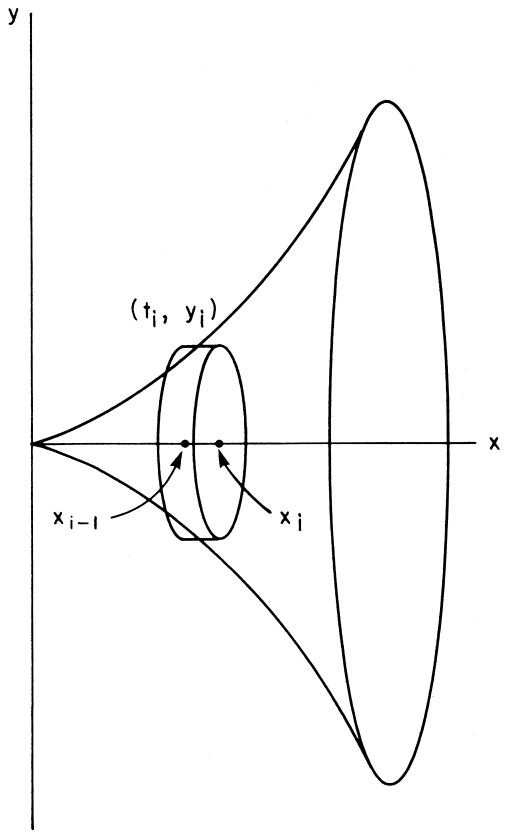


FIGURE 18 The curve $y = x^2$ has been rotated about the x axis to generate a solid figure.

plane $x = 1$ (Fig. 18). A partition of the x axis from 0 to 1 now determines a series of planes that cut off disks, each disk having a volume approximated by $\pi y_i^2(x_i - x_{i-1})$ where $y_i = t_i^2$ for some t_i between x_{i-1} and x_i. Summing these volumes gives

$$\pi t_1^4(x_1 - x_0) + \pi t_2^4(x_2 - x_1) + \cdots + \pi t_n^4(x_n - x_{n-1})$$

which has the form of a Riemann sum for a function $f(x) = \pi x^4$. It converges to an integral that we can evaluate with the help of Eq. (22).

$$\text{Volume} = \int_0^1 \pi x^4 = \pi \int_0^1 x^4 = \pi\left(\frac{1}{5}\right)$$

2. Length of Arc

Suppose the points A and B are connected by a curve that is the graph of $y = g(x)$ for x between a and b (Fig. 19). The length of this curve is clearly approximated by the sum of the lengths of line segments joining points that have as their first coordinates the points of a partition of the x axis from a to b. A typical segment has length

$$l = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}$$
$$= \sqrt{(x_i - x_{i-1})^2 + [g(x_i) - g(x_{i-1})]^2}$$

If the function g is differentiable, then from the Mean Value Theorem, we see that

$$g(x_i) - g(x_{i-1}) = g'(t_i)(x_i - x_{i-1})$$

where t_i is between x_{i-1} and x_i. Thus

$$l = \sqrt{1 + [g'(t_i)]^2}\,(x_i - x_{i-1})$$

and the sum of these lengths is

$$\sqrt{1 + [g'(t_1)]^2}\,(x_1 - x_0) + \cdots$$
$$+ \sqrt{1 + [g'(t_n)]^2}\,(x_n - x_{n-1})$$

This has the form of a Riemann sum for the function $f(x) = \sqrt{1 + [g'(x)]^2}$. It converges to an integral we take to be the desired

$$\text{Length} = \int_a^b \sqrt{1 + [g'(x)]^2}$$

3. Normal Distribution

If in a certain town we measure the heights of all the women, or the IQ scores of all the third graders, or the gallons of water consumed in each single family dwelling, we will find that the readings cluster around a number called the mean, $\bar{x}$. A common display of the readings uses a bar graph, a graph in which the percentage of the readings that fall between x_{i-1} and x_i is indicated by the height of a bar drawn over the appropriate interval (Fig. 20a). The sum of all the heights (all the percentages) should, of course, be 1.

As the size of the intervals is decreased and the number of data points is increased, it happens in a remarkable number of cases that the bars arrange themselves under the so-called normal distribution curve (Fig. 20b) that has an equation of the form

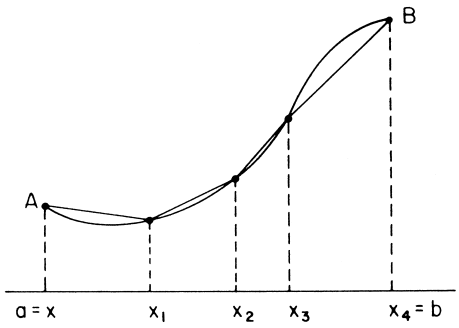


FIGURE 19 An arbitrarily drawn curve, together with a sequence of secant lines being used to (roughly) approximate the length of the curve.

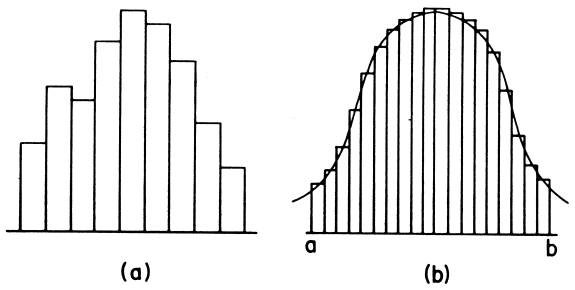


FIGURE 20 A pair of histograms, the second one indicating how an increasing number of columns leads to the concept of a distribution curve.

$$y = de^{-m(x-\bar{x})^2}$$

The constants are related to the relative spread of the bell-shaped curve, and they are chosen so that the area under the curve is 1. The percentage of readings that fall between a and b is then given by

$$\int_a^b de^{-m(x-\bar{x})^2}$$

VI. FUNDAMENTAL THEOREM OF CALCULUS

Up to this point, calculus seems to be neatly separated into two main topics, differential calculus and integral calculus. Historically, the two topics did develop separately, and considerable progress had been made in both areas. The genius of Newton and Leibniz was to see and exploit the connection between the integral and the derivative.

A. Integration by Antidifferentiation

Let $y = f(x)$ be the distance that a moving object has traveled from some fixed starting point in time x. We have seen (Table II) that the velocity of the object at time x is then given by $v = f'(x)$. Since the value of $f'(x)$ is also the slope of the line tangent at (x, y) to the graph of $y = f(x)$, a general sketch of $v = f'(x)$ may be drawn by looking at the graph of $y = f(x)$; see Fig. 21.

From the graph of $y = f(x)$, we see that from time $x = a$ to time $x = b$,

$$\text{Distance traveled} = f(b) - f(a) \qquad (23)$$

At the same time, we argued in setting up Eq. (19) that if the velocity v of a moving object is given by $v = g(t)$, then the distance traveled would be approximated by what we came later to see was a Riemann sum that converged to

$$\int_a^b g$$

Thus, in the present situation with $g = f'$ as the velocity function, we see that from the time $x = a$ to $x = b$,

$$\text{Distance traveled} = \int_a^b f' \qquad (24)$$

Comparison of Eqs. (23) and (24) suggests that

$$\int_a^b f' = f(b) - f(a)$$

The function being integrated is f', the derivative of the function f on the right side of the equal sign. The function f is in turn called the antiderivative of f'. The relationship lies at the heart of the calculus.

Fundamental theorem of calculus. Let F be any antiderivative of f; that is, let F be a function such that $F'(x) = f(x)$. Then

$$\int_a^b f = F(b) - F(a)$$

B. Some Consequences

In Eq. (15), we saw that if $f(x) = x^r$, then $f'(x) = rx^{r-1}$. It follows that the antiderivative of $f(x) = x^r$ would, for $r \neq -1$, be

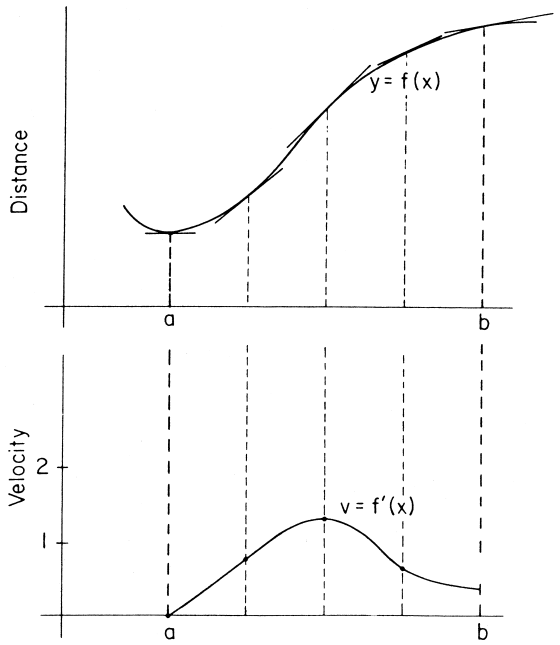


FIGURE 21 A graph of $y = f(x)$ on the top axes with two tangent segments having a slope of 1, an intermediate segment with slope >1, and a low point with slope 0. The lower curve, obtained by plotting the slopes of representative tangent lines, is the graph of $y = f'(x)$.

$$F(x) = \frac{1}{r+1}x^{r+1}$$

Consequently, for any $r \neq -1$ the fundamental theorem gives us

$$\int_a^b x^r = F(b) - F(a) = \frac{1}{r+1}b^{r+1} - \frac{1}{r+1}a^{r+1}$$

For $a = 0$ and r chosen to be a positive integer, we have Cavalieri's result [Eq. (22)].

The case $r = -1$ is of special interest. Since $f(x) = 1/x$ is continuous for $x > 0$,

$$\int_1^b \frac{1}{x}$$

exists for any $b > 0$. Yet the function $f(x) = 1/x$ has no simple antiderivative. Suppose we set

$$L(t) = \int_1^t \frac{1}{x}$$

This function turns out to have the following interesting properties:

$$L'(t) = \frac{1}{t} \qquad L(e^x) = x$$

The second property means that L is the inverse of the exponential function; it exhibits all the properties associated with a logarithm function. Since $L(e) = 1$, we say L is the logarithmic function having e as its base; it is called the natural logarithm.

VII. INFINITE SERIES

In discussing Approximation above, we noted that the value of

$$y = f(x) = \sqrt{x}$$

could be approximated for values of x near 9 by using

$$y = T_1(x) = f(9) + f'(9)(x - 9),$$

the function having as its graph the line tangent at $x = 9$ to the graph of $y = \sqrt{x}$. We might explain the usefulness of T_1 as a way to approximate f near $x = 9$ by observing that T_1 and its first derivative both agree with f at $x = 9$, That is,

$$T_1(9) = f(9) \quad \text{and} \quad T_1'(9) = f'(9)$$

This viewpoint has a wonderfully useful generalization. Consider the function

$$T_2(x) = f(9) + f'(9)(x - 9) + \frac{f''(9)}{2}(x - 9)^2$$

Its first and second derivatives are

$$T_2'(x) = f'(9) + f''(9)(x - 9)$$
$$T_2''(x) = f''(9)$$

This function and its first *two* derivatives agree with f at $x = 9$; that is, we see by substituting in the expressions just given that

$$T_2(9) = f(9), \qquad T_2'(9) = f'(9), \qquad T_2''(9) = f''(9),$$

It seems reasonable, therefore, to guess that T_2 (9.54) will give us a better approximation to $\sqrt{9.54}$ than we got with T_1 (9.54) = 3.09. Indeed, since

$$f''(x) = \left(\frac{1}{2}\right)\left(-\frac{1}{2}\right)x^{-\frac{3}{2}},$$
$$f''(9) = \left(\frac{1}{2}\right)\left(-\frac{1}{2}\right)\frac{1}{3^3} = -\frac{1}{108}$$

and

$$T_2(9.54) = 3 + \frac{1}{6}(0.54) + \frac{1}{2}\left(\frac{-1}{108}\right)(0.54)^2 = 3.08865$$

The actual value correct to five places to the right of the decimal is 3.08869, so we have improved our estimate.

The full generalization is this. Suppose a function and its derivatives at a point x_0 are known or easily calculated. Then the nth degree *Taylor polynomial* is defined to be

$$T_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + \cdots + \frac{f^n(x_0)}{n!}(x - x_0)^n$$

This polynomial can be used to estimate f for values of x near x_0, and the larger n is taken, the better the estimate will be. Taylor polynomials of degree $n = 6$ for some well known functions are seen in Table III.

We are now in a position to get some idea of how calculators determine the values of familiar functions. The following table gives the actual values for the sine and cosine of angles near $x_0 = 0$, and, for comparison, the values of the polynomials $S(x)$ and $C(x)$ defined above. To get the results listed, one must remember that in calculus, angles must be measured in radians.

TABLE III

Function	Taylor polynomial of degree 6
$\sin x$	$S(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!}$
$\cos x$	$C(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!}$
e^x	$E(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^6}{6!}$
Arc $\tan x$	$A(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!}$

TABLE IV

X (degrees)	x(radians)	Sin x	$S(x)$	Cos x	$C(x)$
1	0.0174533	0.0174524	0.0174524	0.9998477	0.9998477
2	0.0349066	0.0348995	0.0348995	0.9993908	0.9993908
3	0.0523599	0.0523360	0.0523360	0.9986295	0.9986295
4	0.0698132	0.0697565	0.0697565	0.9975641	0.9975641
5	0.0872664	0.0871557	0.0871558	0.9961947	0.9961947
10	0.1745329	0.1736482	0.1736483	0.9848078	0.9848077
15	0.2617994	0.2588192	0.2588192	0.9659258	0.9659258
20	0.3490659	0.3420201	0.3420204	0.9396926	0.9396926
25	0.4363323	0.4226183	0.4226190	0.9063078	0.9063077
30	0.5235987	0.5000000	0.5000023	0.8660254	0.8660252
35	0.6108652	0.5735764	0.5735829	0.8191520	0.8191514

In a table of values accurate to seven places to the right of the decimal, one must get to 5 degrees before any difference shows up between $\sin x$ and its polynomial approximation $S(x)$; and the polynomial approximations $S(x)$ and $C(x)$ give six place accuracy for $\sin x$ and $\cos x$ for values of x all the way through 20 degrees.

Though manufacturers of calculators use variations, the idea illustrated in the Table IV gives correct insight into how calculators find values of trigonometric functions; they make use of polynomial approximations. Greater accuracy can be obtained by using Taylor polynomials of higher degree. It is sufficient, of course, to obtain accuracy up to the number of digits to the right of the decimal that are displayed by the calculator.

The last function listed in Table III, $Arc\tan x$, is the inverse tangent function. If $y = Arc\tan x$, then $x = \tan y$. It is included so we can address another question about which people sometimes wonder. Since $\tan 30^\circ = \frac{1}{\sqrt{3}}$ and $30^\circ = \frac{\pi}{6}$ radians, substitution of $x = \frac{1}{\sqrt{3}}$ into the polynomial approximation for $Arc\tan x$ gives

$$\frac{\pi}{6} \approx \frac{1}{\sqrt{3}} - \frac{1}{3(\sqrt{3})^3} + \frac{1}{5(\sqrt{3})^5} - \frac{1}{7(\sqrt{3})^7} = 0.5230$$

Multiplication by 6 gives us the approximation of $\pi \approx 3.138$ Rounding to two decimals to the right of the decimal, we get the familiar approximation of 3.14. More accuracy can be obtained by using a Taylor polynomial of higher degree, and there are tricks that yield much better approximations using just the 7th degree polynomial employed here, but again we stop, having illustrated our point. Familiar constants such as π (and e) can be approximated using Taylor polynomials.

While our discussion of infinite series has given some indication of the usefulness of the idea, it is important to indicate that a great deal more could be said. Thus, while Taylor series provide an important way to represent functions, they should be seen as just one such representation. Fourier series provide another useful representation, and there are others.

Also, while we have approached infinite series by way of representing functions, we might well have started with the representation of individual numbers. The familiar use of $0.333\cdots = \frac{1}{3}$ is really a statement saying that the infinite sum

$$\frac{3}{10} + \frac{3}{10^2} + \frac{3}{10^3} + \cdots$$

gets closer and closer to $\frac{1}{3}$; that is, the finite sums

$$S_n = \frac{3}{10} + \frac{3}{10^2} + \cdots \frac{3}{10^n}$$

get closer to $\frac{1}{3}$ as n, the number of terms increases.

The addition of an infinite number of terms is not a trivial subject. The history of mathematics includes learned discussions of what value to assign to

$$1 - 1 + 1 - 1 + 1 - \cdots$$

Some argued that an obvious grouping shows that

$$(1 - 1) + (1 - 1) + (1 - 1) + \cdots = 0$$

Others countered that

$$1 - (1 - 1) - (1 - 1) - \cdots = 1$$

and Leibniz, one of the developers of calculus, suggested that the proper value would therefore be $\frac{1}{2}$. There are other instances in which famous mathematicians have challenged one another to find the value of an infinite series of constants. James Bernoulli (1654–1705) made it clear that he would like to know the sum of

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots + \frac{1}{n^2} + \cdots.$$

but it wasn't until 1736 that Euler discovered that this sum got closer and closer to $\frac{\pi^2}{6}$

VIII. SOME HISTORICAL NOTES

Archimedes, with his principle of exhaustion, certainly had in hand the notion of approaching a limiting value with a sequence of steps that could be carried on ad infinitum, but he was limited by, among other things, the lack of a convenient notation. The algebraic notation came much later, and the idea of relating algebra to geometry was later still, often attributed to Descartes (1596–1650), but arguably owing more to Fermat (1601–1665). Fermat, Barrow (1630–1677), Newton's teacher, Huygens (1629–1695), Leibniz's teacher, and others set the stage in the first half of the 17th century, but the actual development of the calculus is attributed to Isaac Newton (1642–1727) and Gottfried Wilhelm Leibniz (1646–1716), two geniuses who worked independently and were ultimately drawn into arguments over who developed the subject first.

Evidence seems to substantiate Newton's claim to primacy, but his towering and deserved reputation as one of the greatest thinkers in the history of mankind is surely owing not so much to what he invented, but what he did with it. In Newton's hands, calculus was a tool for changing the way humans understood the universe. Using calculus to extrapolate from his law of universal gravitation and other laws of motion, Newton was able to analyze not only the motion of free falling bodies on earth, but to explain, even predict the motions of the planets. He was widely regarded as having supernatural insight, a reputation the poet Alexander Pope caught with the lines,

Nature and Nature's laws lay hid in night:
God said, "Let Newton be!" and all was light.

Leibniz, on the other hand, developed the notation that made the calculus comprehensible to others, and he gathered around himself the disciples that took the lead away from Newton's followers in England, and made the continent the center of mathematical life for the next century. James (1654–1705) and John (1667–1748) Bernoulli, Leonhard Euler (1707–1783), and others applied calculus to a host of problems and puzzles as well applied problems.

The first real textbook on calculus was *Analyse des Infiniments Petits Pour L'intelligence des Lignes Courbes*, written by the Marquis de l'Hospital in 1696. It wasn't until 1816, however, that a text by Lacroix, *Traité du Calcul Différentiél et du Calcul Intégral*, having been well received in France for many years, was translated into English and so brought the methods used on the continent to the English speaking world. Through all of these books there ran an undercurrent of confusion about the nature of the infinitesimal, and it was not until the work of Cauchy (1789–1857) and Weierstrass (1815–1897) that logical gaps were closed and calculus took the form that we recognize today.

The most recent changes in the teaching of calculus grew out of an effort during the decade from 1985–1995 to reform the way calculus was to be taught in the United States. A summary of that effort is presented in *Calculus, The Dynamics of Change*.

SEE ALSO THE FOLLOWING ARTICLES

ALGEBRAIC GEOMETRY • DIFFERENTIAL EQUATIONS, ORDINARY • DIFFERENTIAL EQUATIONS, PARTIAL • INTEGRAL EQUATIONS • NUMERICAL ANALYSIS • STOCHASTIC PROCESSES

BIBLIOGRAPHY

Apostol, T. M. (1961). "Calculus, Volumes 1 and 2," Blaisdell, Boston.
Boyer, C. B. (1959). "The History of the Calculus and Its Conceptual Development," Dover, New York.
Morgan, F. (1995). "Calculus Lite," A. K. Peters, Wellesley, Massachusetts.
Ostebee, A., and Zorn, P. (1997). "Calculus," Harcourt Brace & Co., San Diego.
Roberts, A. W., ed. "Calculus: The Dynamics of Change," Mathematics Association of America, Washington, DC.
Sawyer, W. W. (1961). "What is Calculus About?," The Mathematical Association of America New Mathematical Library, Washington, DC.
Swokowski, E. W. (1991). "Calculus," 5th ed., PWS-Kent, Boston.
Thomas, G. B., Jr., and Finney, R. L. (1984). "Calculus and Analytic Geometry," 6th ed., Addison-Wesley, Reading, Massachusetts.
Toeplitz, O. (1963). "The Calculus, A Genetic Approach," University of Chicago Press, Chicago.



Complex Analysis

Joseph P. S. Kung
Department of Mathematics, University of North Texas

Chung-Chun Yang
Department of Mathematics, The Hong Kong University of Science and Technology

I. The Complex Plane
II. Analytic and Holomorphic Functions
III. Cauchy Integral Formula
IV. Meromorphic Functions
V. Some Advanced Topics

GLOSSARY

Analytic function A complex function with a power series expansion. An analytic function is holomorphic and conversely.

Argument In the polar form $z = re^{i\theta}$ of a complex number, the argument is the angle θ, which is determined up to an integer multiple of 2π.

Cauchy integral formula A basic formula expressing the value of an analytic function at a point as a line integral along a closed curve going around that point.

Cauchy–Riemann equations The system of first-order partial differential equations $u_x = v_y,\ u_y = -v_x$, which is equivalent to the complex function $f(z) = u(x, y) + iv(x, y)$ being holomorphic (under the assumption that all four first-order partial derivatives are continuous).

Conformal maps A map that preserves angles infinitesimally.

Disk A set of the form $\{z : |z - a| < r\}$ is an open disk.

Domain A nonempty connected open set in the complex plane.

Half-plane The upper half-plane is the set of complex numbers with positive imaginary part.

Harmonic function A real function $u(x, y)$ of two real variables satisfying Laplace's equation $u_{xx} + u_{yy} = 0$.

Holomorphic functions A differentiable complex function. See also Analytic functions.

Laurent series An expansion of a complex function as a doubly infinite sum: $f(z) = \sum_{m=-\infty}^{\infty} a_n(z - c)^m$.

Meromorphic function A complex function that is differentiable (or holomorphic) except at a discrete set of points, where it may have poles.

Neighborhood of a point An open set containing that point.

Pole A point a is a pole of a function $f(z)$ if $f(z)$ is analytic on a neighborhood of a but not at a, $\lim_{z\to a} f(z) = \infty$, and $\lim_{z\to a}(z - a)^k f(z)$ is finite for some positive integer k.

Power series or Taylor series An expansion of a complex function as an infinite sum: $f(z) = \sum_{m=0}^{\infty} a_n(z - c)^m$.

Region A nonempty open set in the complex plane.

Riemann surface A surface (in two real dimensions) obtained by cutting and gluing together a finite or infinite number of complex planes.

TO OVERSIMPLIFY, COMPLEX ANALYSIS is calculus using complex numbers rather than real numbers. Because a complex number $a+ib$ is essentially a pair of real numbers, there is more "freedom of movement" over the complex numbers and many conditions become stronger. As a result, complex analysis has many features that distingush it from real analysis. The most remarkable feature, perhaps, is that a differentiable complex function always has a power series expansion. This fact follows from the Cauchy integral formula, which expresses the value of a function in terms of values of the function on a closed curve going around that point.

Many powerful mathematical techniques can be found in complex analysis. Many of these techniques were developed in the 19th century in conjunction with solving practical problems in astronomy, engineering, and physics. Complex analysis is now recognized as an indispensable component of any applied mathematician's toolkit. Complex analysis is also extensively used in number theory, particularly in the study of the distribution of prime numbers.

In addition, complex analysis was the source of many new subjects of mathematics. An example of this is Riemann's attempt to make multivalued complex functions such as the square-root function single-valued by enlarging the range. This led him to the idea of a Riemann surface. This, in turn, led him to the theory of differentiable manifolds, a mathematical subject that is the foundation of the theory of general relativity.

I. THE COMPLEX PLANE

A. Complex Numbers

A complex number is a number of the form $a+ib$, where a and b are real numbers and i is a square root of -1, that is, i satisfies the quadratic equation $i^2+1=0$. Historically, complex numbers arose out of attempts to solve polynomial equations. In particular, in the 16th century, Cardano, Tartaglia, and others were forced to use complex numbers in the process of solving cubic equations, even when all three solutions are real. Because of this, complex numbers acquired a mystical aura that was not dispelled until the early 19th century, when Gauss and Argand proposed a geometric representation for them as pairs of real numbers.

Gauss thought of a complex number $z=a+ib$ geometrically as a point (a, b) in the real two-dimensional space. This represents the set $\mathbf{C}$ of complex numbers as a real two-dimensional plane, called the *complex plane.* The x-axis is called the *real axis* and the real number a is called the *real part* of z. The y-axis is called the *imaginary axis* and the real number b is called the *imaginary part* of z.

The *complex conjugate* $\bar{z}$ of the complex number $z=a+ib$ is the complex number $a-ib$. The *absolute value* $|z|$ is the real number $\sqrt{a^2+b^2}$. The (*multiplicative*) *inverse* or *reciprocal* of z is given by the formula

$$\frac{1}{z}=\frac{\bar{z}}{|z|}=\frac{a-ib}{\sqrt{a^2+b^2}}.$$

The complex number $z=a+ib$ can also be written in the *polar form*

$$z=re^{i\theta}=r(\cos\theta+i\sin\theta),$$

where

$$r=\sqrt{a^2+b^2}, \qquad \tan\theta=y/x.$$

The angle θ is called the *argument* of z. The argument is determined up to an integer multiple of 2π. Usually, one takes the value θ so that $-\pi<\theta\leq\pi$; this is called the *principal value* of the argument.

B. Topology of the Complex plane

The absolute value defines a *metric* or distance function d on the complex plane $\mathbf{C}$ by

$$d(z_1, z_2)=|z_1-z_2|.$$

This metric satisfies the triangle inequality

$$d(z_1, z_2)+d(z_2, z_3)\geq d(z_1, z_3)$$

and determines a topology on the complex plane $\mathbf{C}$. We shall need the notions of open sets, closed sets, closure, compactness, continuity, and homeomorphism from topology.

C. Curves

A *curve* γ is a continuous map from a real interval $[a, b]$ to $\mathbf{C}$. The curve γ is said to be *closed* if $\gamma(a)=\gamma(b)$; it is said to be *open* otherwise. A *simple closed curve* is a closed curve with $\gamma(t_1)=\gamma(t_2)$ if and only if $t_1=a$ and $t_2=b$.

An intuitively obvious theorem about curves that turned out to be very difficult to prove is the Jordan curve theorem. This theorem is usually not necessary in complex analysis, but is useful as background.

The Jordan Curve Theorem. The image of a simple closed curve (not assumed to be differentiable) separates the extended complex plane into two regions. One region is bounded (and "inside" the curve) and the other is unbounded.

D. The Stereographic Projection

It is often useful to extend the complex plane by adding a point ∞ at infinity. The extended plane is called the *extended complex plane* and is denoted $\bar{\mathbf{C}}$. Using a *stereographic projection*, we can represent the extended plane $\bar{\mathbf{C}}$ using a sphere.

Let $\mathcal{S}$ denote the sphere with radius 1 in real three-dimensional space defined by the equation

$$x_1^2 + x_2^2 + x_3^2 = 1$$

Let the x_1-axis coincide with the real axis of $\mathbf{C}$ and let the x_2-axis coincide with imaginary axes of $\mathbf{C}$. The point $(0, 0, 1)$ on $\mathcal{S}$ is called the *north pole*. Let (x_1, x_2, x_3) be any point on $\mathcal{S}$ not equal to the north pole. The point $z = x + iy$ of intersection of the straight line segment emanating from the north pole N going through the point (x_1, x_2, x_3) with the complex plane $\mathbf{C}$ is the *stereographical projection* of (x_1, x_2, x_3). Going backwards, the point (x_1, x_2, x_3) is the *spherical image* of z. The north pole is the spherical image of the point ∞ at infinity. The stereographic projection is a one-to-one mapping of the extended plane $\bar{\mathbf{C}}$ onto the sphere $\mathcal{S}$. The sphere $\mathcal{S}$ is called the *Riemann sphere*. The stereographical projection has the property that the angles between two (differentiable) curves in $\mathbf{C}$ and the angle between their images on $\mathcal{S}$ are equal.

The mathematical formulas relating points and their spherical images are as follows:

$$x_1 = \frac{z + \bar{z}}{1 + |z|^2}, \quad x_2 = \frac{z - \bar{z}}{i(1 + |z|^2)}, \quad x_3 = \frac{|z|^2 - 1}{1 + |z|^2}$$

and

$$z = \frac{x_1 + ix_2}{1 - x_3}$$

Let z_1 and z_2 be two points in the plane $\mathbf{C}$. The *spherical* or *chordal distance* $\sigma(z_1, z_2)$ between their spherical images on $\mathcal{S}$ is

$$\sigma(z_1, z_2) = \frac{2|z_1 - z_2|}{\sqrt{1 + |z_1|^2}\sqrt{1 + |z_2|^2}}.$$

Let $d\sigma$ and ds be the length of the infinitesimal arc on $\mathcal{S}$ and $\mathbf{C}$, respectively. Then

$$d\sigma = 2(1 + |z|^2)^{-1}\, ds.$$

E. Connectivity

We need two notions of connectedness of sets in the complex plane. Roughly speaking, a set is *connected* if it consists of a single piece. Formally, a set S is connected if S is not the union $A \cup B$ of two disjoint nonempty open (and hence, closed) subsets A and B. A set is *simply connected* if every closed curve in it can be contracted to a point, with the contraction occurring in the set. It can be proved that a set is simply connected in $\mathbf{C}$ if its complement in the extended complex plane is connected. If A is not simply connected, then the connected components (that is, open and closed subsets) of its complement $\bar{\mathbf{C}}$ not containing the point ∞ are the "holes" of A.

A *region* is a nonempty open set of $\mathbf{C}$. A *domain* is a non-empty connected open set of $\mathbf{C}$. When $r > 0$, the set $\{z : |z - c| < r\}$ of all complex numbers at distance strictly less than r from the *center* c is called the *open disk with center c and radius r*. Its closure is the *closed disk* $\{z : |z - c| \le r\}$. Open disks are the most commonly used domains in complex analysis.

II. ANALYTIC AND HOLOMORPHIC FUNCTIONS

A. Holomorphic Functions

Let $f(z)$ denote a complex function defined on a set Ω in the complex plane $\mathbf{C}$. The function $f(z)$ is said to be *differentiable* at the point a in Ω if the limit

$$\lim_{h \to 0} \frac{f(a + h) - f(a)}{h}$$

is a finite complex number. This limit is the *derivative* $f'(a)$ of $f(z)$ at a. Note that the limit is taken over all complex numbers h such that the absolute value $|h|$ goes to zero, so that h ranges over a set that has two real dimensions. Thus, differentiability of complex functions is a much stronger condition than differentiability of real functions. For example, the length of every (infinitesimally) small line segment starting from a is changed under the function $f(z)$ by the same real scaling factor $|f'(a)|$, independently of the angle. The formal rules of differentiation in calculus hold also for complex differentiation.

It is possible to construct complex functions that are differentiable at only one point. To exclude these degenerate cases, we only consider complex functions that are differentiable at every point in a region Ω. Such functions are said to be *holomorphic* on Ω.

B. The Cauchy–Riemann Equations and Harmonic Functions

A complex function $f(z)$ can be expressed in the following way:

$$f(z) = f(x + iy) = u(x, y) + iv(x, y)$$

where $u(x, y)$ is the real part of $f(z)$ and $v(x, y)$ is the imaginary part of $f(z)$. The functions $u(x, y)$ and $v(x, y)$ are real functions of two real variables.

Differentiability of the complex function $f(z)$ can be rewritten as a condition on the real functions $u(x, y)$ and $v(x, y)$. Let $f(z) = u(x, y) + iv(x, y)$ be a complex function such that all four first-order partial derivatives of u and v are continuous in the open set Ω. Then a necessary and sufficient condition for $f(z)$ to be holomorphic on Ω is

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

These equations are the *Cauchy–Riemann equations.* In polar coordinates $z = re^{i\theta}$, the Cauchy–Riemann equations are

$$r\frac{\partial u}{\partial r} = \frac{\partial v}{\partial \theta}, \qquad \frac{\partial u}{\partial \theta} = -r\frac{\partial v}{\partial r}$$

The square of the absolute value of the derivative is the Jacobian of $u(x, y)$ and $v(x, y)$, that is

$$|f'(z)|^2 = \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial u}{\partial y}\frac{\partial v}{\partial x}$$

A *harmonic or potential function* $h(x, y)$ is a real function having continuous second-order partial derivatives in a nonempty open set Ω in $\boldsymbol{R}^2$ satisfying the two-dimensional *Laplace equation*

$$\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} = 0$$

for all $z = x + iy$ in Ω. We shall see that $f(z)$ being holomorphic in Ω implies that every derivative $f^{(n)}(z)$ is holomorphic on Ω, and, in particular, the real functions $u(x, y)$ and $v(x, y)$ have partial derivatives of any order. Hence, by the Cauchy–Riemann equations, $u(x, y)$ and $v(x, y)$ satisfy the Laplace equation and are harmonic functions. The harmonic functions $u(x, y)$ and $v(x, y)$ are called the *conjugate pair* of the holomorphic function $f(z)$.

The two-dimensional Laplace equation governs (incompressible and irrotational) fluid flow and electrostatics in the plane. These give intuitive physical models for holomorphic functions.

C. Power Series and Analytic Functions

Just as in calculus, we can define complex functions using power series. A *power series* is an infinite sum of the form

$$\sum_{m=0}^{\infty} a_m (z - c)^m$$

where the *coefficients* a_m and the *center* c are complex numbers. This series determines a complex number (depending on z) whenever the the series converges. Complex power series work in the same way as power series over the reals.

In particular, a power series has a *radius of convergence,* that is, an extended real number ρ, $0 \le \rho \le \infty$, such that the series converges absolutely whenever $|z - c| < \rho$. The radius of convergence is given explicitly by *Hadamard's formula:*

$$\rho = \frac{1}{\limsup_{n\to\infty} \sqrt[n]{|a_n|}}$$

A function $f(z)$ is *analytic* in the region Ω if for every point c in Ω, there exists an open disk $\{z : |z - c| < r\}$ contained in Ω such that $f(z)$ has a (convergent) power series or *Taylor expansion*

$$f(z) = \sum_{m=0}^{\infty} a_m (z - c)^m$$

When this holds, $f^{(n)}(c)/m! = a_m$.

Polynomials are analytic functions on the complex plane. Other examples of analytic functions on the complex plane are the *exponential function,*

$$e^z = \sum_{m=0}^{\infty} z^n / n!$$

and the two *trigenometric functions,*

$$\cos z = \frac{e^{iz} + e^{-iz}}{2}, \qquad \sin z = \frac{e^{iz} - e^{-iz}}{2i}$$

The inverse under functional composition of e^z is the (*natural*) *logarithm* $\log z$. The easiest way to define it is to put z into polar form. Then

$$\log z = \log re^{i\theta} = \log r + i\theta$$

Since θ is determined up to an integer multiple of 2π, the logarithmic function is multivalued and one needs to extend the range to a Riemann surface (see Section V.D) to make it a function. For most purposes, one takes the value of θ so that $-\pi < \theta \le \pi$. This yields the *principal value* of the logarithm.

III. CAUCHY INTEGRAL FORMULA

A. Line Integrals and Winding Numbers

Line integrals are integrals taken over a curve rather than an interval on a real line. Let $\gamma : [a, b] \to C$ be a piecewise continuously differentiable curve and let $f(z)$ be a continuous complex function defined on the image of γ. Then the *line integral* of $f(z)$ along the path γ is the Riemann integral

$$\int_a^b f(\gamma(t))\gamma'(t)\,dt$$

This integral is denoted by

$$\int_\gamma f(z)\,dz$$

Line integrals are also called *path integrals* or *contour integrals*. Line integrals behave in similar ways to integrals over the real line, except that instead of moving along the real line, we move on a curve.

The *winding number* or *index* $n(\gamma, a)$ of a curve γ *relative to a point* a is the number of time the curve winds or goes around the point a. Formally, we define $n(\gamma; z)$ using a line integral:

$$n(\gamma, a) = \frac{1}{2\pi i}\int_\gamma \frac{d\zeta}{\zeta - a}$$

This definition is consistent with the intuitive definition. One can show, for example, that (a) $n(\gamma, a)$ is always an integer, (b) the winding number of a circle around its center is 1 if the circle goes around in a counterclockwise direction, and -1 if the circle goes around in a clockwise direction, (c) the winding number of a circle relative to a point in its exterior is 0, and (d) if a is a point in the interior of two curves and those two curves can be continuously deformed into one another without going through a, then they have the same winding number relative to a.

B. Cauchy's Integral Formula

In general, line integrals depend on the curve. But if the integrand $f(z)$ is holomorphic, Cauchy's integral theorem implies that the line integral on a simply connected region only depends on the endpoints.

Cauchy's integral theorem. Let $f(z)$ be holomorphic on a simply connected region Ω in $\mathbf{C}$. Then for any closed piecewise continuously differential curve γ in Ω,

$$\int_\gamma f(z)\,dz = 0$$

One way to prove Cauchy's theorem (due to Goursat) is to observe that if the curve is "very small," then the line integral should also be "very small" because holomorphic functions cannot change drastically in a small neighborhood. Hence, we can prove the theorem by carefully decomposing the curve into a union of smaller curves. Another way to think of Cauchy's theorem is that a line integral over a curve γ of a holomorphic function on a region Ω is zero whenever γ can be continuously shrunk to a point in Ω.

Cauchy's integral formula. Let $f(z)$ be holomorphic on a simply connected region Ω in $\mathbf{C}$ and let γ be a simple piecewise continuously differentiable closed path going counterclockwise in Ω. Then for every point z in Ω inside γ,

$$f^{(m)}(z) = \frac{m!}{2\pi i}\int_\gamma \frac{f(\zeta)\,d\zeta}{(\zeta - z)^{m+1}}$$

In particular,

$$f(z) = \frac{1}{2\pi i}\int_\gamma \frac{f(\zeta)\,d\zeta}{\zeta - z}$$

Cauchy's formula for $f(z)$ follows from Cauchy's theorem applied to the function $(f(\zeta) - f(z))/(\zeta - z)$, and the general case follows similarly.

A somewhat more general formulation of Cauchy's formula is in terms of the winding number. If $f(z)$ is analytic on a simply connected nonempty open set Ω and γ is a closed piecewise continuously differentiable curve, then, for every point z in Ω,

$$f^{(m)}(z) = n(\gamma, z)\frac{m!}{2\pi i}\int_\gamma \frac{f(\zeta)\,d\zeta}{(\zeta - z)^{m+1}}$$

Cauchy's integral formula expresses the function value $f(z)$ in terms of the function values around z. Take the curve γ to be a circle $|z - a| = r$ of radius r with center a, where r is sufficiently small so that the circle is in Ω. Using the geometric series expansion

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - a}\sum_{m=0}^{\infty}\left(\frac{z - a}{\zeta - a}\right)^m$$

and interchanging summation and integration (which is valid as all the series converge uniformly), we obtain

$$f(z) = \sum_{m=0}^{\infty}\frac{(z - a)^{m+1}}{2\pi i}\int_{|z-a|=r}\frac{f(\zeta)\,d\zeta}{(\zeta - a)^{m+1}}$$

This gives an explicit formula for the power series expansion of $f(z)$ and shows that a holomorphic function is analytic. Since an analytic function is clearly differentiable, being analytic and being holomorphic are the same property.

C. Geometric Properties of Analytic Functions

Analytic functions satisfy many nice geometric and topological properties. A basic property is the following.

The open mapping theorem. The image of an open set under a non-constant analytic function is open.

Analytic functions also have the nice geometric property that they preserve angles. Let γ_1 and γ_2 be two differentiable curves intersecting at a point z in Ω. The *angle* from γ_1 to γ_2 is the signed angle of their tangent lines at z. A function $f(z)$ is said to be *conformal* at the point z if it preserves the angles of pairs of curves intersecting at z. An analytic function $f(z)$ preserves angles at points z where the derivative $f'(z) \neq 0$. If an analytic function $f(z)$ is conformal at every point in Ω (equivalently, if $f'(z) \neq 0$

for every point z in Ω), then it is said to be a *conformal mapping on* Ω.

Examples of conformal mappings (on suitable regions) are Möbius transformations. Let a, b, c, and d be complex numbers such that $ad - bc \neq 0$. Then the bilinear transformation

$$T(z) = \frac{az + b}{cz + d}$$

is a *Möbius transformation.* Any Möbius transformation can be decomposed into a product of four elementary conformal mappings: translations, rotations, homotheties or dilations, and inversions. In addition to preserving angles, Möbius transformations map circles into circles, provided that a straight line is viewed as a "circle" passing through the point ∞ at infinity.

D. Some Theorems about Analytic Functions

If $f(z)$ is a function on the disk with center 0 and radius ρ and $r < \rho$, let $M_f(r)$ be the maximum value of $|f(z)|$ on the circle $\{z : |z| = r\}$.

Cauchy's inequality. Let $f(z)$ be analytic in the disk with center 0 and radius ρ and let $f(z) = \sum_{m=0}^{\infty} a_n z^n$. If $r < \rho$, then

$$|a_n| r^n \leq M_f(r)$$

A function is *entire* if it is analytic on the entire complex plane. The following are two theorems about entire function. The first follows easily from the case $n = 1$ of Cauchy's inequality.

Liouville's theorem. If $f(z)$ is an entire function and $f(z)$ is bounded on $\mathbf{C}$, then f must be a constant function.

The second is much harder. It implies the fact that if two or more complex numbers are absent from the image of an entire function, then that entire function must be a constant.

Picard's little theorem. An entire function that is not a polynomial takes every value, with one possible exception, infinitely many times.

Applying Liouville's theorem to the reciprocal of a nonconstant polynomial $p(z)$ and using the fact that $p(z) \to \infty$ as $z \to \infty$, one obtains the following important theorem.

The fundamental theorem of algebra. Every polynomial with complex coefficients of degree at least one has a root in $\mathbf{C}$.

It follows that a polynomial of degree n must have n roots in $\mathbf{C}$, counting multiplicities.

The next theorem is a fundamental property of analytic functions.

Maximum principle. Let Ω be a bounded domain in $\mathbf{C}$. Suppose that $f(z)$ is analytic in Ω and continuous in the closure of Ω. Then, $|f(z)|$ attains its maximum value $|f(a)|$ at a boundary point a of Ω. If $f(z)$ is not constant, then

$$|f(z)| < |f(a)|$$

for every point z in the interior of Ω.

Using the maximum principle, one obtains Schwarz's lemma.

Schwarz's lemma. Let $f(z)$ be analytic on the disk $D = \{z : |z| < 1\}$ with center 0 and radius 1. Suppose that $f(0) = 0$ and $|f(z)| \leq 1$ for all points z in D. Then,

$$|f'(0)| \leq 1$$

and for all points z in D,

$$|f(z)| \leq |z|$$

If $|f'(0)| = 1$ or $|f(a)| = |a|$ for some nonzero point a, then $f(z) = \alpha z$ for some complex number α with $|\alpha| = 1$.

Another theorem, which has inspired many generalizations in the theory of partial differential equations, is the following.

Hadamard's three-circles theorem. Let $f(z)$ be an analytic function on the annulus $\{z : \rho_1 \leq |z| \leq \rho_3\}$ and let $\rho_1 < r_1 \leq r_2 \leq r_3 < \rho_3$. Then

$$\log M_f(r_2) \leq \frac{\log r_3 - \log r_2}{\log r_3 - \log r_1} \log M_f(r_1) + \frac{\log r_2 - \log r_1}{\log r_3 - \log r_1} \log M_f(r_3)$$

It follows from the three-circles theorem that $\log M_f(r)$ is a convex function of $\log r$.

Cauchy's integral theorem has the following converse.

Morera's theorem. Let $f(z)$ be a continuous function on a simply connected region Ω in $\mathbf{C}$. Suppose that the line integral

$$\int_{\partial \triangle} g(z)\, dz$$

over the boundary $\partial\triangle$ of every triangle in Ω is zero. Then $f(z)$ is analytic in Ω.

E. Analytic Continuation

An analytic function $f(z)$ is usually defined initially with a certain formula in some region D_1 of the complex plane. Sometimes, one can extend the function $f(z)$ to a function $\hat{f}(z)$ that is analytic on a bigger region D_2 containing D_1 such that $\hat{f}(z) = f(z)$ for all points z on D_1. Such an extension is called *analytic continuation.* Expanding the function as a Taylor (or Laurent series) is one possible

way to extend a function locally from a neighborhood of a point. Contour integration is another way.

Analytic continuation results in a unique extended function when it is possible. It also preserves identities between functions.

The uniqueness theorem for analytic continuation. Let $f(z)$ and $g(z)$ be analytic in a region Ω. If the set of points z in Ω where $f(z) = g(z)$ has a limit point in Ω, then $f(z) = g(z)$ for all z in Ω. In particular, if the set of zeros of $f(z)$ in Ω has a limit point in Ω, then $f(z)$ is identically zero in Ω.

Permanence of functional relationships. If a finite number of analytic functions in a region Ω satisfy a certain functional equation in a part of Ω that has a limit point, then that functional equation holds everywhere in Ω.

For example, the Pythagorean identity

$$\sin^2 x + \cos^2 x = 1$$

holds for all real numbers x. Thus, it holds for all complex numbers.

Deriving a functional equation is often the key step in analytic continuation. A famous example is the Riemann functional equation relating the gamma function and the Riemann zeta function.

A quick and direct way (when it works) is the following method.

Schwarz's reflection principle. Let Ω be a domain in the upper half-plane that contains a line segment L on the real axis. Let $f(z)$ be a function analytic on Ω and continuous on L. Then the function extended by defining $f(z) = f(\bar{z})$ for z in the "reflection" $\tilde{\Omega} = \{z | \bar{z} \in \Omega\}$ is an analytic continuation of $f(z)$ from Ω to to the bigger domain $\Omega \cup \tilde{\Omega}$.

We note that not all analytic functions have proper analytic continuations. For example, when $0 < |a| < 1$, the *Fredholm series*

$$\sum_{m=1}^{\infty} a^n z^{n^2}$$

converges absolutely on the closed unit disk $\{z : |z| \le 1\}$ and defines an analytic function $f(z)$ on the open disk $\{z : |z| < 1\}$. However, it can be shown that $f(z)$ has no analytic continuation outside the unit disk. Roughly speaking, the Fredholm functions are not extendable because of "gaps" in the powers of z. The sharpest "gap" theorem is the following.

The Fabry gap theorem. Let

$$f(z) = \sum_{m=0}^{\infty} a_m z^{b_m}$$

where b_m is a sequence of increasing nonnegative integers such that

$$\lim_{m \to \infty} b_m / m = \infty$$

Suppose that the power series has radius of convergence ρ. Then $f(z)$ has no analytic continuation outside the disk $\{z : |z| < \rho\}$.

IV. MEROMORPHIC FUNCTIONS

A. Poles and Meromorphic Functions

A point a is an *isolated singularity* of the analytic function $f(z)$ if $f(z)$ is analytic in a neighborhood of a, except possibly at the point itself. For example, the function $f(z) = 1/z$ is analytic on the entire complex plane, except at the isolated singularity $z = 0$. If the limit $\lim_{z \to a} f(z)$ is a finite complex number c, then we can simply define $f(a) = c$ and $f(z)$ will be analytic on the entire neighborhood. Such an isolated singularity is said to be *removable.*

If the limit $\lim_{z \to a} f(z)$ is infinite, but for some positive real number α, $\lim_{z \to a} |z - a|^{\alpha} |f(z)|$ is finite, then a is a *pole* of $f(z)$. It can be proved that if the last condition holds, then the smallest such real number α must be a positive integer k. In this case, the pole a is said to have order k. Equivalently, k is the smallest positive integer such that $(z - a)^k f(z)$ is analytic on an entire neighborhood of a (including a itself).

If an isolated singularity is neither removable nor a pole, it is said to be *essential.* Weierstrass and Casorati proved that an analytic function comes arbitrarily close to every complex number in every neighborhood of an isolated essential singularity. A refinement of this is a deep theorem of Picard.

Picard's great theorem. An analytic function takes on every complex value with one possible exception in every neighborhood of an essential singularity.

These results say that near an isolated singularity, an analytic function behaves very wildly. Thus, the study of isolated singularities has concentrated on analytic functions with poles. A complex function $f(z)$ is *meromorphic* in a region Ω if it is analytic except at a discrete set of points, where it may have poles.

The *residue* $\mathrm{Res}(f, a)$ of a meromorphic function $f(z)$ at the isolated singularity a is defined by

$$\mathrm{Res}(f, a) = \frac{1}{2\pi i} \int_{\gamma} f(\zeta)\, d\zeta$$

If one allows negative powers, then analytic functions can be expanded as power series at isolated singularities. The idea is to write a meromorphic function $f(z)$ in a neighborhood of a pole a as a sum of an analytic part and a singular part. Suppose the function $f(z)$ is analytic in a region containing the annulus $\{z : \rho_1 < |z - a| < \rho_2\}$.

Then we can define two functions $f_1(z)$ and $f_2(z)$ by:

$$f_1(z) = \frac{1}{2\pi i}\int_{\{\zeta:|\zeta-a|=r\}} \frac{f(\zeta)\,d\zeta}{\zeta - z}$$

where r satisfies $|z-a| < r < \rho_2$, and

$$f_2(z) = -\frac{1}{2\pi i}\int_{\{\zeta:|\zeta-a|=r\}} \frac{f(\zeta)\,d\zeta}{\zeta - z}$$

where r satisfies $\rho_1 < r < |z-a|$. The function $f_1(z)$ is analytic in the disk $\{z : |z-a| < \rho_2\}$ and the function $f_2(z)$ is analytic in the complement $\{z : |z-a| > \rho_1\}$. By the Cauchy integral formula, $f(z) = f_1(z) + f_2(z)$ and this representation is valid in the annulus $\{z : \rho_1 < |z-a| < \rho_2\}$.

The functions $f_1(z)$ and $f_2(z)$ can each be expanded as Taylor series. Using the transformation $z - a \mapsto 1/z$ and some simple calculation, we obtain the *Laurent series* expansion

$$f(z) = \sum_{m=-\infty}^{\infty} a_m (z-a)^m$$

where

$$a_m = \frac{1}{2\pi i}\int_{\{\zeta:|\zeta-a|=r\}} \frac{f(\zeta)\,d\zeta}{(\zeta - z)^{m+1}}$$

valid in the annulus $\{z : \rho_1 < |z-a| < \rho_2\}$. Note that

$$\mathrm{Res}(f, a) = a_{-1},$$

and the point a is a pole of order k if and only if $a_{-k} \neq 0$ and every coefficient a_{-m} with $m > k$ is zero. The polynomial

$$\sum_{m=1}^{k} \frac{a_m}{(z-a)^m}$$

in the variable $1/(z-a)$ is called the *singular* or *principal part* of $f(z)$ at the pole a.

B. Elliptic Functions

Important examples of meromorphic functions are elliptic functions. Elliptic functions arose from attempts to evaluate certain integrals. For example, to evaluate the integral

$$\int_0^1 \sqrt{1-x^2}\,dx$$

which gives the area $\pi/2$ of a semicircle with radius 1, one can use the substitution $x = \sin\theta$. However, to evaluate integrals of the form

$$\int_0^1 \sqrt{(1-x^2)(1-k^2x^2)}\,dx$$

we need elliptic functions. Elliptic functions are doubly periodic generalizations of trigonometric functions.

Let ω_1 and ω_2 be two complex numbers whose ratio ω_1/ω_2 is not real and and let L be the 'integer lattice' $\{c\omega_1 + d\omega_2\}$, where c and d range over all integers. An *elliptic function* is a meromorphic function on the complex plane with two (independent) periods, ω_1 and ω_2, that is, $f(z+\omega_1) = f(z)$, $f(z+\omega_2) = f(z)$, and every complex number ω such that $f(z+\omega) = f(z)$ for all points z in the complex plane is a number in the lattice L.

A specific example of an elliptic function is the *Weierstrass $\wp$-function* defined by the formula

$$\wp(z) = \frac{1}{z^2} + \sum_{\omega\in L\setminus\{0\}} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}\right)$$

This defines a meromorphic function on the complex plane that is doubly periodic with periods ω_1 and ω_2. The $\wp$-function has poles exactly at points in L. Weierstrass proved that every elliptic function with periods ω_1 and ω_2 can be written as a rational function of $\wp$ and its derivative $\wp'$.

C. The Cauchy Residue Theorem

A simple but useful generalization of Cauchy's integral formula is the Cauchy residue theorem.

The Cauchy residue theorem. Let Ω be a simply connected region in the complex plane, let $f(z)$ be a function analytic on Ω except at the isolated singularities a_m, and let let γ be a closed piecewise continuously differentiable curve in Ω that does not pass through any of the points a_m. Then

$$\frac{1}{2\pi i}\int_\gamma f(z)dz = \sum n(\gamma, a_m)\mathrm{Res}(f, a_m)$$

where the sum ranges over all the isolated singularities inside the curve γ.

Cauchy's residue theorem has the following useful corollary.

The argument principle. Let $f(z)$ be a meromorphic function in a simply connected region Ω, let $a_1, a_2, \ldots,$ be the zeros of $f(z)$ in Ω, and let $b_1, b_2, \ldots,$ be the poles of $f(z)$ in Ω. Suppose the zero a_m has multiplicity s_m and the pole b_n has order t_n. Let γ be a closed piecewise continuously differentiable curve in Ω that does not pass through any poles or zeros of $f(z)$. Then

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(\zeta)\,d\zeta}{f(\zeta)} = \sum s_m n(\gamma, a_m) - \sum t_j n(\gamma, b_j)$$

where the sum ranges over all the zeros and poles of $f(z)$ contained in the curve γ.

The name "argument principle" came from the following special case. When γ is a circle, the argument principle says that the change in the argument of $f(z)$ as z traces the circle in a counterclockwise direction, equals

$$Z(f) - P(f)$$

the difference between the number $Z(f)$ of zeros and the number $P(f)$ of poles of $f(z)$ inside γ, counting multiplicities and orders.

D. Evaluation of Real Integrals

Cauchy's residue theorem can be used to evaluate real definite integrals that are otherwise difficult to evaluate.

For example, to evaluate an integral of the form

$$\int_0^{2\pi} R(\cos\theta, \sin\theta)\, d\theta$$

where R is a rational function of $\cos\theta$ and $\sin\theta$, let $z = e^{i\theta}$. If we make the substitutions

$$\cos\theta = \frac{z + z^{-1}}{2}, \qquad \sin\theta = \frac{z - z^{-1}}{2i}$$

the integral becomes a line integral over the unit circle of the form

$$\int_{|z|=1} S(z)\, dz$$

where $S(z)$ is a rational function of z. By Cauchy's residue theorem, this integral equals $2\pi i$ times the sum of the residues of the poles of $S(z)$ inside the unit circle. Using this method, one can prove, for example, that if $a > b > 0$,

$$\int_0^{2\pi} \frac{d\theta}{(a + b\cos^2\theta)^2} = \frac{\pi(2a + b)}{a^{3/2}(a + b)^{3/2}}$$

One can also evaluate improper integrals, obtaining formulas such as the following formula due to Euler: For $-1 < p < 1$ and $-\pi < \alpha < \pi$,

$$\int_0^{\infty} \frac{x^{-p}\, dx}{1 + 2x\cos\alpha + x^2} = \frac{\pi \sin p\alpha}{\sin p\pi \sin\alpha}$$

E. Location of Zeros

It is often useful to locate zeros of polynomials in the complex plane. An elegant theorem, which can be proved by elementary arguments, is the following result.

Lucas' theorem. Let $p(z)$ be a polynomial of degree at least 1. All the zeros of the derivative $p'(z)$ lie in the convex closure of the set of zeros of $p(z)$.

Deeper results usually involve using some form of Rouché's theorem, which is proved using the argument principle.

Rouché's theorem. Let Ω be a region bounded by a simple closed piecewise continuously differentiable curve. Let $f(z)$ and $g(z)$ be two functions meromorphic in an open set containing the closure of Ω. If $f(z)$ and $g(z)$ satisfy

$$|f(z) - g(z)| < |f(z)| + |g(z)|$$

for every point z on the curve bounding Ω, then

$$Z(f) - P(f) = Z(g) - P(g)$$

Hurwitz's theorem. Let $(f_n(z) : n = 1, 2, \ldots)$ be a sequence of functions analytic in a region Ω bounded by a simple closed piecewise continuously differentiable curve such that $f_n(z)$ converges uniformly to a nonzero (analytic) function $f(z)$ on every closed subset of Ω. Let a be an interior point of Ω. If a is a limit point of the set of zeros of the functions $f_n(z)$, then a is a zero of $f(z)$. If a is a zero of $f(z)$ with multiplicity m, then every sufficiently small neighborhood K of a contains exactly m zeros of the functions $f_n(z)$, for all n greater than a number N depending on K.

F. Infinite Products, Partial Fractions, and Approximations

A natural way to write a meromorphic function is in terms of its zeros and poles. For example, because $\sin\pi z$ has zeros at the integers, we expect to be able to "factor" it into product. Indeed, Euler wrote down the following product expansion:

$$\sin\pi z = \pi z \prod_{j=1}^{\infty} \left(1 - \frac{z}{n}\right)\left(1 + \frac{z}{n}\right)$$

With complex analysis, one can justify such expansions rigorously.

The question of convergence of an infinite product is easily resolved. By taking logarithms, one can reduce it to a question of convergence of a sum. For example, the product

$$\prod_{m=1}^{\infty} (1 + a_m)$$

converges absolutely if and only if the sum $\sum_{m=1}^{\infty} |\log(1 + a_m)|$ converges absolutely. Since $|\log(1 + a_m)|$ is approximately $|a_m|$, the product converges absolutely if and only if the series $\sum_{m=1}^{\infty} |a_m|$ converges absolutely.

The following theorem allows us to construct an entire function with a prescribed set of zeros.

The Weierstrass product theorem. Let $(a_j : j = 1, 2, \ldots)$ be a sequence of nonzero complex numbers in which no complex number occurs infinitely many times. Suppose that the set $\{a_j\}$ has no (finite) limit point in the complex plane. Then there exists an entire function $f(z)$ with a zero of multiplicity m at 0, zeros in the set $\{a_j\}$ with the correct multiplicity, and no other zeros. This function can be written in the form

$$f(z) = z^m e^{g(z)} \prod_{j=1}^{\infty} \left(1 - \frac{z}{a_j}\right) \times e^{a_j/z + (1/2)(a_j/z)^2 + \cdots + (1/m_j)(a_j/z)^{m_j}}$$

where m_j are positive integers depending on the set $\{a_j\}$, and $g(z)$ is an entire function.

From this theorem, we can derive the following representation of a meromorphic function.

Theorem. A meromorphic function on the complex plane is the quotient of two entire functions. The two entire functions can be chosen so that they have no common zeros.

In particular, one can think of meromorphic functions as generalizations of rational functions.

The *gamma function* $\Gamma(z)$ is a useful function which can be defined by a product formula. Indeed,

$$\Gamma(z) = \frac{e^{-\gamma z}}{z} \prod_{m=1}^{\infty} \left(1 + \frac{z}{m}\right)^{-1} e^{z/m}$$

where γ is the *Euler–Mascheroni constant* defined by

$$\gamma = \lim_{n\to\infty} \left[\left(\sum_{m=1}^{n} \frac{1}{m} \right) - \log n \right]$$

It equals $0.57722\ldots$. The gamma function interpolates the integer factorials. Specifically, $\Gamma(n) = (n-1)!$ for a positive integer n and $\Gamma(z)$ satisfies the functional equation

$$\Gamma(z+1) = z\Gamma(z).$$

Another useful functional equation is the *Legendre formula*

$$\sqrt{\pi}\Gamma(2z) = 2^{2z-1}\Gamma(z)\Gamma(z+1/2)$$

This gives the following useful value: $\Gamma(1/2) = \sqrt{\pi}$.

Rational functions can be represented as partial fractions; so can meromorphic functions.

Mittag-Leffler's theorem. Let $\{b_j : j = 1, 2, \ldots\}$ be a set of complex numbers with no finite limit point in the complex plane, and let $p_j(z)$ be given polynomials with zero constant terms, one for each point b_j. Then there exist meromorphic functions in the complex plane with poles at b_m with singular parts $p_j(1/z - b_j)$. These functions have the form

$$g(z) + \sum_{j=1}^{\infty} \left(p_j\left(\frac{1}{z-b_j}\right) - q_j(z) \right)$$

where $q_j(z)$ are suitably chosen polynomials depending on $p_j(z)$, and $g(z)$ is an entire function.

Taking logarithmic derivatives and integrating, one can derive Weierstrass's product theorem from Mittag-Leffler's theorem.

Two examples of partial fraction exapnsions of meromorphic functions are

$$\frac{\pi^2}{\sin^2 \pi z} = \sum_{m=-\infty}^{\infty} \frac{1}{(z-n)^2}$$

and

$$\pi \cot \pi z = \frac{1}{z} + \sum_{m=1}^{\infty} \frac{2z}{z^2 - n^2}$$

Runge's approximation theorem says that a function analytic on a bounded region Ω with holes can be uniformly approximated by a rational function all of whose poles lie in the holes. Runge's theorem can be proved using a Cauchy's integral formula for compact sets.

Runge's approximation theorem. Let $f(z)$ be an analytic function on a region Ω in the complex plane; let K be a compact subset of Ω. Let $\epsilon > 0$ be a given (small) positive real number. Then there exists a rational function $r(z)$ with all its poles outside K such that

$$|f(z) - r(z)| < \epsilon$$

for all z in K.

V. SOME ADVANCED TOPICS

A. Riemann Mapping Theorem

On the complex plane, most simply connected regions can be mapped conformally in a one-to-one way onto the open unit disk. This is a useful result because it reduces many problems to problems about the unit disk.

Riemann mapping theorem. Let Ω be a simply connected region that is not the entire complex plane, and let a be a point in Ω. Then there exists a unique analytic function $f(z)$ on Ω such that $f(a) = 0$, $f'(a) > 0$, and $f(z)$ is a one-to-one mapping of Ω onto the open disk $\{z : |z| < 1\}$ with radius 1.

For example, the upper half-plane $\{\zeta : \mathrm{Im}(\zeta) > 0\}$ and the unit disk $\{z : |z| < 1\}$ are mapped conformally onto each other by the Möbius transformations

$$\zeta = \frac{i(1+z)}{1-z}, \qquad z = \frac{\zeta - i}{\zeta + i}$$

The Schwarz–Christoffel formula gives an explicit formula for one-to-one onto conformal maps from the open unit disk or the upper half-plane to polygonal domains, which are sets in the complex plane bounded by a a closed simple curve made up of a finite number of straight line segments. It is rather complicated to state, and we refer the reader to the books by Ahlfors and Nehari in the bibliography. As an example, one can map the upper-half complex plane into a triangle with interior angles $\alpha_1\pi$, $\alpha_2\pi$, and $\alpha_3\pi$ (where $\alpha_1 + \alpha_2 + \alpha_3 = 1$) by the inverse of the function

$$F(z) = \int_0^z t^{\alpha_1 - 1}(t-1)^{\alpha_2 - 1}\, dt$$

The function $F(z)$ is called the *triangle function of Schwarz.*

B. Univalent Functions and Bieberbach's Conjecture

Univalent or *Schlicht* functions are one-to-one analytic functions. They have been extensively studied in complex analysis. A famous result in this area was the Bieberbach conjecture, which was proved by de Branges in 1984.

Theorem. Let $f(z)$ be a univalent analytic function on the unit disk $\{z : |z| < 1\}$ with power series expansion

$$f(z) = z + a_2 z^2 + a_3 z^3 + \cdots$$

(that is, $f(0) = 0$ and $f'(0) = 1$). Then

$$|a_n| \leq n.$$

When equality holds, $f(z) = e^{-i\theta} K(e^{i\theta} z)$, where

$$K(z) = \frac{z}{(1-z)^2} = z + 2z^2 + 3z^3 + 4z^4 + \cdots$$

The function $K(z)$ is called *Koebe's function.*

Another famous result in this area is due to Koebe.

Koebe's 1/4-theorem. Let $f(z)$ be a univalent function. Then the image of the unit disk $\{z : |z| < 1\}$ contains the disk $\{z : |z| < 1/4\}$ with radius $1/4$.

The upper bound $1/4$ (the "Koebe constant") is the best possible.

C. Harmonic Functions

Harmonic functions, defined in Section II, are real functions $u(x, y)$ satisfying Laplace's equation. We shall use the notation: $u(x, y) = u(x + iy) = u(z)$. Harmonic functions have several important properties.

The mean-value property. Let $u(z)$ be a harmonic function on a region Ω. Then for any disk D with center a and radius r whose closure is contained in Ω,

$$u(a) = \frac{1}{2\pi} \int_0^{2\pi} u(a + re^{i\theta})\, d\theta$$

The maximum principle for harmonic functions. Let $u(z)$ be a harmonic function on the domain Ω. If there is a point a in Ω such that $u(a)$ equals the maximum $\max\{u(z) : z \in \Omega\}$, then $u(z)$ is a constant function.

An important problem in the theory of harmonic function is *Dirichlet's problem.* Given a simply connected domain Ω and a piecewise continuous function $g(z)$ on the boundary $\partial\Omega$, find a function $u(z)$ in the closure $\bar{\Omega}$ such that $u(z)$ is harmonic in Ω, and the restriction of $u(z)$ to the boundary $\bar{\Omega} \backslash \Omega$ equals $g(z)$. (The *boundary* $\partial\Omega$ is the set $\bar{\Omega} \backslash \Omega$.)

For general simply connected domains, Dirichlet's problem is difficult. It is equivalent to finding a Green's function. For a disk, the following formula is known.

The Poisson formula. Let $g(z) = g(e^{i\phi})$, $-\pi < \phi \leq \pi$, be a piecewise continuous function on the boundary $\{z : |z| = 1\}$ of the unit disk. Then the function

$$u(z) = u(re^{i\theta}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\frac{1 - r^2}{1 + r^2 - 2r\cos(\phi - \theta)} \right) g(e^{i\phi})\, d\phi$$

is a solution to Dirichlet's problem for the unit disk.

D. Riemann Surfaces

A *Riemann surface* $\mathcal{S}$ is a one-dimensional complex connected paracompact Hausdorff space equipped with a *conformal atlas*, that is, a set of maps or *charts* $\{h : D \to N\}$, where D is the open disk $\{z : |z| < 1\}$ and N is an open set of $\mathcal{S}$, such that

1. The union of all the open sets N is $\mathcal{S}$.
2. The chart $h : D \to N$ is a homeomorphism of the disk D to N.
3. Let N_1 and N_2 be neighborhoods with charts h_1 and h_2. If the intersection $N_1 \cap N_2$ is nonempty and connected, then the composite mapping $h_2^{-1} \circ h_1$, defined on the inverse image $h_1^{-1}(N_1 \cap N_2)$, is conformal.

Riemann surfaces originated in an attempt to make a "multivalued" analytic function single-valued by making its range a Riemann surface. Examples of multivalued functions are *algebraic functions.* These are functions $f(z)$ satisfying a polynomial equation $P(f(z)) = 0$. A specific example of this is the square-root function $f(z) = \sqrt{z}$, which takes on two values $\pm\sqrt{z}$ except when $z = 0$. This can be made into a single-valued function using the Riemann surface obtained by gluing together two sheets or copies of the complex plane cut from 0 to ∞ along the positive real axis. Another example is the logarithmic function $\log z$, which requires a Riemann surface made from countably infinitely many sheets.

Intuitively, the *genus* of a Riemann surface $\mathcal{S}$ is the number of "holes" it has. The genus can be defined as the maximum number of disjoint simple closed curves that do not disconnect $\mathcal{S}$. For example, the extended complex plane has genus 0 and an annulus has genus 1. There are many results about Riemann surfaces. The following are two results that can be simply stated.

Picard's theorem. Let $P(u, v)$ be an irreducible polynomial with complex coefficients in two variables u and v. If there exist nonconstant entire functions $f(z)$ and $g(z)$ satisfying $P(f(z), g(z)) = 0$ for all complex numbers z,

then the Riemann surface associated with the algebraic equation $P(u, v) = 0$ has genus 0.

Koebe's uniformization theorem. If $\mathcal{S}$ is a simply connected Riemann surface, then $\mathcal{S}$ is conformally equivalent to

1. (Elliptic type) The Riemann sphere. In this case, $\mathcal{S}$ is the sphere.
2. (Parabolic type) The complex plane $\mathbf{C}$. In this case, $\mathcal{S}$ is biholomorphic to $\mathbf{C}$, $\mathbf{C}\backslash\{0\}$, or a torus.
3. (Hyperbolic type) The unit disk $\{z : |z| < 1\}$.

Complex manifolds are higher-dimensional generalizations of Riemann surfaces. They have been extensively studied.

E. Other Topics

Complex analysis is a vast and ever-expanding area. "Nine lifetimes" do not suffice to cover every topic. Some interesting areas that we have not covered are complex differential equations, complex dynamics, Montel's theorem and normal families, value distribution theory, and the theory of complex functions in several variables. Several books on these topics are listed in the Bibliography.

SEE ALSO THE FOLLOWING ARTICLES

CALCULUS • DIFFERENTIAL EQUATIONS • NUMBER THEORY • RELATIVITY, GENERAL • SET THEORY • TOPOLOGY, GENERAL

BIBLIOGRAPHY

Ahlfors, L. V. (1979). "Complex Analysis," McGraw-Hill, New York.

Beardon, A. F. (1984). "A Primer on Riemann Surfaces," Cambridge University Press, Cambridge, U.K.

Blair, D. E. (2000). "Inversion Theory and Conformal Mappings," American Mathematical Society, Providence, RI.

Carleson, L., and Gamelin, T. W. (1993). "Complex Dynamics," Springer-Verlag, Berlin.

Cartan, H. (1960). "Elementary Theory of Analytic Functions of One or Several Complex Variables," Hermann and Addison-Wesley, Paris and Reading, MA.

Cherry, W., and Ye, Z. (2001). "Nevanlinna's Theory of Value Distribution," Springer-Verlag, Berlin.

Chuang, C.-T., and Yang, C.-C. (1990). "Fix-Points and Factorization of Meromorphic Functions," World Scientific, Singapore.

Duren, P. L. (1983). "Univalent Functions," Springer-Verlag, Berlin.

Farkas, H. M., and Kra, I. (1992). "Riemann Surfaces," Springer-Verlag, Berlin.

Gong, S. (1999). "The Bieberbach Conjecture," American Mathematical Society, Providence, RI.

Gunning, R. (1990). "Introduction to Holomorphic Functions of Several Variables," Vols. I, II, and III. Wadsworth and Brooks/Cole, Pacific Grove, CA.

Hayman, W. K. (1964). "Meromorphic Functions," Oxford University Press, Oxford.

Hille, E. (1962, 1966). "Analytic Function Theory," Vols. I and II. Ginn-Blaisdell, Boston.

Hille, E. (1969). "Lectures on Ordinary Differential Equations," Addison-Wesley, Reading, MA.

Hu, P.-C., and Yang, C.-C. (1999). "Differential and Complex Dynamics of One and Several Variables," Kluwer, Boston.

Hua, X.-H., and Yang, C.-C. (1998). "Dynamics of Transcendental Functions," Gordon and Breach, New York.

Kodiara, K. (1984). "Introduction to Complex Analysis," Cambridge University Press, Cambridge, U.K.

Krantz, S. G. (1999). "Handbook of Complex Variables," Birkhäuser, Boston.

Laine, I. (1992). "Nevanlinna Theory and Complex Differential Equations," De Gruyter, Berlin.

Lang, S. (1987). "Elliptic Functions," 2nd ed., Springer-Verlag, Berlin.

Lehto, O. (1987). "Univalent Functions and Teichmüller Spaces," Springer-Verlag, Berlin.

Marden, M. (1949). "Geometry of Polynomials," American Mathematical Society, Providence, RI.

McKean, H., and Moll, V. (1997). "Elliptic Curves, Function Theory, Geometry, Arithmetic," Cambridge University Press, Cambridge, U.K.

Morrow, J., and Kodiara, K. (1971). "Complex Manifolds," Holt, Rinehart and Winston, New York.

Nehari, Z. (1952). "Conformal Mapping," McGraw-Hill, New York; reprinted, Dover, New York.

Needham, T. (1997). "Visual Complex Analysis," Oxford University Press, Oxford.

Palka, B. P. (1991). "An Introduction to Complex Function Theory," Springer-Verlag, Berlin.

Pólya, G., and Szegö, G. (1976). "Problems and Theorems in Analysis," Vol. II. Springer-Verlag, Berlin.

Protter, M. H., and Weinberger, H. F. (1984). "Maximum Principles in Differential Equations," Springer-Verlag, Berlin.

Remmert, R. (1993). "Classical Topics in Complex Function Theory," Springer-Verlag, Berlin.

Rudin, W. (1980). "Function Theory in the Unit Ball of $\mathbf{C}^n$," Springer-Verlag, Berlin.

Schiff, J. L. (1993). "Normal Families," Springer-Verlag, Berlin.

Schwerdtfeger, H. (1962). "Geometry of Complex Numbers," University of Toronto Press, Toronto. Reprinted, Dover, New York.

Siegel, C. L. (1969, 1971, 1973). "Topics in Complex Function Theory," Vols. I, II, and III, Wiley, New York.

Smithies, F. (1997). "Cauchy and the Creation of Complex Function Theory," Cambridge University Press, Cambridge, U.K.

Steinmetz, N. (1993). "Rational Iteration, Complex Analytic Dynamical Systems," De Gruyter, Berlin.

Titchmarsh, E. C. (1939). "The Theory of Functions," 2nd ed., Oxford University Press, Oxford.

Vitushkin, A. G. (ed.). (1990). "Several Complex Variables I," Springer-Verlag, Berlin.

Weyl, H. (1955). "The Concept of a Riemann Surface," 3rd ed., Addison-Wesley, Reading, MA.

Whitney, H. (1972). "Complex Analytic Varieties," Addison-Wesley, Reading, MA.

Whittaker, E. T., and Watson, G. N. (1969). "A Course of Modern Analysis," Cambridge University Press, Cambridge, U.K.

Yang, L. (1993). "Value Distribution Theory," Springer-Verlag, Berlin.



Computer-Based Proofs of Mathematical Theorems

C. W. H. Lam
Concordia University

I. Mathematical Theories
II. Computer Programming
III. Computer As an Aid to Mathematical Research
IV. Examples of Computer-Based Proofs
V. Proof by Exhaustive Computer Enumeration
VI. Recent Development: RSA Factoring Challenge
VII. Future Directions

GLOSSARY

Axiom A statement accepted as true without proof.
Backtrack search A method of organizing a search for solutions by a systematic extension of partial solutions.
Computer-based proof A proof with a heavy computer component; one which is impossible to do by hand.
Computer programming A process of creating a sequence of instructions to be used by a computer.
Enumerative proof A proof method by exhibiting and analyzing all possible cases.
Monte Carlo method A method of estimation by performing random choices.
Optimizing a computer program A process of fine-tuning a computer program so that it runs faster.
Predicate A statement whose truth value depends of the values of its arguments.
Proof A demonstration of the truth of a statement.
Proposition A statement which is either true or false, but not both.
Search tree A pictorial representation of the partial solutions encountered in a backtrack search.

EVER SINCE the arrival of computers, mathematicians have used them as a computational aid. Initially, they were used to perform tedious and repetitious calculations. The tremendous speed and accuracy of computers enable mathematicians to perform lengthy calculations without fear of making careless mistakes. Their main application, however, has been to obtain insight into various mathematical problems, which then led to conventional proofs independent of the computer. Recently, there was a departure from this traditional approach. By exploiting the speed of a computer, several famous and long-standing problems were settled by lengthy enumerative proofs, a technique

only suitable for computers. Two notable examples are the four-color theorem and the nonexistence of a finite projective plane of order 10. Both proofs required thousands of hours of computing and gave birth to the term "a computer-based proof." The organization of such a proof requires careful estimation of the necessary computing time, meticulous optimization of the computer program, and prudent control of all possible computer errors. In spite of the difficulty in checking these proofs, mathematicians are starting to accept their validity. As computers are getting faster, many other famous open problems will be solved by this new approach.

I. MATHEMATICAL THEORIES

What is mathematics? It is not possible to answer this question precisely in this short article, but a generally acceptable definition is that *mathematics is a study of quantities and relations using symbols and numbers*. The starting point is often a few undefined objects, such as a *set* and its *elements*. A mathematical theory is then built by assuming some axioms which are statements accepted as true. From these basic components, further properties can then be derived.

For example, the study of geometry can start from the undefined objects called *points*. A *line* is then defined as a set of points. An axiom may state: "Two distinct lines contain at most one common point." This is one of the axioms in Euclidean geometry, where it is possible to have parallel lines. The complete set of axioms of Euclidean geometry was classified by the great German mathematician David Hilbert in 1902. Using these axioms, further results can be derived. Here is an example:

Two triangles are congruent if the three sides of one are equal to the three sides of the other.

To show that these derived results follow logically from the axioms, a system of well-defined principles of mathematical reasoning is used.

A. Mathematical Statements

The ability to demonstrate the truth of a statement is central to any mathematical theory. This technique of reasoning is formalized in the study of *propositional calculus*.

A *proposition* is a statement which is either true or false, but not both. For example, the following are propositions:

- $1 + 1 = 2$.
- 4 is a prime number.

The first proposition is true, and the second one is false.

The statement "$x = 3$" is not a proposition because its truth value depends on the value of x. It is a *predicate*, and its truth value depends of the value of the *variable* or *argument* x. The expression $P(x)$ is often used to denote a predicate P with an argument x. A predicate may have more than one argument, for example, $x + y = 1$ has two arguments. To convert a predicate to a proposition, values have to be assigned to all the arguments. The process of associating a value to a variable is called a *binding*.

Another method of converting a predicate to a proposition is by the quantification of the variables. There are two common quantifiers: *universal* and *existential*. For example, the statement

$$\text{For all } x,\ x < x + 1$$

uses the universal quantifier "for all" to provide bindings for the variable x in the predicate $x < x + 1$. If the predicate is true for every possible value of x, then the proposition is true. The symbol $\forall$ is used to denote the phrase "for all." Thus, the above proposition can also be written as

$$\forall x,\ x < x + 1.$$

The set of possible values for x has to come from a certain *universal* set. The universal set in the above example may be the set of all integers, or it may be the set of all reals. Sometimes, the actual universal set can be deduced from context and, consequently, not stated explicitly in the proposition. A careful mathematician may include all the details, such as

$$\forall \text{ real } x > 1,\ x > \sqrt{x}.$$

The choice is often not a matter of sloppiness, but a conscious decision depending on whether all the exact details will obscure the main thrust of the result.

An existential quantifier asserts that a predicate $P(x)$ is true for one or more x in the universal set. It is written as

$$\text{For some } x,\ P(x),$$

or, using the symbol $\exists$, as

$$\exists x,\ P(x).$$

This assertion is false only if for all x, $P(x)$ is false. In other words,

$$\neg[\exists x,\ P(x)] \equiv [\forall x,\ \neg P(x)].$$

B. Enumerative Proof

A proof of a mathematical result is a demonstration of its truth. To qualify as a proof, the demonstration must be absolutely correct and there can be no uncertainty nor ambiguity in the proof.

Many interesting mathematical results involve quantifiers such as $\exists$ or $\forall$. There are many techniques to prove quantified statements, one of which is the enumerative proof. In this method, the validity of $\forall x, P(x)$ is established by investigating $P(x)$ for every value of x one after another. The proposition $\forall x, P(x)$ is only true if $P(x)$ has been verified to be true for all x. However, if $P(x)$ is false for one of the values, then it is not necessary to consider the remaining values of x, because $\forall x, P(x)$ is already false. Similarly, the validity of $\exists x, P(x)$ can be found by evaluating $P(x)$ for every value of x. The process of evaluation can stop once a value of x is found for which $P(x)$ is true, because $\exists x, P(x)$ is true irrespective of the remaining values of x. However, to establish that $\exists x, P(x)$ is false, it has to be shown that $P(x)$ is false for all values of x.

To ensure a proof that is finite in length, enumerative proofs are used where only a finite number of values of x have to be considered. It may be a case where x can take on only a finite number of values, or a situation where an infinite number of values can be handled by another proof technique, leaving only a finite number of values for which $P(x)$ are unknown.

An enumerative proof is a seldom-used method because it tends to be too long and tedious for a human being. A proof involving a hundred different cases is probably the limit of the capacity of the human mind. Yet, it is precisely in this area that a computer can help, where it can evaluate millions or even billions of cases with ease. Its use can open up new frontiers in mathematics.

C. Kinds of Mathematical Results

Mathematicians love to call their results *theorems*. Along the way, they also prove lemmas, deduce corollaries, and propose conjectures. What are these different classifications? The following paragraphs will give a brief answer to this question.

Mathematical results are classified as lemmas, theorems, and corollaries, dependent on their importance. The most important results are called theorems. A *lemma* is an auxiliary result which is useful in proving a theorem. A *corollary* is a subsidiary result that can be derived from a theorem. These classifications are quite loose, and the actual choice of terminology is based on the subjective evaluation of the discoverer of the results. There are cases where a lemma or a corollary have, over a period of time, attained an importance surpassing the main theorem.

A *conjecture*, on the other hand, is merely an educated guess. It is a statement which may or may not be true. Usually, the proposer of the conjecture suspects that it is highly probable to be true but cannot prove it. Many famous mathematical problems are stated as conjectures. If a proof is later found, then it becomes a theorem.

D. What Makes an Interesting Theorem

Of course, a theorem must be interesting to be worth proving. However, whether something is interesting is a subjective judgement. The following list contains some of the properties of an interesting result:

- Useful
- Important
- Elegant
- Provides insight

While the usefulness of some theorems is immediately obvious, the appreciation of others may come only years later. The importance of a theorem is often measured by the number of mathematicians who know about it or who have tried proving it. An interesting theorem should also give insights about a problem and point to new research directions. A theorem is often appreciated as if it is a piece of art. Descriptions such as "It is a beautiful theorem" or "It is an elegant proof" are often used to characterize an interesting mathematical result.

II. COMPUTER PROGRAMMING

Figure 1 shows a highly simplified view of a computer. There are two major components: the central processing unit (CPU) and memory. Data are stored in the memory. The CPU can perform operations which changes the data. A program is a sequence of instructions which tell the CPU what operations to perform and it is also stored in the computer memory.

Computer programming is the process of creating the sequence of instructions to be used by the computer. Most programming today is done in high-level languages such as Fortran, Pascal, or C. Such languages are designed for the ease of creating and understanding a complex program by human beings. A program written in one of these languages has to be translated by a compiler before it can be used by the computer.

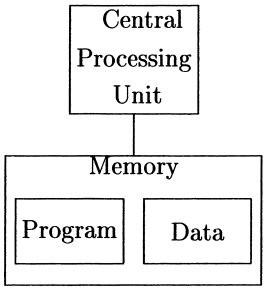


FIGURE 1 A simplified view of a computer.

Due to the speed with which a computer can perform operations it is programmed to do and its capacity in dealing with huge amounts of information, it is well suited to do lengthy or repetitive tasks that human beings are not good at doing.

III. COMPUTER AS AN AID TO MATHEMATICAL RESEARCH

Ever since the arrival of computers, mathematicians have used them as a computational aid. Initially, computers were used to perform tedious and repetitious calculations. The tremendous speed and accuracy of computers enable mathematicians to perform lengthy calculations without fear of making careless mistakes. For example, π can be computed to several billion digits of accuracy, and prime numbers of ever-increasing size are being found continually by computers.

Computer can also help mathematicians manipulate complicated formulae and equations. MACSYMA is the first such symbol manipulation program developed for mathematicians. A typical symbol manipulation program can perform integration and differentiation, manipulate matrices, factorize polynomials, and solve systems of algebraic equations. Many programs have additional features such as two- and three-dimensional plotting or even the generation of high-level language output.

Even though these programs have been used to prove a number of results directly, their main application has been to obtain insight into the behavior of various mathematical objects, which then has led to conventional proofs. Mathematicians still prefer computer-free proofs, if possible. A proof that involves using the computer is difficult to check and so its correctness cannot be determined absolutely. The term "a computer-based proof" is used to refer to a proof where part of it involves extensive computation, for which an equivalent human argument may take millions of years to make.

IV. EXAMPLES OF COMPUTER-BASED PROOFS

In the last 20 years, two of the best known mathematical problems were solved by lengthy calculations: the four-color conjecture and the existence question of a finite projective plane of order 10.

A. Four-Color Conjecture

The *four-color conjecture* says that four colors are sufficient to color any map drawn in the plane or on a sphere so that no two regions with a common boundary line are colored with the same color.

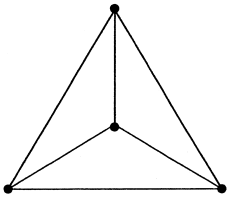


FIGURE 2 The complete graph K_4.

Francis Guthrie was given credit as the originator of this problem while coloring a map of England. His brother communicated the conjecture to DeMorgan in October 1852. Attempts to prove the four-color conjecture led to the development of a major branch of mathematics: graph theory.

A graph $G(V, E)$ is a structure which consists of a set of vertices $V = \{v_1, v_2, \ldots\}$ and a set of edges $E = \{e_1, e_2, \ldots\}$; each edge e is incident on two distinct vertices u and v. For example, Fig. 2 is a graph with four vertices and six edges. The technical name for this graph is the *complete graph* K_4. It is *complete* because there is an edge incident on every possible pair of vertices. Figure 3 is another graph with six vertices and nine edges. It is the *complete bipartite* graph $K_{3,3}$. In a *bipartite* graph, the set of vertices is divided into two classes, and the only edges are those that connect a vertex from one class to one of the other class. The graph $K_{3,3}$ is complete because it contains all the possible nine edges of the bipartite graph. A graph is said to be *planar* if it can be drawn on a plane in such a way that no edges cross one another, except, of course, at common vertices. The graph K_4 in Fig. 2 is drawn with no crossing edges and it is obviously planar. The graph $K_{3,3}$ can be shown to be not planar, no matter how one tries to draw it.

One can imagine that a planar graph is an abstract representation of a map. The vertices represent the capitals of the countries in the map. Two vertices are joined by an edge if the two countries have a common boundary. Instead of coloring regions of a map, colors can be assigned to the vertices of a graph representing the map. A graph is *k-colorable* if there is a way to assign colors to the graph such that no edge is incident on two vertices with the same

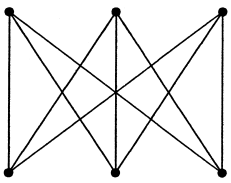


FIGURE 3 The complete bipartite graph $K_{3,3}$.

color. So, the four-color conjecture can also be stated as every planar graph is 4-colorable.

The graph K_4 is not 3-colorable. This can be proved by contradiction. Suppose it is 3-colorable. Since there are four vertices, two of the vertices must have the same color. Since the graph is complete, there is an edge connecting these two vertices of the same color, which is a contradiction.

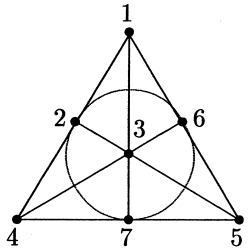
Kempe in 1879 published an incorrect proof of the four-color conjecture. Heawood in 1890 pointed out Kempe's error, but demonstrated that Kempe's method did prove that every planar graph is 5-colorable. Since then, many famous mathematicians have worked on this problem, leading to many significant theoretical advances. In particular, it was shown that the validity of the four-color conjecture depended only on a finite number of graphs. These results laid the foundation for a computer-based proof by Appel and Haken in 1976. Their proof depended on a computer analysis of 1936 graphs which took 1200 h of computer time and involved about 10^{10} separate operations. Finally, the four-color conjecture became the four-color theorem.

B. Projective Plane of Order 10

The question of the possible existence of a projective plane of order 10 was also settled recently by using a computer. A *finite projective plane of order n*, with $n > 0$, is a collection of $n^2 + n + 1$ lines and $n^2 + n + 1$ points such that

1. Every line contains $n + 1$ points.
2. Every point is on $n + 1$ lines.
3. Any two distinct lines intersect at exactly one point.
4. Any two distinct points lie on exactly one line.

The smallest example of a finite projective plane is one of order 1, which is a triangle. The smallest nontrivial example is one of order 2, as shown in Fig. 4. There are seven points labeled from 1 to 7. There are also seven lines labeled $L1$ to $L7$. Six of them are straight lines, but $L6$ is represented by the circle through points 2, 6, and 7.



$$L1 = \{1, 2, 4\}$$
$$L2 = \{2, 3, 5\}$$
$$L3 = \{3, 4, 6\}$$
$$L4 = \{4, 5, 7\}$$
$$L5 = \{5, 6, 1\}$$
$$L6 = \{6, 7, 2\}$$
$$L7 = \{7, 1, 3\}$$

FIGURE 4 The finite projective plane of order 2.

	1	2	3	4	5	6	7
L1	1	1	0	1	0	0	0
L2	0	1	1	0	1	0	0
L3	0	0	1	1	0	1	0
L4	0	0	0	1	1	0	1
L5	1	0	0	0	1	1	0
L6	0	1	0	0	0	1	1
L7	1	0	1	0	0	0	1

FIGURE 5 An incidence matrix for the plane of order 2.

The earliest reference to a finite projective plane was in an 1856 book by von Staudt. In 1904, Veblen used the projective plane of order 2 as an exotic example of a finite object satisfying all the Hilbert's axioms for geometry. He also proved that this plane of order 2 cannot be drawn using only straight lines. In a series of papers from 1904 to 1907, Veblen, Bussey, and Wedderburn established the existence of most of the planes of small orders. Two of the smallest orders missing are 6 and 10. In 1949, the celebrated Bruck-Ryser theorem gave an ingenious theoretical proof of the nonexistence of the plane of order 6. The nonexistence of the plane of order 10 was established in 1988 by a computer-based proof.

In the computer, lines and points are represented by their incidence relationship. The *incidence matrix* $A = [a_{ij}]$ of a projective plane of order n is an $n^2 + n + 1$ by $n^2 + n + 1$ matrix where the columns represent the points and the rows represent the lines. The entry a_{ij} is 1 if point j is on line i; otherwise, it is 0. For example, Fig. 5 gives the incidence matrix for the projective plane of order 2. In terms of an incidence matrix, the properties of being a projective plane are translated into

1. A has constant row sum $n + 1$.
2. A has constant column sum $n + 1$.
3. The inner product of any two distinct rows of A is 1.
4. The inner product of any two distinct columns of A is 1.

These conditions can be encapsuled in the following matrix equation:

$$AA^T = nI + J,$$

where A^T denotes the transpose of the matrix A, I denotes the identity matrix, and J denotes the matrix of all 1's.

As a result of intensive investigation by a number of mathematicians, it was shown that the existence question of the projective plane of order 10 can be broken into four starting cases. Each case gives rise to a number of geometric configurations, each corresponding to a partially completed incidence matrix. Starting from these partial matrices, a computer program tried to complete them to a full

plane. After about 2000 computing hours on a CRAY-1A supercomputer in addition to several years of computing on a number of VAX-11 and micro-VAX computers, it was shown that none of the matrices could be completed, which implied the nonexistence of the projective plane of order 10. About 1012 different subcases were investigated.

V. PROOF BY EXHAUSTIVE COMPUTER ENUMERATION

The proofs of both the four-color theorem and the nonexistence of a projective plane order 10 share one common feature: they are both enumerative proofs. This approach to a proof is often avoided by humans, because it is tedious and error prone. Yet, it is tailor-made for a computer.

A. Methodology

Exhaustive computer enumeration is often done by a programming technique called *backtrack search*. A version of the search problem can be defined as follows:

Search problem:

Given a collection of sets of candidates $C_1, C_2, C_3, \ldots, C_m$ and a boolean *compatibility* predicate $P(x, y)$ defined for all $x \in C_i$ and $y \in C_j$, find an *m-tuple* $(x_1, \ldots, x_m)$ with $x_i \in C_i$ such that $P(x_i, x_j)$ is true for all $i \neq j$.

A m-tuple satisfying the above condition is called a *solution*.

For example, if we take $m = n^2 + n + 1$ and let C_i be the set of all candidates for row i of the incidence matrix of a projective plane, then $P(x, y)$ can be defined as

$$P(x, y) = \begin{cases} \text{true} & \text{if } \langle x, y \rangle = 1 \\ \text{false} & \text{otherwise,} \end{cases}$$

where $\langle x, y \rangle$ denotes the inner product of rows x and y. A solution is then a complete incidence matrix.

In a backtrack approach, we generate *k-tuples* with $k \leq m$. A k-tuples $(x_1, \ldots, x_k)$ is a *partial solution* at level k if $P(x_i, x_j)$ is true for all $i \neq j \leq k$. The basic idea of the backtrack approach is to *extend* a partial solution at level k to one at level $k + 1$, and if this extension is impossible, then to go back to the partial solution at level $k - 1$ and attempt to generate a different partial solution at level k.

For example, consider the search for a projective plane of order 1. In terms of the incidence matrix, the problem is to find a 3×3 (0,1)-matrix A satisfying the matrix equation

$$AA^T = I + J.$$

Suppose the matrix A is generated row by row. Since each row of A must have two 1's, the candidate sets C_i are all equal to {[110], [101], [011]}. A partial solution at level 1 can be formed by choosing any of these three rows as x_1. If x_1 is chosen to be [110], then there are only two choices for x_2 which would satisfy the predicate $P(x_1, x_2)$, namely, [101] and [011]. Each of these choices for x_2 has a unique extension to a solution.

A nice way to organize the information inherent in the partial solutions is the backtrack *search tree*. Here, the empty partial solution ( ) is taken as the root, and the partial solution $(x_1, \ldots, x_k)$ is represented as the child of the partial solution $(x_1, \ldots, x_{k-1})$. Following the computer science terminology, we call a partial solution a *node*. It is also customary to label a node $(x_1, \ldots, x_k)$ by only x_k because this is the choice made at this level. The full partial solution can be read off the tree by following the branch from the root to the node in question. Figure 6 shows the search tree for the projective plane of order 1. The possible candidates are labeled as r_1, r_2, and r_3. The right-most branch, for example, represents choosing r_3 or [011] as the first row, r_2 as the second row, and r_1 as the third row.

It is often true that the computing cost of processing a node is independent of its level k. Under this assumption, the total computing cost of a search is equal to the number of nodes in the search tree times the cost of processing a node. Hence, the number of nodes in the search tree is an important parameter in a search. This number can be obtained by counting the nodes in every level, with α_i defined as the node count at level i of the tree. For example, in the search tree for the projective plane of order 1 shown in Fig. 6, $\alpha_1 = 3$, $\alpha_2 = 6$, and $\alpha_3 = 6$.

B. Estimation

It is very difficult to predict *a priori* the running time of a backtrack program. Sometimes, one program runs to completion in less than 1 sec, while other programs seem to take forever. A minor change in the strategy used in the

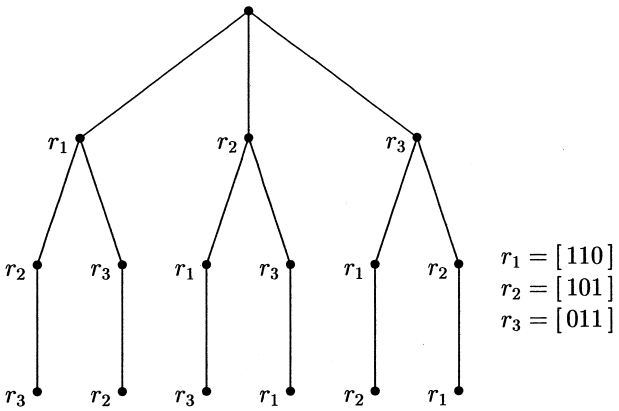


FIGURE 6 Search tree for the projective plane of order 1.

backtrack routine may change the total running time by several orders of magnitude. Some "minor improvements" may speed up the program by a factor of 100, while some other "major improvements" actually slow down the program. It is useful to have a simple and reliable method to estimate the running time of such a program. It can be used to

1. Decide whether the problem can be solved with the available resources.
2. Compare the various methods of "improvement."
3. Design an optimized program to reduce the running time.

A good approximation to the running time of a backtrack program is to multiply the number of nodes in the search tree by a constant representing the time required to process each node. Hence, a good estimate of the running time can be obtained by finding a good estimate to the number of nodes in the search tree. The size of the tree can be approximated by summing up the estimated α_i's, the node counts at each level.

An interesting solution to the estimation problem is a *Monte Carlo* approach developed by Knuth. The idea is to run a number of experiments, with each experiment consisting of performing the backtrack search with a randomly chosen candidate at each level of the search. Suppose we have a partial solution $(x_1, \ldots, x_k)$ for $0 \leq k < n$, where n is the depth of the search tree. We let

$$C'_{k+1} = \{x_{k+1} \in C_{k+1} | P(x_1, \ldots, x_{k+1}) \text{ is true}\}$$

be the set of acceptable candidates which extend $(x_1, \ldots, x_k)$. We choose x_{k+1} at random from C'_{k+1} such that each of the $|C'_{k+1}|$ possibilities are equally likely to be chosen. We let $d_k = |C'_k|$ be the number of elements in C'_k for $k = 1, \ldots, n$. Then the node count at level i, α_i, can be estimated by

$$\alpha_i \approx d_1 \ldots d_i.$$

Now, the total estimated size of the search tree is

$$\sum_{k=1}^{n} d_1 \ldots d_k.$$

The cost of processing a node can be estimated by running a few test cases, counting the nodes, and dividing the running time by the number of nodes.

These estimated values of the node counts can best be presented by plotting the logarithm of α_i, or equivalently the number of digits in α_i, as a function of i. A typical profile is shown in Fig. 7. The value of i for which $\log \alpha_i$ is maximum is called the *bulge* of the search tree. A backtrack search spends most of its time processing nodes near the bulge.

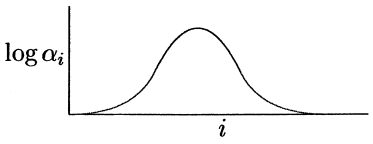


FIGURE 7 Typical shape of a search tree.

C. Optimization

A computer-based proof using backtracking may take a large amount of computing time. Optimization is a process of fine-tuning the computer program so that it runs faster. The optimization methods are divided into two broad classes:

1. Those whose aim is to reduce the size of the search tree
2. Those whose aim is to reduce the cost of the search tree by processing its nodes more efficiently

As a general rule, methods that reduce the size of the search tree, a process also called *pruning* the search tree, can potentially reduce the search by many orders of magnitude, whereas improvements obtained by trying to process nodes more efficiently are often limited to 1 or 2 orders of magnitude. Thus, given a choice, one should first try to reduce the size of the search tree.

There are many methods to prune the search tree. One possibility is to use a more effective compatibility predicate, while preserving the set of solutions. In a search tree, there are many branches which do not contain any solution. If these branches can be identified early, then they can be eliminated, hence reducing the size of the tree.

Another method to prune the search tree is by *symmetry pruning*. Technically, a symmetry is a property preserving operation. For example, two columns of a matrix A can be interchanged without affecting the product AA^T. Consider again the search tree for a projective plane of order 1. The interchange of columns 1 and 2 will induce a relabeling of r_2 as r_3 and vice versa. Thus, after considering the partial solution $x_1 = [101]$, there is no need to consider the remaining partial solution $x_1 = [011]$, because its behavior will be a duplicate of the earlier case. In combinatorial problems, the size of symmetries tends to be large and symmetry pruning can be very effective.

Methods to reduce the cost of processing a node of the search tree are often adaptions of well-known methods in computer programming. For example, one can use a better algorithm such as replacing a linear search by a binary search. We can also replace the innermost loop by an assembly language subroutine, or a faster computer can be used.

One common optimization technique is to move invariant operations from the inside of a loop to the outside. This idea can be applied to a backtrack search in the following manner. We try to do as little as possible for nodes near the bulge, at the expense of more processing away from the bulge. For example, suppose we have a tree of depth 3 and that $\alpha_1 = 1$, $\alpha_2 = 1000$, and $\alpha_3 = 1$. If the time required to process each node is 1 sec, then the processing time for the search tree is 1002 sec. Suppose we can reduce the cost of processing the nodes at level 2 by a factor of 10 at the expense of increasing the processing cost of nodes at levels 1 an 3 by a factor of 100. Then, the total time is reduced to 300 sec.

D. Practical Aspects

There are many considerations that go into developing a computer program which runs for months and years. Interruptions, ranging from power failures to hardware maintenance, are to be expected. A program should not have to restart from the very beginning for every interruption; otherwise, it may never finish. Fortunately, an enumerative computer proof can easily be divided into independent runs. If there is an interruption, just look up the last completed run and restart from that point. Thus, the disruptive effect of an untimely interrupt is now limited to the time wasted in the incomplete run. Typically, the problem is divided into hundreds or even millions of independent runs in order to minimize the time wasted by interruptions.

Another advantage of dividing a problem into many independent runs is that several computers can be used to run the program simultaneously. If a problem takes 100 years to run on one computer, then by running on 100 computers simultaneously the problem can be finished in 1 year.

E. Correctness Considerations

An often-asked question is, "How can one check a computer-based proof?" After all, a proof has to be absolutely correct. The computer program itself is part of the proof, and checking a computer program is no different from checking a traditional mathematical proof. Computer programs tend to be complicated, especially the ones that are highly optimized, but their complexity is comparable to some of the more difficult traditional proofs.

The actual execution of the program is also part of the proof, and the checking of this part is difficult or impossible. Even if the computer program is correct, there is still a very small chance that a computer makes an error in executing the program. In the search for a projective plane of order 10, this error probability is estimated to be 1 in 100,000.

So, it is impossible to have an absolute error-free, computer-based proof! In this sense, a computer-based proof is an experimental result. As scientists in other disciplines have long discovered, the remedy is an independent verification. In a sense, the verification completes the proof.

VI. RECENT DEVELOPMENT: RSA FACTORING CHALLENGE

Recently, there has been a lot of interest in factorizing big numbers. It all started in 1977 when Rivest, Shamir, and Adleman proposed a public-key cryptosystem based on the difficulty of factorizing large numbers. Their method is now known as the RSA scheme. In order to encourage research and to gauge the strength of the RSA scheme, RSA Data Security, Inc. in 1991 started the RSA Factoring Challenge. It consists of a list of large composite numbers. A cash prize is given to the first person to factorize a number in the list. These challenge numbers are identified by the number of decimal digits contained in the numbers. In February 1999, the 140-digit RSA-140 was factorized. In August 1999, RSA-155 was factorized. The best known factorization method divides the task into two parts: a sieving part to discover relations and a matrix reduction part to discover dependencies. The sieving part has many similarities with an enumerative proof. One has to try many possibilities, and the trials can be divided into many independent runs. In fact, for the factorization of RSA-155, the sieving part took about 8000 MIPS years and was accomplished by using 292 individual computers located at 11 different sites in 3 continents. The resulting matrix had 6,699,191 rows and 6,711,336 columns. It took 224 CPU hours and 3.2 Gbytes of central memory on a Cray C916 to solve. Fortunately, there is never any question about the correctness of the final answer, because one can easily verify the result by multiplying the factors together.

VII. FUTURE DIRECTIONS

When the four-color conjecture was first settled by a computer, there was some hesitation in mathematics circles to accept it as a proof. There is first the question of how to check the result. There is also the aesthetic aspect: "Is a computer-based proof elegant?" The result itself is definitely interesting. The computer-based proof of the nonexistence of a projective plane of order 10 again demonstrated the importance of this approach. A lengthy enumerative proof is a remarkable departure

from traditional thinking in terms of simple and elegant proofs. As computers are getting faster, many other famous open problems may be solved by this approach. Mathematicians are slowly coming to accept a computer-based proof as another proof technique. Before long, it will be treated as just another tool in a mathematician's tool box.

SEE ALSO THE FOLLOWING ARTICLES

COMPUTER ARCHITECTURE • DISCRETE MATHEMATICS AND COMBINATORICS • GRAPH THEORY • MATHEMATICAL LOGIC • PROBABILITY

BIBLIOGRAPHY

Cipra, B. A. (1989). "Do mathematicians still do math?" Res. News, Sci. **244,** 769–770.

Kreher, D. L., and Stinson, D. R. (1999). "Combinatorial Algorithms, Generation, Enumeration, and Search," CRC Press, Boca Raton, FL.

Lam, C. W. H. (1991). "The search for a finite projective plane of order 10," *Am. Math. Mon.* **98,** 305–318.

Lam, C. W. H. (1989). "How reliable is a computer-based proof?" *Math. Intelligencer* **12,** 8–12.

Odlyzko, A. M. (1985). Applications of Symbolic Algebra to Mathematics, In "Applications of Computer Algebra" (R. Pavelle, ed.), pp. 95–111, Kluwer-Nijhoff, Boston.

Saaty, T. L., and Kainen, P. C. (1977). "The Four-Color Problem," McGraw-Hill, New York.



Computer-Generated Proofs of Mathematical Theorems

David M. Bressoud
Macalester College

I. The Ideal versus Reality
II. Hypergeometric Series Identities
III. The WZ Method
IV. Extensions, Future Work, and Conclusions

GLOSSARY

Algorithm A recipe or set of instructions that, when followed precisely, will produce a desired result.

Binomial coefficient The binomial coefficient $\binom{n}{k}$ is the coefficient of x^k in the expansion of $(1+x)^n$. It counts the number of ways of choosing k objects from a set of n objects.

Computer algebra system A computer package that enables the computer to do symbolic computations such as algebraic simplification, formal differentiation, and indefinite integration.

Diophantine equation An equation in several variables for which only integer solutions are accepted.

Hypergeometric series A finite or infinite summation, $1+a_1+\cdots+a_k+\cdots$, in which the first term is 1 and the ratio of successive summands, a_{k+1}/a_k, is a quotient of polynomials in k.

Hypergeometric term A function of, say, k, that is the kth summand in a hypergeometric series.

Proof certificate A piece of information about a mathematical statement that makes it possible to prove the statement easily and quickly.

Proper hypergeometric term A function of two variables such as n and k that is of the following form: a polynomial in n and k times $x^k y^n$ for some fixed x times a product of quotients of factorials of the form $(an+bk+c)!$ where a, b, and c are fixed integers.

Rising factorial A finite product of numbers in an arithmetic sequence with difference 1. It is written as $(a)_n = a(a+1)(a+2)\cdots(a+n-1)$.

IN A COMPUTER-BASED PROOF, the computer is used as a tool to help guess what is happening, to check cases, to do the laborious computations that arise. The person who is creating the proof is still doing most of the work. In contrast, a computer-generated proof is totally automated. A person enters a carefully worded mathematical statement for which the truth is in doubt, hits the RETURN key, and within a reasonable amount of time the computer responds either that the statement is true or that it is false. A step beyond this is to have the computer do its own searching for reasonable statements that it can test.

Such fully automated algorithms for determining the truth or falsehood of a mathematical statement do exist.

With Doron Zeilberger's program EKHAD, one can enter the statement believed or suspected to be correct. If it is true, the computer will not only tell you so, it is capable of writing the paper ready for submission to a research journal. Even the search for likely theorems has been automated. A good deal of human input is needed to set parameters within which one is likely to find interesting results, but computer searches for mathematical theorems are now a reality.

The possible theorems to which this algorithm can be applied are strictly circumscribed, so narrowly defined that there is still a legitimate question about whether this constitutes true computer-generated proof or is merely a powerful mathematical tool. What is not in question is that such algorithms are changing the kinds of problems that mathematicians need to think about.

I. THE IDEAL VERSUS REALITY

A. What Cannot Be Done

Mathematics is frequently viewed as a formal language with clearly established underlying assumptions or axioms and unambiguous rules for determining the truth of every statement couched in this language. In the early decades of the twentieth century, works such as Russell and Whitehead's *Principia Mathematica* attempted to describe all mathematics in terms of the formal language of logic. Part of the reason for this undertaking was the hope that it would lead to an algorithmic procedure for determining the truth of each mathematical statement. As the twentieth century progressed, this hope receded and finally vanished. In 1931, Kurt Gödel proved that no axiomatic system comparable to that of Russell and Whitehead could be used to determine the truth or falsehood of every mathematical statement. Every consistent system of axioms is necessarily incomplete.

One broad class of theorems deals with the existence of solutions of a particular form. Given the mathematical problem, the theorem either exhibits a solution of the desired type or states that no such solution exists. In 1900, as the tenth of his set of twenty-three problems, David Hilbert challenged the mathematical community: "Given a Diophantine equation with any number of unknown quantities and with rational integral numerical coefficients: To devise a process according to which it can be determined by a finite number of operations whether the equation is solvable in rational integers." A well-known example of such a Diophantine equation is the Pythagorean equation, $x^2 + y^2 = z^2$, with the restriction that we only accept integer solutions such as $x = 3$, $y = 4$, and $z = 5$. Another problem of this type is Fermat's Last Theorem. This theorem asserts that no such positive integer solutions exist for the equation $x^n + y^n = z^n$ when n is an integer greater than or equal to 3. We know that the last assertion is correct, thanks to Andrew Wiles.

For a Diophantine equation, if a solution exists then it can be found in finite (though potentially very long) time just by trying all possible combinations of integers, but if no solution exists then we cannot discover this fact just by trying possibilities. A proof that there is no solution is usually very hard. In 1970, Yuri Matijasevič proved that Hilbert's algorithm could not exist. It is impossible to construct an algorithm that, for every Diophantine equation, is able to determine whether it does or does not have a solution.

There have been other negative results. Let E be an expression that involves the rational numbers, π, $\ln 2$, the variable x, the functions sine, exponential, and absolute value, and the operations of addition, multiplication, and composition. Does there exist a value of x where this expression is zero? As an example, is there a real x for which

$$e^x - \sin(\pi \ln 2) = 0?$$

For this particular expression the answer is "yes" because $\sin(\pi \ln 2) > 0$, but in 1968, Daniel Richardson proved that it is impossible to construct an algorithm that would determine in finite time whether or not, for every such E, there exists a solution to the equality $E = 0$.

B. What Can Be Done

In general, the problem of determining whether or not a solution of a particular form exists is extremely difficult and cannot be automated. However, there are cases where it can be done. There is a simple algorithm that can be applied to each quadratic equation to determine whether or not it has real solutions, and if it does, to find them.

That $x^2 - 4312x + 315 = 0$ has real solutions may not have been explicitly observed before now, but it hardly qualifies as a theorem. The theorem is the statement of the quadratic formula that sits behind our algorithm. The conclusion for this particular equation is simply an application of that theorem, a calculation whose relevance is based on the theory.

But as the theory advances and the algorithms become more complex, the line between a calculation and a theorem becomes less clear. The Risch algorithm is used by computer algebra systems to find indefinite integrals in Liouvillian extensions of difference fields. It can answer whether or not an indefinite integral can be written in a suitably defined closed form. If such a closed form exists, the algorithm will find it. Most people would still classify a specific application of this algorithm as a calculation, but it is no longer always so clear-cut. Even definite integral evaluations can be worthy of being called theorems.

Freeman Dyson conjectured the following integral evaluation for positive integer z in 1962:

$$(2\pi)^{-n}\int_0^{2\pi}\cdots\int_0^{2\pi}\prod_{1\le j<k\le n}\left|e^{i\theta_j}-e^{i\theta_k}\right|^{2z}d\theta_1\cdots d\theta_n = \frac{(nz)!}{(z!)^n}.$$

Four proofs have since been published. Dyson's conjecture cannot be proven by the Risch or any other general integral evaluation algorithm because the dimension of the space over which the integral is taken is a variable, but its proof is now close to the boundary of what can be totally automated.

Most of this article will focus on the WZ method developed by Wilf and Zeilberger in the early 1990s. Given a suitable hypergeometric series, the WZ method will determine whether or not it has a closed form. If it does, the algorithm will find it. It can even be used to find new hypergeometric series that can be expressed in closed form. Again, the important mathematics is the theory that is used to create and justify the algorithm, but specific applications now look very much like theorems. One example of a result that can be proved by the WZ method is the following identity, discovered and proved by J. C. Adams in the nineteenth century. Let $P_n(x)$ be the Legendre polynomial defined by

$$P_n(x) := \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}^2(x-1)^k(x+1)^{n-k},$$

and let $A_k = \binom{2k}{k}$, then

$$\int_{-1}^{1}P_m(x)P_n(x)P_{m+n-2k}(x)\,dx = \frac{1}{(m+n+1/2-k)}\cdot\frac{A_kA_{m-k}A_{n-k}}{A_{m+n-k}}. \qquad (1)$$

Note that the term-by-term integration is not difficult for a computer algebra system. What distinguishes this particular identity is that the number of terms in each summation is left as a variable.

Given a hypergeometric series, the WZ method can be used to find the closed expression that it equals, provided such an expression exists. We take as an example

$$f(n) = \sum_{0\le k\le n/3}2^k\frac{n}{n-k}\binom{n-k}{2k}.$$

The algorithm produces a recursion satisfied by $f(n)$:

$$f(n+3)-2f(n+2)+f(n+1)-2f(n)=0.$$

This is a particularly nice example because the coefficients are constants and standard techniques can be applied to discover that

$$\sum_{0\le k\le n/3}2^k\frac{n}{n-k}\binom{n-k}{2k} = 2^{n-1}+\frac{1}{2}(i^n+(-i)^n), \qquad n\ge 2.$$

In general, the coefficients in the recursion will be polynomials in n. In 1991 Marko Petkovšek created an algorithm that will find a closed form solution for such a recursion, or prove that no such formula exists. The combination of the WZ method with Petkovšek's algorithm gives an automated proof that a particular type of solution cannot exist, or else it finds such a solution. As an example, there is a computer-generated proof of the fact that

$$\sum_{k=0}^{n}\binom{n}{k}^2\binom{n+k}{k}^2$$

cannot be written as a linear combination of hypergeometric terms in n.

The WZ method combined with Petkovšek's algorithm is producing fully automated proofs of results that, until recently, have required considerable human ingenuity. Significantly, it replies not just with a statement that a particular identity is true, but also with a proof certificate, a critical insight that enables anyone with pencil and paper and a little time to verify that this identity is correct. At the very least, these algorithms have moved the line of demarcation between what constitutes a proof and what is only a computation.

II. HYPERGEOMETRIC SERIES IDENTITIES

A. What Is a Hypergeometric Series?

A series, $1+a_1+a_2+a_3+\cdots$, is called hypergeometric if the ratio of consecutive terms, a_{n+1}/a_n, is a rational function of n, say $a_{n+1}/a_n = P(n)/Q(n)$, where P and Q are polynomials. Most of the commonly encountered power series are hypergeometric or can be expressed in terms of hypergeometric series (see Fig. 1). A hypergeometric term is a function of n that is a summand of a hypergeometric series indexed by n. In particular, a hypergeometric term is of the form

$$a_k = \prod_{n=0}^{k-1}\frac{a_{n+1}}{a_n} = \prod_{n=0}^{k-1}\frac{P(n)}{Q(n)},$$

for some pair of polynomials P and Q.

If we factor P and Q,

$$P(n) = c_1(n+\alpha_1)(n+\alpha_2)\cdots(n+\alpha_m),$$

$$Q(n) = c_2(n+\beta_1)(n+\beta_2)\cdots(n+\beta_{n+1}),$$

Exponential function:

$$e^x = 1 + \sum_{n=1}^{\infty} \frac{x^n}{n!}, \quad \frac{a_{n+1}}{a_n} = \frac{x}{n+1},$$

Sine function:

$$\sin x = x\left(1 + \sum_{n=1}^{\infty} \frac{(-1)^n x^{2n}}{(2n+1)!}\right), \quad \frac{a_{n+1}}{a_n} = \frac{-x^2}{4(n+1)(n+3/2)},$$

Bessel function of the first kind:

$$J_k(x) = \frac{x^k}{\Gamma(k+1)}\left(1 + \sum_{n=1}^{\infty} \frac{(-1)^n x^{2n}}{4^n\, n!\, (k+1)_n}\right), \quad \frac{a_{n+1}}{a_n} = \frac{-x^2}{4(n+1)(n+k+1)},$$

Error function:

$$\operatorname{erf}(x) = \frac{2x}{\sqrt{\pi}}\left(1 + \sum_{n=1}^{\infty} \frac{(-1)^n x^{2n}}{(2n+1)\, n!}\right), \quad \frac{a_{n+1}}{a_n} = -x^2\, \frac{(2n+1)\, n}{(2n+3)(n+1)}.$$

FIGURE 1 Examples of common functions expressed in terms of hypergeometric series.

then the hypergeometric term can be written as

$$a_k = c\, \frac{(\alpha_1)_k (\alpha_2)_k \cdots (\alpha_m)_k}{(\beta_1)_k (\beta_2)_k \cdots (\beta_{n+1})_k},$$

where $c = c_1/c_2$ and $(\alpha)_k$ is the rising factorial:

$$(\alpha)_k = \alpha(\alpha+1)(\alpha+2)\cdots(\alpha+k-1).$$

B. The Chu–Vandermonde Identity

A large part of the impetus behind the development of the WZ method and the reason why it has become such an influential tool is that there is a rich and ever-expanding store of useful identities for hypergeometric series. These recur throughout mathematics, playing important roles in the solutions of both theoretical and applied problems.

The binomial theorem was the first and is the most fundamental of these identities. It is the foundation upon which all others are proved. Mathematicians have been building upon the binomial theorem for many years. In 1303, Chu Shih-Chieh wrote *Precious Mirror of the Four Elements* (*Ssu Yü Chien*), in which he may have been the first person to state the fundamental result:

$$\sum_{i=0}^{\infty} \binom{a}{i}\binom{b}{k-i} = \binom{a+b}{k}. \tag{2}$$

In Chu's identity, a, b, and k are positive integers. Note that all summands will be zero once i is greater than a or k. Equation (2) is easily derived from the binomial theorem by comparing the coefficients of x^k in

$$(1+x)^a (1+x)^b = \sum_{i=0}^{a} \binom{a}{i} x^i \sum_{j=0}^{b} \binom{b}{j} x^j,$$

and

$$(1+x)^{a+b} = \sum_{k=0}^{a+b} \binom{a+b}{k} x^k.$$

Equation (2) was rediscovered by Alexandre Vandermonde in 1772 and is today known as the *Chu–Vandermonde Identity*.

The ratio of successive terms in the summation is

$$\binom{a}{n+1}\binom{b}{k-n-1} \Big/ \binom{a}{n}\binom{b}{k-n} = \frac{(n-a)(n-k)}{(n+1)(n+b-k+1)}.$$

If we divide both sides of Eq. (2) by the first summand, $\binom{b}{k}$, it can be rewritten in terms of rising factorials as

$$1 + \sum_{n=1}^{\infty} \frac{(-a)_n (-k)_n}{n!(b-k+1)_n} = \frac{(a+b)!\,(b-k)!}{(a+b-k)!\,b!}. \tag{3}$$

In 1797, Johann Friedrich Pfaff showed that, subject only to convergence conditions, Eq. (3) holds for complex values, in which case it can be expressed as

$$1 + \sum_{n=1}^{\infty} \frac{(\alpha)_n (\beta)_n}{n!(\gamma)_n} = \frac{\Gamma(\gamma-\alpha-\beta)\Gamma(\gamma)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)}. \tag{4}$$

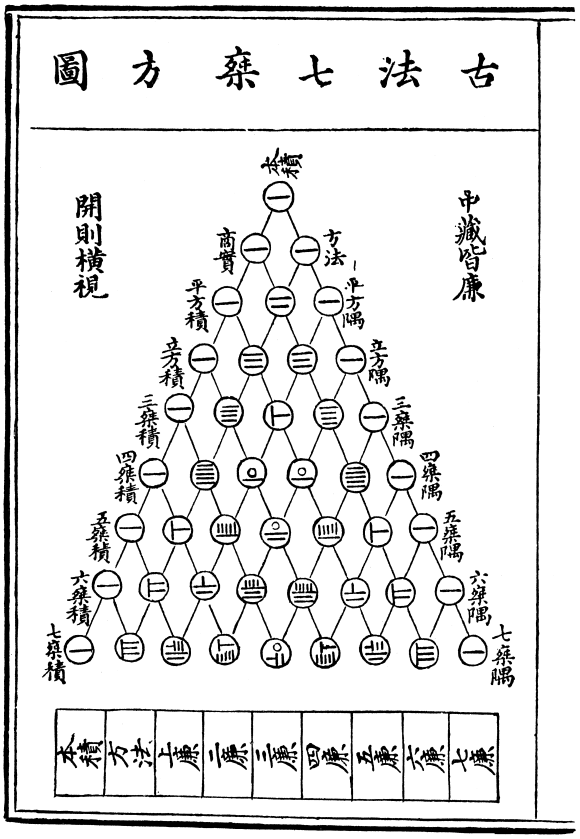


FIGURE 2 The representation of "Pascal's" triangle in Chu's *Precious Mirror of the Four Elements* of 1303. (Reprinted with the permission of Cambridge University Press.)

Pfaff's student, Carl Friedrich Gauss, used hypergeometric series in his astronomical work and advanced their study. Among his contributions, he found sharp criteria for whether or not a hypergeometric series converges. Throughout the nineteenth and twentieth century, a great number of identities for hypergeometric series were discovered, many of which were collected in the Bateman Manuscript Project published as *Higher Transcendental Functions* in 1953–1955.

C. Standardized Notation

Most hypergeometric series can be written as sums of rational products of binomial coefficients, but this representation is problematic because it is not unique. As an example,

$$\sum_{n=0}^{m} 2^{m-k-2n}\binom{m}{n}\binom{m-n}{n+k} = \binom{2m}{m+k}$$

appears to be different from the Chu-Vandermonde identity [Eq. (2)]. But if we look at the ratio of consecutive summands, it is

$$\frac{1}{4}\frac{(m-2n-k-1)(m-2n-k)}{(n+1)(n+k+1)}$$

$$= \frac{(n+(k+1-m)/2)(n+(k-m)/2)}{(n+1)(n+k+1)}.$$

This is simply the Chu-Vandermonde identity with $\alpha = (k+1-m)/2$, $\beta = (k-m)/2$, and $\gamma = k+1$.

There is clearly an advantage to using the rising factorial notation, in which case we write

$${}_mF_n\left(\begin{matrix}\alpha_1,\ldots,\alpha_m\\ \beta_1,\ldots,\beta_n\end{matrix};\, x\right) := 1+\sum_{k=1}^{\infty}\frac{(\alpha_1)_k\cdots(\alpha_m)_k}{k!(\beta_1)_k\cdots(\beta_n)_k}\,x^k.$$

Even with this standardized notation, there are equivalent identities that look different because there are nontrivial transformation formulas for hypergeometric series. As an example, provided the series in question converge, we have that

$${}_2F_1\left(\begin{matrix}a,\ b\\ 2a\end{matrix};\, x\right) = \left(1-\frac{x}{2}\right)^{-b}$$

$$\times\,{}_2F_1\left(\begin{matrix}b/2,\ (b+1)/2\\ a+1/2\end{matrix};\, \left(\frac{x}{2-x}\right)^2\right).$$

This is why, even if all identities for hypergeometric series were already known, it would not be enough to have a list of them against which one could compare the candidate in question. Just establishing the equivalence of two identities can be a very difficult task. This makes the WZ method all the more remarkable because it is independent of the form in which the identity is given and can even be used to verify (or disprove) a conjectured transformation formula.

III. THE WZ METHOD

A. Sister Celine's Technique

The WZ method for finding and proving identities for hypergeometric series builds on a succession of developments that began with the Ph.D. thesis of Sister Mary Celine Fasenmyer at the University of Michigan in 1945. We consider a sum of the form

$$f(n) = \sum_k F(n,k),$$

where $F(n,k)$ is a proper hypergeometric term. This means that it is a polynomial in n and k times $x^k y^n$, for fixed x and y, times a product of quotients of factorials of the form $(an+bk+c)!$, where a and b, and c are fixed integers. As an example,

$$\sum_{0\le k\le n/3} 2^k \frac{n}{n-k}\binom{n-k}{2k} = \sum_{0\le k\le n/3} n\cdot 2^k \cdot \frac{(n-k-1)!}{(2k)!\,(n-3k)!}.$$

is such a series.

Every such sum of proper hypergeometric terms will satisfy a finite recursion of the form

$$\sum_{j=0}^{J} a_j(n) f(n+j) = 0.$$

Sister Celine showed how to reduce the problem of finding these coefficients to one of solving a system of linear equations. It was Doron Zeilberger who realized that this gives us an algorithm for proving hypergeometric series identities because we need only verify that each side satisfies the same recursion and the same initial conditions. The problem with using Sister Celine's approach is that her particular algorithm for finding the coefficients is slow. Later developments would speed it up considerably, though in the process would lose the easy generalization of Sister Celine's technique to summations over several indices.

B. Gosper's algorithm

In 1977 and 1979, R. W. Gosper, Jr., took a different approach and became one of the first people to use computers to discover and check identities for hypergeometric series. Given a proper hypergeometric term $F(n,k)$, Gosper showed how to automate a search for a proper hypergeometric term $G(n,k)$ with the property that

$$G(n,k+1) - G(n,k) = F(n,k).$$

If such a G could be found, then

$$\begin{aligned} f(n) &= \sum_{k=0}^{n} (G(n,k+1) - G(n,k)) \\ &= G(n,n+1) - G(n,0). \end{aligned}$$

An example of the application of this algorithm is the computer-generated proof of an identity discovered and first proved by J. S. Lomont and John Brillhart: Let $1 \le m \le n$, where $n \ge 2$ and $1 \le s \le \min(m, n-1)$, then

$$\sum_{j=0}^{s} \left[(-1)^j (m+n-2j)\binom{m}{j}\binom{m-j}{m-s}\binom{n}{j}\binom{n-j}{n-s} \times \binom{m+n}{j}\binom{m+n-s-j-1}{s-j} \Big/ \binom{s}{j}^2 \right] = 0.$$

Given this conjecture, the program EKHAD replies with the proof certificate:

$$-s,\ j * (m+n-s-j)/(m+n-2*j).$$

This means that if $F(n,j)$ is the summand in the conjectured identity, then

$$-sF(n,j) = G(n,j+1) - G(n,j),$$

where $G(n,j) = j(m+n-s-j)F(n,j)/(m+n-2j)$. The sum over j of $G(n,j+1) - G(n,j)$ telescopes, and therefore the original summation equals $[G(n,s+1) - G(n,0)]/(-s) = 0$.

Gosper's algorithm is a fertile approach that is often applicable, but it is limited by the fact that such a G does not always exist.

C. Wilf and Zeilberger

Major progress was made by Doron Zeilberger who, starting in 1982, began to combine the ideas of Sister Celine and William Gosper. In the early 1990s, Herbert Wilf joined Zeilberger in extending and refining these methods into a fully automated proof machine that is now known as the WZ method. If $F(n,k)$ is a proper hypergeometric term, then there always is a proper hypergeometric term $G(n,k)$ such that $G(n,k+1) - G(n,k)$ is equal to a linear combination of $\{F(n+j,k) \mid 0 \le j \le J\}$ for some explicitly computable J,

$$\sum_{j=0}^{J} a_j(n) F(n+j,k) = G(n,k+1) - G(n,k), \qquad (5)$$

where the $a_j(n)$ are polynomials in n. If $f(n) = \sum_{k=0}^{K} F(n,k)$, then we can sum both sides of Eq. (5) over $0 \le k \le K$. The right side telescopes, and we are left with the recursive formula

$$\sum_{j=0}^{J} a_j(n) f(n+j) = G(n,K+1) - G(n,0).$$

Gosper's technique—which is very fast—can be used to find the function G. The coefficients $a_j(n)$ are then found by solving a system of linear equations.

Gosper's algorithm is the special case of the WZ method in which $J = 0$. The other case of particular interest is when $J = 1$ and $a_1 = -a_0 = 1$. Consider the conjectured identity:

$$\sum_k \binom{n}{2k}\binom{2k+1}{k}\frac{n+2}{2k+1}2^{n-2k-1} = \binom{2n+1}{n}. \qquad (6)$$

If we divide each side by $\binom{2n+1}{n}$, this can be rewritten as

$$f(n) = \sum_k \binom{n}{2k}\binom{2k+1}{k}\frac{n+2}{2j+1}2^{n-2k-1} \Big/ \binom{2n+1}{n} = 1.$$

If this is true, then $f(n)$ satisfies the recursion $f(n+1)-f(n)=0$. Let $F(n,k)$ be the summand,

$$F(n,k)=\binom{n}{2k}\binom{2k+1}{k}\frac{n+2}{2j+1}2^{n-2k-1}\Big/\binom{2n+1}{n}.$$

If we could find a proper hypergeometric term $G(n,k)$ for which

$$F(n+1,k)-F(n,k)=G(n,k+1)-G(n,k), \quad (7)$$

then it would follow that $f(n+1)-f(n)=0$, and so $f(n)$ would be constant. It would be enough to check that $f(0)=1$.

In fact, such a G does exist. The WZ method finds it. The proof certificate is the rational function,

$$\frac{G(n,k)}{F(n,k)}=\frac{4k(k+1)}{(2n+3)(2k-n-1)}.$$

To check Eq. (6), we only need to verify that F and G, which we now know, do indeed satisfy Eq. (7).

In general, the WZ method returns either the ratio $G(n,k)/F(n,k)$ [if the recursion is of the form given in Eq. (7)], or it returns the actual recursive formula satisfied by $f(n)$. The only drawback to the WZ method is that the number of terms in the recursive formula may be too large for practical use.

D. Petkovšek and Others

In his Ph.D. thesis of 1991, Marko Petkovšek showed how to find a closed form solution—or to show that such a solution does not exist—for any recursive formula of the form

$$\sum_{j=0}^{J}a_j(n)f(n+j)=g(n),$$

in which $g(n)$ and the $a_j(n)$ are polynomials in n. By closed form, we mean a linear combination of a fixed number of hypergeometric terms.

Combined with the WZ method, Petkovšek's algorithm implies that in theory if not in practice, given any summation of proper hypergeometric terms, there is a completely automated computer procedure that will either find a closed form for the summation or prove that no such closed from exists. A full account of the WZ method and Petkovšek's algorithm is given in the book $A=B$ by Petkovšek, Wilf, and Zeilberger.

Others have worked on implementing and extending the ideas of the WZ method. One of the centers for this work has been a group headed by Peter Paule at the University of Linz in Austria. Ira Gessel has been at the forefront of those who have used this algorithm to implement computer searches that both discovered and proved a large number of new identities for hypergeometric series.

IV. EXTENSIONS, FUTURE WORK, AND CONCLUSIONS

A. Extensions and Future Work

All of the techniques described in this article have been extended to q-hypergeometric series such as

$$1+\sum_{k=0}^{n}q^{k^2}\frac{(1-q^n)(1-q^{n-1})\cdots(1-q^{n-k+1})}{(1-q)(1-q^2)\cdots(1-q^k)},$$

in which the ratio of consecutive summands is a rational function of q^k.

Many general determinant evaluations can be reduced to problems that can be solved using the WZ method. Work is progressing on fully automating proofs of such results.

There are other theorems that appear to be amenable to an automated computer attack. These include results on real closed fields using techniques of George Collins and geometrical theorems proved using algebraic techniques such as Gröbner bases.

B. Conclusions

The net effect of the algorithms that prove identities for hypergeometric series is that a piece of mathematics that once could only be done by those with cleverness and insight has been turned into a purely mechanical calculation. Rather than limiting the scope of mathematics, the WZ method has widened it. Problems that had once been intractable are now within reach. The situation is different in degree but not in kind from the invention of calculus. This was the discovery of mechanical procedures that enabled scientists to shift their attention away from laborious and ingenious techniques for finding areas and tangent lines and to begin addressing the really interesting questions.

Perhaps this will always be the fate of computer-generated proofs. That one class of problems has been moved into the category of those that can be solved by computers means that we are freed to direct our attention to those questions that are most important.

Web Sites for the WZ Method and Related Algorithms

Home page for the book $A=B$:
http://www.cis.upenn.edu/~wilf/AeqB.html
Wilf and Zeilberger's programs:
http://www.cis.upenn.edu/~wilf/progs.html
Programs of the RISC group at the University of Linz:
http://www.risc.uni-linz.ac.at/research/combinat/risc/

SEE ALSO THE FOLLOWING ARTICLES

COMPUTER ALGORITHMS • COMPUTER ARCHITECTURE • DISCRETE MATHEMATICS AND COMBINATORICS • GRAPH THEORY • MATHEMATICAL LOGIC • PROBABILITY

BIBLIOGRAPHY

Collins, G. E. (1975). *In* "Automata Theory and Formal Languages," Lecture Notes in Computer Science, (H. Brakhage, ed.) Vol. 33, pp. 134–183, Springer, Berlin.

Hilbert, D. (1902). "Mathematical Problems," *Bulletin of the American Mathematical Society* **8**, pp. 1–34. (Translation of the original German article.)

Nemes, I., Petkovšek, M., Wilf, H., and Zeilberger, D. (1997). "How to do MONTHLY problems with your computer," *American Mathematical Monthly* **104,** 505–519.

Petkovšek, M., Wilf, H., and Zeilberger, D. (1996). "$A = B$," A K Peters, Wellesley, MA.

Wilf, H., and Zeilberger, D. (1992). "An algorithmic proof theory for hypergeometric (ordinary and "q") multisum/integral identities," *Inventiones Mathematicae* **108,** 575–633.



Convex Sets

A. C. Thompson
Dalhousie University

I. Introduction
II. Definitions
III. Examples
IV. Descriptions of Convex Sets
V. New Convex Sets from Old
VI. Spaces of Convex Sets
VII. Basic Theorems
VIII. Volumes and Mixed Volumes
IX. Inequalities
X. Special Classes of Convex Sets

GLOSSARY

Affine map The composition of a linear map and a translation; i.e., if T is linear, then $A(x) := T(x + x_0) = T(x) + T(x_0)$ is an affine map.

Compact set A closed and bounded set in $\mathbb{R}^n$; in a metric space, a set C such that if (x_k) is a sequence of elements of C there is a subsequence that converges to a point of C.

Dual space The collection of all linear functions from a vector space to the set of real numbers. These functions can be added and multiplied by real numbers in a point-by-point fashion which makes this collection into another vector space.

Hyperplane A *level set* of a function in the dual space; i.e., if f is a linear function from a vector space X to $\mathbb{R}$ and if α is a number, then $H_f^\alpha := \{x \in X : f(x) = \alpha\}$ is a typical hyperplane.

Line segment The set of points (vectors) $\{x : x = a + \lambda(b - a),\ \ 0 \le \lambda \le 1\} = \{x : x = (1 - \lambda)a + \lambda b,\ \ 0 \le \lambda \le 1\}$ is called the *line segment* joining a and b and is denoted by $[a, b]$.

Linear map (transformation) A function T between vector spaces that respects the vector operations of addition and multiplication by numbers; i.e., $T(\alpha x + \beta y) = \alpha T(x) + \beta T(y)$.

$\mathbb{R}^n$ The most usual vector spaces consisting of n-tuples of real numbers that are added and multiplied by numbers in a coordinate-by-coordinate fashion.

Vector space A collection of things called *vectors* or *points* that can be added (via a parallelogram law) and multiplied by numbers (also called *scalars*). Here the numbers will be real but in other contexts complex or other number systems are possible.

I. INTRODUCTION

Relatively few shapes in the natural world are convex. When they do occur— for example, soap bubbles, drops of dew, smoothly worn stones on the beach, single crystals of amethyst and salt— we find them to be esthetically pleasing. Among manufactured objects, rectangles, circles, hexagons, cubes, cylinders, and cones are quite ubiquitous. We first encounter them as children and enjoy the shapes of wooden building blocks and colored tiles.

Convexity is the study of these shapes. Two-dimensional convex shapes (circles, ellipses, triangles, polygons) and the regular Platonic solids have been objects of mathematical study for a very long time. The study of convexity as a specific mathematical topic dates back only to the end of the 19th century. The primary influence was the pioneering work of Minkowski for which one should consult his collected works.[3] For a good historical summary, see the article by Peter Gruber in the *Handbook of Convex Geometry*.[2] This chapter covers only some of the topics that fall under the heading of convexity. Other aspects can be found in articles in Reference 2. There is not space to mention the many interactions of convexity with other branches of mathematics. The book by Roger Webster[5] is a readable, elementary introduction to the subject and has some interesting applications.

To define convex sets precisely we need an ambient space in which they may exist. This space requires one type of mathematical structure and usually comes equipped with another.

The structure that is required is that of a *vector* or *linear space*; the most familiar vector spaces are those that we denote by $\mathbb{R}^2$, $\mathbb{R}^3$, and, in general, $\mathbb{R}^n$. This space consists of n-tuples of numbers $(\xi_1, \xi_2, \ldots, \xi_n)$ whose entries are called the *coordinates* of the vector. It is customary to write single vectors as *rows*, but matrices (and other functions that operate on vectors) are usually written to the left of the vector. This means that the vectors should 'really' be viewed as *columns*.

If a vector y is expressed in the form:

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k$$

then it is said to be a *linear combination* of the vectors $\{x_i : i = 1 \ldots, k\}$. The usual basis for $\mathbb{R}^n$ consists of the vectors:

$$e_1 = (1, 0, 0, \ldots, 0), \quad e_2 = (0, 1, 0, \ldots, 0), \quad \ldots,$$
$$e_n = (0, 0, 0, \ldots, 1)$$

where the ith vector has a 1 as the ith coordinate and a 0 elsewhere. Any vector $x = (\xi_1, \xi_2, \ldots, \xi_n)$ can be expressed uniquely as:

$$x = \xi_1 e_1 + \xi_2 e_2 + \cdots + \xi_n e_n.$$

A set of vectors $\mathcal{B}$ with the property that every vector has a unique representation as a finite linear combination of the vectors in $\mathcal{B}$ is called a *basis* for X. Each vector space has a basis, and all bases for the same space have the same number of elements. That number is called the *dimension* of the space. The dimension of $\mathbb{R}^n$ is n. The space is said to be *finite dimensional* if it has a basis with a finite number of elements and is *infinite dimensional* otherwise. We shall be concerned almost entirely with finite dimensional spaces.

A linear map is called an *isomorphism* if it is one to one and onto. If X is finite dimensional, then, corresponding to a basis $(x_1, x_2, \ldots, x_n)$, there is an isomorphism, T, of X onto $\mathbb{R}^n$ defined as follows: each $x \in X$ has a unique representation as $x = \sum \alpha_i x_i$; set $T(x) := (\alpha_1, \alpha_2, \ldots, \alpha_n)$. In this sense, there is no real loss of generality if, when dealing with finite dimensional spaces, we restrict attention to $\mathbb{R}^n$.

The linear mappings from X to $\mathbb{R}$ are given the special name of *linear functionals*. The set of all of them is called the *dual space* of X and denoted by X^*. If X is finite dimensional and if it is given a basis, then the linear functionals can be represented by $1 \times n$ matrices (i.e., row vectors of the same size as the column vectors from X). Thus, in the finite dimensional case, the dimension of X^* is the same as that of X.

In $\mathbb{R}^n$, the length of a vector $x = (\xi_1, \xi_2, \ldots, \xi_n)$ is denoted by $\|x\|$ and is defined to be the number:

$$\|x\| := \left(\sum_{i=1}^{n} |\xi_i|^2 \right)^{1/2}.$$

The structure that is usually present in discussions of convex sets in addition to the vector space structure just outlined is that of a *topological space*. Most frequently the topology derives from a metric or distance. This is the case in $\mathbb{R}^n$ where the distance between two vectors x and y, $d(x, y)$, is defined to be the length of $x - y$:

$$d(x, y) := \|x - y\|.$$

The topological structure allows one to talk about convergence, continuity, and such concepts as open sets, closed sets, connected sets, and compact sets. The Heine-Borel Theorem asserts that C is a compact subset of $\mathbb{R}^n$ if and only if it is both closed and bounded. This is no longer true in infinite dimensional spaces.

II. DEFINITIONS

Definition. A set K in a vector space X is said to be *convex* if whenever $x, y \in K$ then the line segment $[x, y]$ is contained in K. Thus, in Fig. 1, the set (a) is convex while (b) is not.

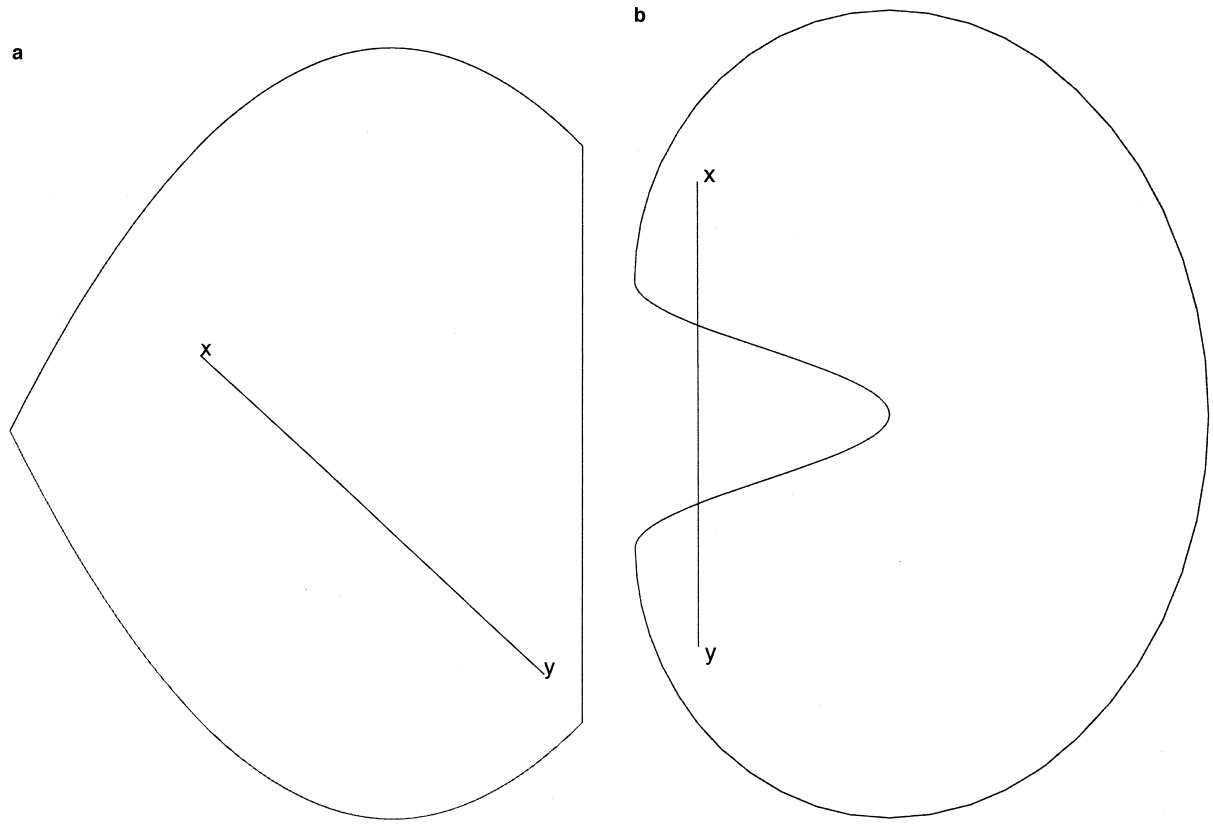


FIGURE 1

A point of the line segment $[a, b]$— a point of the form:

$$a + \lambda(b - a) = (1 - \lambda)a + \lambda b,\ 0 \le \lambda \le 1,$$

is said to be a convex combination of a and b.

A related notion is that of being *star shaped*. A set S is star shaped about a point x_0 if for all x in S the line segment $[x, x_0]$ is contained in S. In Fig. 2, the set (a) is star shaped about x but not about y, and (b) is not star shaped about any point. A set is convex if and only if it is star shaped about every point. Each convex set (and each star-shaped set) is connected.

In the one-dimensional space $\mathbb{R}$ the collections of convex sets, star-shaped sets, and connected sets coincide. Each of these is the class of intervals (closed, open, half-open, bounded, and unbounded). Therefore, in order to have an interesting theory of convexity, the dimension of the underlying space should be at least 2.

There is also a definition of convexity as an adjective that applies to functions rather than sets:

Definition. A real valued function f defined on a convex set K is said to be *convex* if, for all x and y in K,

$$f(\lambda x + (1 - \lambda)y) \le \lambda f(x) + (1 - \lambda) f(y).$$

Note that the convexity of K is needed to ensure that $\lambda x + (1 - \lambda)y$ is in the domain of f when x and y are.

The relation between the definitions of convexity for a function and for a set is twofold. The *graph* of f is a subset of $K \times \mathbb{R}$ and is defined as:

$$\text{graph}(f) := \{(x, \eta) : f(x) = \eta\}.$$

Extending this idea, the *epigraph* of f is the set that lies above the graph of f:

$$\text{epigraph}(f) := \{(x, \eta) : f(x) \le \eta\}.$$

Then, f is convex (as a function) if and only if epigraph(f) is convex as a set. Secondly, if f is convex then, for each (extended) real number α, the sets $\{x : f(x) \le \alpha\}$ and $\{x : f(x) < \alpha\}$ are convex. This illustrates the connection between convexity and certain types of inequality.

A function is sublinear if it is both subadditive—$f(x + y) \le f(x) + f(y)$—and *non-negatively homogeneous*—$f(\alpha x) = \alpha f(x)$ for all $\alpha \ge 0$. From now on, "homogeneity" will mean "non-negative homogeneity." All linear functionals are sublinear and all sublinear functions are convex. Hence, if f is sublinear then sets of the form $\{x : f(x) \le \alpha\}$ are convex.

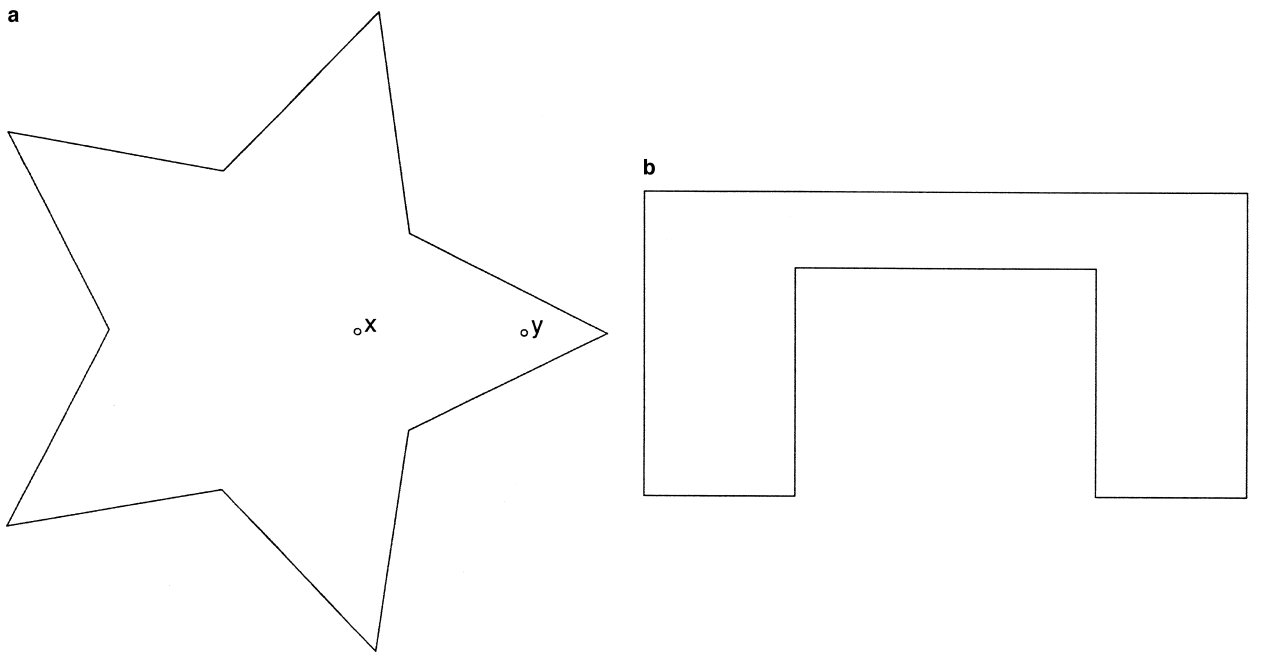


FIGURE 2

III. EXAMPLES

If K is a convex set and if A is an affine mapping, then $A(K) := \{Ax : x \in K\}$ is also convex. Affine mappings include translations, rotations, dilations, and reflections. Therefore, it is often convenient to think of examples as being located at some particular point in space (centered at the origin, for example) or with some particular orientation or with some particular scaling. These are not relevant to the property of being convex and may not be very relevant to other properties of the convex sets.

1. A single point is a convex set.

2. Lines, line segments (with or without the end points), and rays (sets of the form $\{x : x = a + \lambda b$, with $\lambda \geq 0\}$) are convex sets. As indicated in Section II, in $\mathbb{R}$ the only convex sets are intervals.

3. The ball of radius 1 centered at the origin (briefly, the unit ball),

$$B := B[0, 1] := \{x : \|x\| \leq 1\}$$

is convex. While this is geometrically clear in two and three dimensions, the proof (the same in all dimensions) is not so immediate. We need to show that the norm is a sublinear functional. That it is homogeneous is easy; the fact that it is subadditive,

$$\|x + y\| \leq \|x\| + \|y\|,$$

is usually referred to as the *triangle inequality* and is a consequence of the Cauchy-Schwarz inequality.

Any closed ball $B[x_0, r] := \{x : \|x - x_0\| \leq r\}$ of center x_0 and radius r is convex. An open ball, $B(x_0, r) := \{x : \|x - x_0\| < r\}$ is also convex (in general, the interior of a convex set is also convex). Note that we may consider, for example, a two-dimensional ball (a disc) as a subset of $\mathbb{R}^3$ or a higher dimensional space. It is still convex, but whereas as a subset of $\mathbb{R}^2$ it has interior points, in $\mathbb{R}^3$ it does not. We shall say more about the idea of relative interior in Section IV. Closed and bounded convex sets are called *convex bodies*. Some authors use this term to also imply that the set has a non-empty interior.

4. The image of the unit ball under any invertible affine map is convex. This gives the important class of convex bodies known as *ellipsoids*.

5. The *unit cube* is defined to be the set $\{x = (\xi_1, \xi_2, \ldots, \xi_n) : 0 \leq \xi_i \leq 1\}$. If a cube centered at the origin is required, we often consider one that is dilated by a factor of 2 and call it the *standard cube*, C_n:

$$C_n := \{x = (\xi_1, \xi_2, \ldots, \xi_n) : -1 \leq \xi_i \leq 1\}.$$

This set is a convex body in $\mathbb{R}^n$. The image of a standard cube under an invertible affine map is called a *parallelotope*.

6. The *standard simplex*, S_n, is defined by the following equation:

$$S_n := \left\{x = (\xi_1, \xi_2, \ldots, \xi_n) : \xi_i \geq 0 \text{ and } \sum \xi_i \leq 1\right\}.$$

A general n-simplex is the image of S_n under an invertible affine map.

7. The *standard cross-polytope* in $\mathbb{R}^n$, $\mathcal{O}_n$, is defined by:

$$\mathcal{O}_n := \left\{x = (\xi_1, \xi_2, \ldots, \xi_n) : \sum |\xi_i| \leq 1\right\}.$$

The letter $\mathcal{O}$ is used because in $\mathbb{R}^3$ this set is a regular octahedron. Note that $\mathcal{O}_2$ and C_2 are both squares but they are oriented differently.

8. A *hyperplane* $H_f^\alpha := \{x : f(x) = \alpha\}$ (where f is a linear functional) is convex. A (closed) *half-space* is a set of the form $H_f^{\alpha-} := \{x : f(x) \leq \alpha\}$ and is described as one side of the hyperplane H_f^α. This set is convex as is its interior (the open half-space for whose definition $\leq$ is replaced by $<$). One also has the half-space $H_f^{\alpha+} := \{x : f(x) \geq \alpha\} = H_{-f}^{-\alpha-}$.

9. In addition to the standard Euclidean ball B in $\mathbb{R}^n$ we may consider the ℓ_p ball $B(p)$ which is defined by:

$$B(p) := \left\{x : \sum |\xi_i|^p \leq 1\right\}.$$

If $1 \leq p$, then this set is convex.

There will be more examples of convex sets in Sections IV and V, where we discuss general methods of constructing them and of getting new sets from old ones.

IV. DESCRIPTIONS OF CONVEX SETS

How should a convex set be specified? How does one decide whether a given point is inside a particular convex set or not? There are more sophisticated computational versions of these questions: What are the most efficient algorithms for describing a convex set or for deciding whether a point is inside or not? These are difficult questions which will not be tackled here. Instead, we give a variety of answers to the more general questions, any one of which may be the best for a particular situation.

The intersection of an arbitrary collection of convex sets is convex. Since any set is contained in at least one convex set (the whole vector space in which it sits), it follows that any set, A, is contained in a smallest convex set, namely the intersection of *all* the convex sets that contain A. It is called the *convex hull* of A and is written co$A$. Thus,

$$\text{co}A := \bigcap K$$

where the intersection is taken over all convex sets K with $A \subseteq K$. One visualizes the convex hull as the set obtained by stretching a rubber sheet around the set A or, in two dimensions, an elastic band (see Fig. 3).

Therefore, one may describe a convex set either as the intersection of certain simpler convex sets or as the convex hull of some simpler set.

The convex hull of a finite set is called a *convex polytope*. Sometimes more general types of polytopes may be considered but here the word will always denote a closed, convex set and extra adjectives will be dropped.

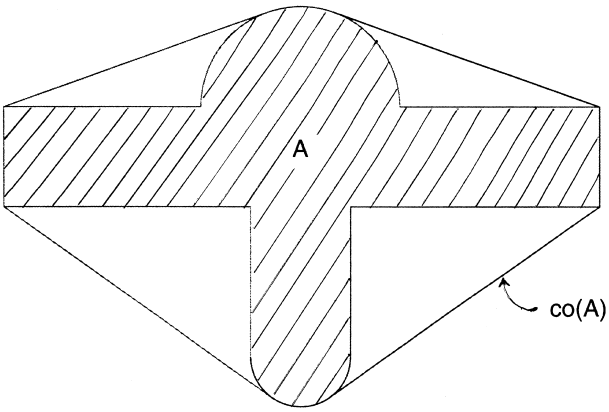


FIGURE 3

Alternatively, a polytope may be described as the intersection of finitely many closed half-spaces. There is a difference between the two notions. The convex hull of finitely many points is always bounded; the intersection of half-spaces may not be. A bounded polytope that has an interior may be described either by the points of which it is the convex hull or by the bounding hyperplanes. This is the first example of the duality relationship discussed in Section V.

Examples. The standard simplex is the convex hull of the finite set $\{0, e_1, e_2, \ldots, e_n\}$. The standard octahedron is the convex hull of the finite set $\{\pm e_1, \pm e_2, \ldots, \pm e_n\}$. The standard cube is the intersection of the following half-spaces: $\{x : f_i(x) \leq 1\}$ and $\{x : -f_i(x) \leq 1\}$, where $f_i(x) = \xi_i$ and $i = 1, 2, \ldots, n$.

Here we digress to discuss the dimension of a convex set. An affine combination of vectors $\{x_1, x_2, \ldots, x_k\}$ is a linear combination $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k$ in which $\sum \alpha_i = 1$. If K is a convex set, then the set of all affine linear combinations of elements of K is called the *affine hull* of K. If $0 \in K$, then the affine hull of K is the same as the set of all linear combinations (because one can add a suitable multiple of 0 to make the coefficients sum to 1). Hence, in this case, the affine hull of K is a subspace (containing K). In the general case, the affine hull of K is a flat (the older term is *affine variety*)—i.e., the translate of a subspace. Subspaces and hence flats have a well-defined dimension. The dimension of a convex set is the dimension of its affine hull.

Related to this notion are the following concepts. First, a finite set of points with n elements is said to be in general position if its affine hull (or, equivalently, its convex hull) has dimension $(n-1)$ which is the maximum possible. Two distinct points are always in general position, three points if they are not collinear, four points if they are not

coplanar, and so on. The convex hull of $(n+1)$ points in general position (necessarily in $\mathbb{R}^m$ with $m \geq n$) is an n-simplex. To see that there is an affine map of this set onto S_n observe that there is a translation that takes one point to 0 and then a linear map that takes the remaining n to the usual basis vectors. Second, the relative interior of K is the interior when K is regarded as a subset of its affine hull. The relative interior of a non-empty convex set is always non-empty.

Definition. A *face* F of a convex set K is a convex subset of K with the property that if y is in F and if y can be represented in the form $y = \alpha x_1 + (1-\alpha)x_2$ with x_1 and x_2 in K, then, in fact, x_1 and x_2 are in F. A more geometrical description is that if an open line segment in K contains points of F then the whole line segment lies in F. If P is an n-dimensional polytope in $\mathbb{R}^n$, then its 0-dimensional faces are called *vertices*, its one-dimensional faces are called *edges*, and the $(n-1)$ dimensional faces will be called *facets* (this latter term is not universally used).

Examples. The faces of the standard cube C_3 in $\mathbb{R}^3$ are the cube itself, the six facets of the cube (the sets where one fixed coordinate has a prescribed value from $\{-1, 1\}$), the 12 edges of the cube (the sets where two fixed coordinates have prescribed values from $\{-1, 1\}$), and the eight vertices (the sets where all three coordinates have prescribed values from $\{-1, 1\}$).

The faces of the standard simplex S_3 in $\mathbb{R}^3$ are S_3; the four facets of the simplex (the intersections of S_3 with the planes $\{x : \xi_i = 0\}$ and $\{x : \sum \xi_i = 1\}$); the six edges (the line segments $[0, e_i]$, $[e_1, e_2]$, $[e_2, e_3]$, $[e_3, e_1]$); and the four vertices $\{0\}$, $\{e_i\}$.

Faces of a convex set K that are single points $\{z\}$ are called *extreme points* of K.

Theorem (Minkowski). A convex body in $\mathbb{R}^n$ is the convex hull of its extreme points.

There is an infinite dimensional extension of this theorem due to Krein and Milman which states that each compact convex set in a normed linear space is the closed convex hull of its extreme points.

A face of a convex set that is a ray is called an *extreme ray*. Another generalization of Minkowski's theorem is that any closed convex set that contains no lines is the convex hull of its extreme points and extreme rays.

The set of extreme points of a compact set need not be closed. This is shown by the following example.

Example. In $\mathbb{R}^3$ let K be the convex hull of the circle $\{x : \xi_1^2 + \xi_2^2 = 1\}$ and the points $(1, 0, 1)$ and $(1, 0, -1)$. K looks like a double slanted cone. The extreme points are $(1, 0, 1)$, $(1, 0, -1)$, and all the points of the circle except $(1, 0, 0)$.

The next way to describe a convex set is to first suppose that the origin is an interior point of the set. This effectively means that the set has interior points because one can always either translate the set or choose the origin appropriately. Then one describes the set by saying how far in any direction the boundary is from the origin.

Definition. If K is a convex set with 0 as an interior point, then the *radial function* of K, $r_K(x)$, is defined by:

$$r_K(x) := \sup\{\lambda : \lambda x \in K\}.$$

A slight variant of this is to say how much K has to be dilated to contain a given vector.

Definition. If K is a convex set with 0 as an interior point, then the *gauge function* (or Minkowski functional) of K, $g_K(x)$, is defined by:

$$g_K(x) := \inf\{\lambda \geq 0 : x \in \lambda K\}.$$

It is evident that $1/r_K(x) = g_K(x)$ (with $1/\infty = 0$ if K is unbounded). The function g_K has the advantage of being both homogeneous and subadditive (i.e., sublinear). If K is bounded, then the radial function is finite for all $x \neq 0$ and the gauge function is non-zero for $x \neq 0$. If, in addition, K is symmetric about the origin so that $g_K(-x) = g_K(x)$, then the gauge function has the properties of a norm. This leads to the use of convexity in the study of normed spaces.

If K is closed, it is described as $K = \{x : g_K(x) \leq 1\}$. The boundary of K (denoted by ∂K) is the set of points for which $g_K(x) = r_K(x) = 1$.

Instead of thinking of the boundary of K as a set of points, we may also regard it as an envelope of half-spaces (in the same way that a polytope was described as a finite intersection of half-spaces).

Definition. A hyperplane H_f^α is said to be a *support hyperplane* for a convex body K if (1) $K \cap H_f^\alpha \neq \emptyset$, and (2) K is contained in one of the half-spaces $H_f^{\alpha+}$, $H_f^{\alpha-}$.

If K is a convex body, if f is a linear functional, and if $\alpha := \max\{f(x) : x \in K\}$, then H_f^α is a support hyperplane of K. This leads to the idea of the support function of K.

Definition. If K is a closed convex set then the *support function* of K, $h_K(f)$, is defined by:

$$h_K(f) := \sup\{f(x) : x \in K\}.$$

The support function is a sublinear function. If K is bounded, then h_K is finite for all x and we may replace sup by max. If $0 \in K$, then h_K is non-negative, and if 0 is an interior point, then h_K is strictly positive. If K is symmetric, then h_K is a norm on the dual space. There are generalizations of this idea to infinite dimensional spaces, but there

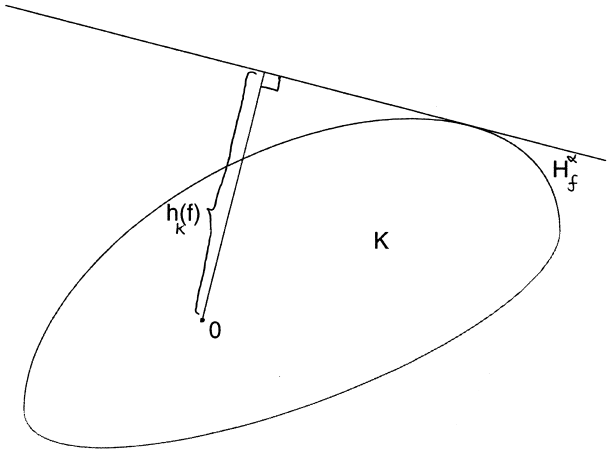


FIGURE 4

the dual space has a slightly different connotation—one must restrict attention to continuous linear functionals.

Most books on convexity that restrict attention to finite dimensional spaces identify the dual space $(\mathbb{R}^n)^*$ with $\mathbb{R}^n$ by means of the inner product. By that we mean that for each linear functional f there is a vector y_f such that $f(x) = \langle y_f, x\rangle$ where the symbol $\langle\,.\,,\,.\,\rangle$ denotes the inner product. In this case, the definition of support function is given in terms of the inner product.

To interpret the support function geometrically, we restrict our attention to linear functionals f with length 1. In this special case, the α that appears in the equation of the hyperplane:

$$f(x) = \alpha$$

represents the perpendicular distance from the origin to the hyperplane. Therefore, the support function (restricted to linear functions of length 1) represents the perpendicular distance from the origin to the supporting hyperplane of K that is in the direction f (see Fig. 4).

The following theorems are important.

Theorem. If f is a sublinear functional on $\mathbb{R}^n$, then there is a convex body K such that $f = h_K$.

Theorem. If K is a non-empty convex body, then $K = \{x : f(x) \le h_K(f)$ for all linear functionals $f\}$.

The proof of the second theorem requires the separation theorem from Section VII. There are a variety of proofs of the first theorem (see Schneider[4]).

V. NEW CONVEX SETS FROM OLD

A. The Convex Hull

We begin this section with a second look at the operation of the convex hull, the intersection of all convex sets containing A. Rather than cut down to the convex hull from outside the set, we may also build up the convex hull from inside.

If $x_1, x_2, \ldots, x_k$ is a finite set of vectors, then y is said to be a *convex combination* of these vectors if:

$$y = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k \text{ with } \lambda_i \ge 1 \text{ and } \sum \lambda_i = 1.$$

The definition of convexity uses only two points but, by induction, one shows that if K is convex then every convex combination of points in K is also in K. If A is not convex, every convex combination of points from A is in co$A$. Finally, one shows that the set of all convex combinations of points from A is a convex set. Therefore, co$A$ is precisely the set of all convex combinations of points from A.

Note that k, the length of the convex combination, is arbitrary but Carathéodory's Theorem in Section VII shows that in finite dimensional spaces this is really not so.

Theorem. The convex hull of a compact set is compact and of an open set is open.

Theorem. If A is a non-empty bounded set in $\mathbb{R}^n$, then the diameter of co$A$ is the same as the diameter of A. (Here, "diameter" means the supremum of the distances between points of A.)

A nice application of the ideas here is the Gauss-Lucas Theorem. A short and elegant proof can be found in Webster.[5]

Theorem. If $p(z)$ is a non-constant polynomial, then the roots of $p'(z)$ (the derivative of p) are contained in the convex hull of the roots of $p(z)$.

B. The Polar of a Convex Set

By definition, the convex hull operator assigns to each set a convex set. The next very important operation does the same thing.

Definition. If A is a non-empty subset of X, then the *polar* of A, denoted by A° (or A^*), is defined by:

$$A^\circ := \{f \in X^* : f(x) \le 1 \text{ for all } x \in A\}.$$

If $A \subseteq X$, then $A^\circ \subseteq X^*$. Often, the distinction between X and X^* is obscured and the inner product is used in the definition of A°. Similarly, for a non-empty set B in X^* we have:

$$B^\circ := \{x \in X : f(x) \le 1 \text{ for all } f \in B\}.$$

Repeated applications of this operation are indicated without parentheses, thus $A^{\circ\circ}$, $B^{\circ\circ}$, and so on. For all sets A (in either X or X^*) we have $A \subseteq A^{\circ\circ}$. The operation

reverses inclusions: If $A_1 \subseteq A_2$, then $A_2^\circ \subseteq A_1^\circ$. It follows that $A^\circ = A^{\circ\circ\circ}$ always.

The definition of B° reveals it to be an intersection of closed half-spaces so it is always a closed, convex set in X that contains 0. The same holds for A° in X^*. If $A = \{0\}$, then $A^\circ = X^*$ and $X^\circ = \{0\}$. If A is a single point other than 0, then A° is a half-space. Thus, the duality between points and half-spaces (encountered in Section IV) is implemented by this operation. However, if A is a half-space, then A° is not a singleton but a line segment joining the expected point to 0. Because A° is always a closed, convex set containing the origin, in order to get an exact duality we must restrict attention to this class of sets.

Theorem. If K is a closed, convex set with $0 \in K$, then $K = K^{\circ\circ}$.

This is a most important theorem whose proof relies on the separation theorem in Section VII. For an arbitrary set A, $A^{\circ\circ}$ is the closed convex hull of A and 0.

On the collection of closed convex sets that contain 0, the polar mapping is one to one and maps this collection onto the corresponding collection in X^*. If 0 is an interior point of K, then K° is compact. If K is compact, then K° has 0 as an interior point. Thus, the polar operation also maps the class of compact convex sets with 0 as an interior point onto the same class in X^*. If K is also symmetric about 0, then so is K°. In this last case, K plays the role of the unit ball in a normed space and K° is the dual ball in the dual space X^*.

Examples. If B is the unit ball in X, then B° is the unit ball in X^*.

If C_n is the standard cube in $\mathbb{R}^n$, then C_n° is the standard cross-polytope and conversely.

If S_n is the standard simplex, then S_n° is the following unbounded set. If $f = (\phi_1, \phi_2, \ldots, \phi_n)$, then $f \in S_n^\circ$ if and only if $\phi_i \leq 1$ for all i.

Finally, this duality also connects the gauge (and radial function) of K with the support function of K°.

Theorem. If K is a convex body with 0 as an interior point, then the gauge function of K° is the support function of K, and the gauge function of K is the support function of K°.

Since the radial function is the reciprocal of the gauge function, this is readily rewritten using the radial functions.

C. The Collection of Convex Sets as a Lattice

The intersection of two convex sets is again a (possibly empty) convex set. It is the largest convex set contained in both of them. The union of two convex sets need not be convex, but the convex hull of the union is the smallest convex set that contains both of them. Therefore, we define the following binary operations:

$$K_1 \wedge K_2 := K_1 \cap K_2$$

$$K_1 \vee K_2 := \mathrm{co}(K_1 \cup K_2).$$

With these operations, the collection of convex sets is a *lattice*. The underlying order relation is that of inclusion.

There are two important sublattices: the collection of closed convex sets that contain 0 and the collection of compact convex sets with 0 as an interior point. On each of these sublattices, the polar map is a bijection that reverses inclusion and hence reverses the lattice operations:

$$(K_1 \wedge K_2)^\circ = (K_1 \cap K_2)^\circ = K_1^\circ \vee K_2^\circ = \mathrm{co}\left(K_1^\circ \cup K_2^\circ\right)$$

$$(K_1 \vee K_2)^\circ = (\mathrm{co}(K_1 \cup K_2))^\circ = K_1^\circ \wedge K_2^\circ = K_1^\circ \cap K_2^\circ.$$

D. Algebraic Operations on Convex Sets

In addition to the operations just discussed, vector operations can be performed on the collection of convex sets. These are done "elementwise."

Definition. If K_1 and K_2 are convex sets in a vector space X and λ is a non-negative scalar, then:

$$K_1 + K_2 := \{x : x = x_1 + x_2 \text{ with } x_i \in K_i\}$$

and

$$\lambda K_1 := \{\lambda x : x \in K_1\}.$$

The reason for the restriction to $\lambda \geq 0$ is twofold. First, although we may define $(-1)K$, it may be (if K is symmetric) that $(-1)K = K$. Furthermore, the distributive law:

$$(\lambda + \mu)K = \lambda K + \mu K$$

only holds generally if $\lambda, \mu \geq 0$. Hence, there is no such thing as $K + (-1)K = \{0\}$. With the restriction to non-negative scalars, all the usual algebraic identities are valid. Despite this, we shall occasionally have need to talk about $-K := (-1)K$ below. Any non-negative linear combination of convex sets is again convex. This is the basis of what is called the *Brunn-Minkowski Theory* of convex sets (see Schneider[4]).

Examples. The sum of a convex set K with a single point $\{x_0\}$ is the same as the translation of K by x_0. In general, $K + L$ is the union of translates of K by points in L (or vice versa). Hence, one way to think of this operation is to visualize one of the convex sets K with its boundary ∂K and the other one L with the origin as a reference point somewhere in L. First translate L so that the reference point lies on ∂K and then slide L round K (just

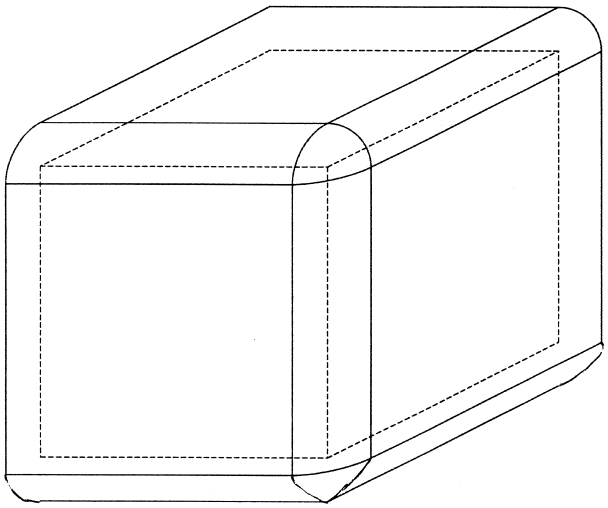


FIGURE 5

by translation) so that the reference point stays on ∂K and eventually has traversed all of it. The convex body swept out in this process is $K + L$.

In $\mathbb{R}^2$, the sum of the line segments $[-e_1, e_1]$ and $[-e_2, e_2]$ is the square C_2. If, in $\mathbb{R}^3$, we now add the line segment $[-e_3, e_3]$, we get the standard cube in $\mathbb{R}^3$. In general, the cube C_n is the sum of the line segments $[-e_i, e_i]$ for $i = 1, 2, \ldots, n$.

In $\mathbb{R}^2$, if we add the line segment $[(-1, -1),\ (1, 1)]$ to the square C_2 we get a hexagon with vertices at $(2, 0),\ (2, 2),\ (0, 2)$ and their negatives.

The sum of a finite number of line segments is a special type of polytope called a *zonotope*. These have a number of significant properties (see Section X). The sum of the cube C_3 and a multiple λB of the unit ball is shown in Fig. 5.

One of the reasons that the support function is so important is that it behaves well for these operations:

$$h_{K_1+K_2} = h_{K_1} + h_{K_2} \text{ and } h_{\lambda K} = \lambda h_K.$$

For the gauge and radial functions, we have:

$$g_{\lambda K} = \lambda^{-1} g_K \text{ and } r_{\lambda K} = \lambda r_K.$$

For the polar operation, likewise:

$$(\lambda K)^\circ = \lambda^{-1}(K^\circ).$$

Finally, these operations can be defined for arbitrary sets, and then the convex hull operation is also well behaved:

$$\text{co}(A + B) = \text{co}(A) + \text{co}(B) \text{ and } \text{co}\lambda A = \lambda \text{co}(A).$$

E. Operations that Increase Dimension

So far we have been concerned with operations that take place within one fixed vector space X or from that space to its dual X^*. We now consider two vector spaces X and Y and their Cartesian product $X \times Y$, which has dimension equal to the sum of the dimensions of X and Y (assuming all are finite dimensional). There is a natural embedding of X into $X \times Y$ that sends a vector x to $(x, 0)$ and similarly for Y. Then $X \times Y$ is the direct sum of these embedded spaces.

If K is a convex set in X and L is one in Y, then $K \times L$ (the Cartesian product) is a convex set in $X \times Y$. However, if we apply the embedding maps to K and L so that both can be thought of as lying in $X \times Y$, then $K \times L = K + L$.

In this sense, the cube C_n can be thought of either as a sum of line segments (as above) or as the Cartesian product of the interval $[-1, 1]$ with itself n times.

Examples. The product of the unit ball in $\mathbb{R}^2$ and the interval $[-1, 1]$ in $\mathbb{R}$ is the standard circular cylinder in $\mathbb{R}^3$. Likewise, the product of any convex set K with a line segment is a cylinder. In the case when K is a polytope, the word "prism" is often used instead of cylinder.

Another operation that can be performed on K and L as above is the suspension operation. As just indicated, we may think of both K and L as embedded in $X \times Y$ and then $K * L$, the suspension of K and L, is defined to be $\text{co}(K \cup L) = K \vee L$.

Example. The standard cross-polytope is built up from the interval $[-1, 1]$ by repeated suspension operations.

If we are careful to interpret the polar operations in the appropriate spaces, we then get:

$$(K + L)^\circ = K^\circ * L^\circ \text{ and } (K * L)^\circ = K^\circ + L^\circ.$$

F. Symmetrizing Operations

The question of symmetry is an important one. A convex set is said to be *symmetric* about the origin if $K = -K$. It is often useful to operate on a convex set in such a way as to make it more symmetrical. For example, to prove certain inequalities where some quantity is maximized by a ball, one may be able to show that the quantity increases under a symmetrizing operation.

The first of these is quite simple. We define the *difference set* of K to be the convex set $D(K) := K - K := K + (-1)K$. The set $D(K)$ is always symmetric about 0. One readily checks that the support function of $-K$ is given by $h_{-K}(f) = h_K(-f)$, hence $h_{D(K)}(f) = h_K(f) + h_K(-f)$. For linear functionals f of norm 1, this last quantity is defined to be the *width* of K in the direction f and is denoted by $w_K(f)$. It represents the distance between parallel supporting hyperplanes with normal direction f (see Fig. 6).

A set is said to be of constant width if $w_K(f)$ is constant for all f with $\|f\| = 1$; that is, if $h_{D(K)}$ (restricted to those f with $\|f\| = 1$) is constant, which is so if and only if $D(K) = B$. We shall say more about these sets in Section X. For now, note that the only convex set of constant width that is symmetric about 0 is the ball.

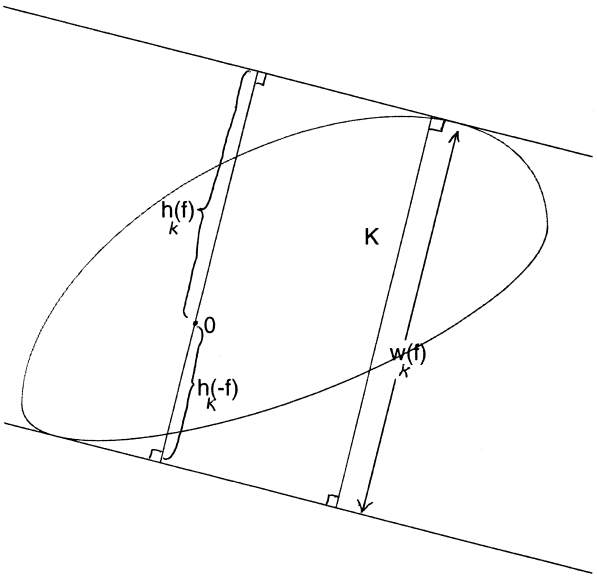


FIGURE 6

Examples. From some points of view, the most asymmetric convex set is the standard simplex. The difference body is rather regular. The difference body of the standard two-simplex S_2 is an affine regular hexagon and that of S_3 is an affine regular cuboctahedron (see Fig. 7).

The most well-known and simplest example of a non-symmetric set of constant width is the Reuleaux triangle (see Section X for a definition of this set; see Fig. 8 for a representation of it.)

The second symmetrizing operation is more complicated and requires an inner product and hence a notion of orthogonality. It is due to Jakob Steiner, is called *Steiner symmetrization*, and deals with symmetry with respect to a hyperplane.

Let K be a compact convex set in X. We wish to symmetrize K with respect to a hyperplane H in X. For each point $x \in H$ let $\ell(x)$ denote the line through x orthogonal to H. The intersection $\ell(x) \cap K$ is a line segment, a point, or is empty. If it is a line segment, let $S(x)$ be the translation of $\ell(x) \cap K$ whose midpoint is at x; if the intersection is a point, then $S(x) = x$, otherwise, $S(x)$ is empty. The Steiner symmetral of K with respect to H, $S_H(K)$, is defined by the equation $S_H(K) := \cup S(x)$.

It turns out that $S_H(K)$ is convex. If K has interior, then so does $S_H(K)$. If K and L are compact convex sets, then $S_H(K) + S_H(L) \subseteq S_H(K + L)$. The most important property of this symmetrization is that, given any convex body K, one may, by symmetrizing successively with a suitable sequence of hyperplanes, obtain a convex body that approximates a ball to within any degree of accuracy. Although we have not defined these measures, we also point out here that Steiner symmetrization leaves volume unchanged and does not increase either the surface area or the diameter of the set.

The third such operation is similar but reverses the roles of line and hyperplane. If ℓ is a line, for each point $x \in \ell$ let $H(x)$ be the hyperplane through x orthogonal to ℓ. If $H(x) \cap K \neq \emptyset$, let $B(x)$ be the $(n-1)$-dimensional ball centered at x and lying in $H(x)$ whose $(n-1)$-dimensional volume (see Section VIII) is the same as that of $H(x) \cap K$. Let $S_\ell(K) := \cup B(x)$. This object is also convex. It is called the *Schwarz symmetral* of K.

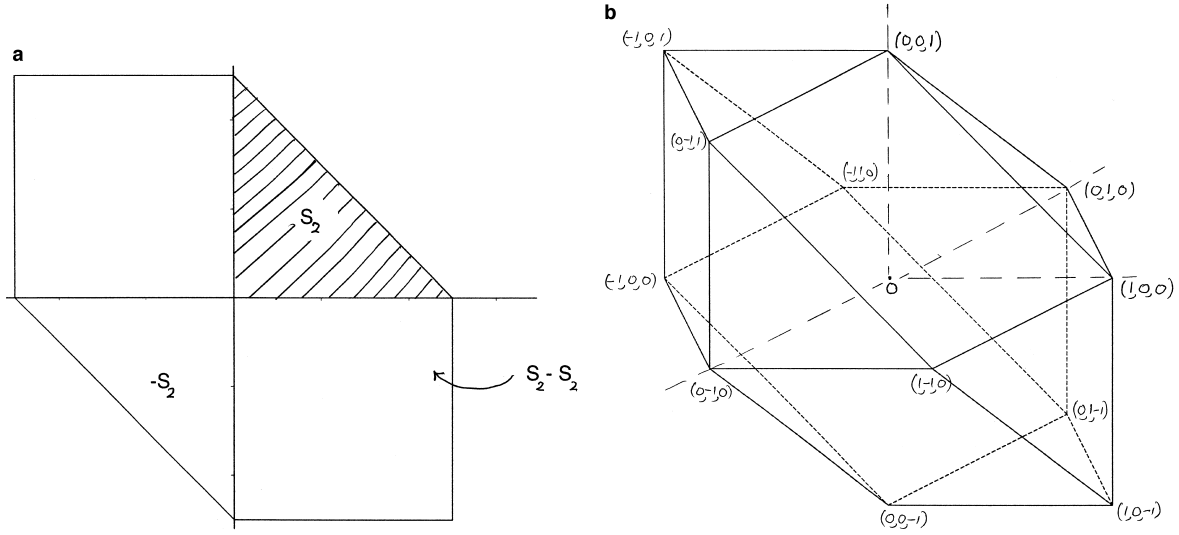


FIGURE 7

G. Other Operations

Two more ways to get new convex sets from old ones ought to be mentioned. The first is the projection body operator Π. There are several ways to define this operation; the simplest way is to use orthogonal projections. If K is a convex body in X, for each unit vector u in X consider the projection of K onto the hyperplane H_u orthogonal to u. This projection is the set $p_u(K)$ consisting of those points $x \in H_u$ such that the line $x + \lambda u$ has a non-empty intersection with K. Now let $h(u)$ be the $(n-1)$-dimensional volume (see Section VIII) of $p_u(K)$. It turns out that this function is subadditive. If it is extended to all of X by (positive) homogeneity, then it is sublinear and hence is the support function of a convex body in X^* called the *projection body* of K, $\Pi(K)$.

This is a most interesting object. It is always symmetric with respect to the origin. If K is a polytope, then $\Pi(K)$ is not just a polytope but is a zonotope (a sum of line segments) which has all of its faces centrally symmetric.

Examples. The construction (explained in Section X) of projection bodies of polytopes is relatively easy. The projection body of C_3 is (up to a scaling) C_3 itself. The projection body of $\mathcal{O}_3$ is a rhombic dodecahedron. This object has eight vertices that coincide with those of C_3 and six that are at $\pm 2e_1, \pm 2e_2, \pm 2e_3$. It has 12 facets that are all alike and are rhombi, hence the name.

The second operation is (in some not entirely clear way) a dual construction to Π. We begin with a convex body K such that $0 \in K$. Let f be a norm 1 linear functional in X^*. Let H_f be the hyperplane that is the kernel of f. Let $i_f(K) := K \cap H_f$ and let $r(f)$ be the $(n-1)$-dimensional volume (see Section VIII) of $i_f(K)$. Now let $I(K)$ be the set in X^* consisting of all the line segments $[0, r(f)f]$. In other words, by again extending r, we make it the radial function of $I(K)$. It is an important theorem (whose proof is difficult) of Busemann that if K is symmetric about 0 then $I(K)$ is convex. It is called the *intersection body* of K. For more general convex sets, $I(K)$ is star shaped. Lutwak has shown that the "proper" setting for this operation is the class of star-shaped sets.

VI. SPACES OF CONVEX SETS

In the last section, various operations on the collection of all convex sets were considered. With the algebraic operations of addition and multiplication by (non-negative) scalars, the collection of convex sets has many of the attributes of a vector space. In this section, we show that the collection of compact convex sets is also a metric space. In fact, there are two metrics that can be considered. One is defined on all compact convex sets in a given vector space regardless of their dimension. The second distinguishes between points, line-segments, two-dimensional convex sets, and so on.

A. The Hausdorff Metric

If K and L are two compact convex sets, then they are bounded. Hence, there exist non-negative scalars λ, μ such that:

$$K \subseteq L + \lambda B \text{ and } L \subseteq K + \mu B.$$

Now let λ_0 be the infimum (in fact, the minimum) of all such λ's and similarly for μ_0. The numbers λ_0 and μ_0 may also be defined directly in terms of the norm (or distance) on X by:

$$\lambda_0 = \min_{x \in L} \max_{y \in K} \|x - y\|$$

and similarly for μ_0.

Definition. The *Hausdorff metric* δ on the set of compact convex sets is defined by the equation:

$$\delta(K, L) := \max\{\lambda_0, \mu_0\}.$$

Theorem. The function δ is a metric on the set of all compact convex sets in X.

If the support functions of convex sets are restricted to the unit ball in X^*, then the support function of a multiple λB of the unit ball is just the constant λ; hence, $K \subseteq (L + \lambda B)$ if and only if $h_K \leq h_L + \lambda$, which implies that the Hausdorff metric between the sets is the same as the uniform metric between the support functions. All the operations discussed in the previous section are continuous with respect to this metric as are several important functions that are defined in later sections.

The most important fact about this metric is a compactness result, known as the *Blaschke selection theorem* (the name refers to the selection of a convergent subsequence from a given bounded sequence).

Theorem (Blaschke). The set of compact convex sets contained in some ball $B[x, r]$ and equipped with the Hausdorff metric is compact.

Various collections of convex sets, for example the collection of all ellipsoids, form closed subsets in this metric. Therefore, a bounded sequence of ellipsoids has a subsequence that converges to another ellipsoid (provided we allow degenerate cases such as points and line segments).

Other important classes are dense in this metric. For example, the collection of all polytopes is dense. Therefore, any convex body can be approximated as closely as we please (with respect to the Hausdorff metric) by a polytope. This, together with the continuity of various functions, means that the proof of a theorem can often be

accomplished by first proving the result for polytopes and then extending it to all convex bodies "by continuity."

A convex set is said to be strictly convex if its boundary does not contain any line segment (of positive length). A convex set is said to be smooth if there is a unique supporting hyperplane at each point of its boundary. The sets of smooth, of strictly convex, and of both smooth and strictly convex bodies are all dense in the set of all convex bodies.

B. The Banach-Mazur Metric

It is sometimes appropriate to consider that convex sets of different dimension are infinitely far apart. This section is concerned with metrics of this sort. We limit our attention to convex bodies with 0 as an interior point.

If K and L are two convex bodies with 0 in their interiors, then there are scalars λ and μ such that:

$$K \subseteq \lambda L \text{ and } L \subseteq \mu K.$$

As before, we can now take λ_0 and μ_0 to be the minimal such λ and μ. Then set $\Phi(K, L) := \lambda_0\mu_0$. The fundamental result is now John's Theorem.

Theorem (John). If K is a centrally symmetric convex body in an n-dimensional space X and if 0 is an interior point of K, then there is an ellipsoid E such that:

$$E \subseteq K \subseteq \sqrt{n}E$$

The standard cube C_n and cross polytope $\mathcal{O}_n$ show that $\sqrt{n}$ cannot be improved. If we remove the condition that K be centrally symmetric, then we must replace $\sqrt{n}$ by n. This bound is attained by the simplex S_n.

The ellipsoid E that appears in this result is of considerable interest. It is the ellipsoid of maximal volume contained in K and is called the *Löwner-John ellipsoid*. It occurs in linear and nonlinear programming in Khachiyan's polynomial time algorithm and in Schor's algorithm.

The functional Φ is not a metric for two reasons. First, the construction is multiplicative rather than additive; e.g., $\Phi(K, K) = 1$ rather than 0. If we want a genuine metric we must take the logarithm of Φ. Since not all authors do so, one should be careful when reading the literature to see the precise definition.

Second, $\log(\Phi(K, \alpha K)) = 0$; Therefore, one should consider equivalence classes of multiples of K. However, it is more appropriate to enlarge the equivalence classes and say that the sets K_1 and K_2 are equivalent if there is an invertible linear map T such that $T(K_1) = K_2$. Finally, the definition of the Banach-Mazur metric is $\Delta(K, L) := \inf\{\log[\Phi(K, T(L))] : T \text{ is an invertible linear map}\}$.

An infimum is used here because it is also a useful definition in the case of infinite dimensional spaces. Restricted to the finite dimensional case, the infimum is attained. The functional Δ is a metric on the equivalence classes of convex sets under the above equivalence relation. If we consider the norms generated by K and L, then the equivalence relation is one of isometry between normed spaces and Δ measures the distance between equivalence classes of normed spaces.

John's Theorem now says that the distance between the equivalence class containing K and the set of ellipsoids (which is the equivalence class containing B) is no more than $(\log n)/2$. Hence, the distance between any two equivalence classes is no more than $\log n$. If we allow non-symmetric sets, then these numbers are doubled. It is surprising that the exact diameter of these metric spaces is only known in the case of two dimensions.

VII. BASIC THEOREMS

A. Separation and Support Theorems

The notion of separation involves placing a hyperplane between two convex sets. There are varying degrees of separation that can be considered.

Definition. A hyperplane $H = H_f^\alpha$ is said to *separate* the convex sets K and L if K lies in one of the closed half-spaces determined by H, and L lies in the other. The separation is *proper* if it is not the case that both sets lie in H. The separation is *strict* if one can replace "closed" by "open" in the first sentence. Finally, the separation is *strong* if there exist α and β with $\alpha < \beta$ and $K \subseteq H_f^{\alpha-}$ and $L \subseteq H_f^{\beta+}$ (or vice versa).

Separation Theorem. Let K be a convex set in $\mathbb{R}^n$ and suppose x is not in K, then K and x can be separated. If K is closed, then K and x can be strongly separated.

It follows from the second statement that every closed convex set can be represented as an intersection of closed half-spaces. It follows from the first statement that if K has a non-empty interior then at each point x of the boundary of K there is a supporting hyperplane obtained by separating x from the interior of K.

There is a converse to this theorem. If A is a closed set with a non-empty interior and if, at each boundary point, there is a supporting hyperplane, then A is convex.

The separation theorem can be generalized to the following: If K and L are two convex sets whose relative interiors are disjoint, then K and L can be separated. If one set is closed and the other is compact, then the separation is strong.

The proof of the seemingly more general statement follows from the earlier one because K and L can be (strongly) separated if and only if $K - L$ and $\{0\}$ can be (strongly) separated. Also, if one set is compact and the other closed, then $K - L$ is closed. This is *not* true in general for two closed sets.

Example. In $\mathbb{R}^2$ the sets $K := \{(x, y) : y = 0\}$ and $L := \{(x, y) : x \geq 0 \text{ and } y \geq 1/x\}$ are both closed and convex. They cannot be strongly (or even strictly) separated, and the set $K - L$ is open.

These theorems have very important analogs in infinite dimensional spaces. In that setting, they are all consequences of the Hahn-Banach Theorem, which is one of the most important theorems of functional analysis.

B. Carathéodory's Theorem and Its Relatives

The following theorems are all closely related, but the Carathéodory result appears the most fundamental.

Theorem (Carathéodory). If A is a subset of an n-dimensional space and if $x \in \text{co}A$, then x can be expressed as a convex combination of $(n + 1)$ or fewer points.

Other ways of phrasing the conclusion is to say that x is a convex combination of a set of points in general position. Another is to say that x lies in a simplex whose vertices are in A. Thus, when constructing the convex hull, the length of the convex combinations needed is bounded (in finite dimensional spaces).

Theorem (Radon). If a finite set F of points in an n-dimensional space is *not* in general position, then it may be decomposed into two disjoint subsets F_1 and F_2 such that $\text{co}(F) \cap \text{co}(G) \neq \emptyset$.

In particular, this is true of any set of at least $(n + 2)$ points.

Theorem (Helly). Let $K_1, K_2, \ldots, K_m$ be a finite family of convex sets in an n-dimensional space ($m \geq n + 1$). If every subfamily with exactly $(n + 1)$ members has a non-empty intersection, then the whole family has a non-empty intersection.

Eggleston[1] shows how Helly's Theorem can be derived from Carathéodory's and conversely.

Since any family of compact sets has a non-empty intersection if every finite subfamily does, there is an easy extension to infinite families of compact convex sets. If an arbitrary family of compact convex sets in an n-dimensional space is such that every subfamily with $(n + 1)$ members has a non-empty intersection, then so does the whole family. A transversal for a family of sets is a line that meets every member of the family.

Corollary. If $\mathcal{F}$ is a finite family of parallel line segments in $\mathbb{R}^2$ such that every three of them has a transversal, then the whole family has a transversal.

Corollary. Let $\mathcal{F}$ be a finite family of convex sets in $\mathbb{R}^n$ and let K be a convex set. If for every finite subfamily with $(n + 1)$ elements there is a translate of K that intersects each member of the subfamily, then there is a single translate of K which intersects every member of the whole family.

Theorem (Kirchberger). Let F_1 and F_2 be finite sets in an n-dimensional space such that $F_1 \cup F_2$ has at least $(n + 2)$ elements. Suppose that for every subset F of $F_1 \cup F_2$ with exactly $(n + 2)$ points the sets $F \cap F_1$ and $F \cap F_2$ can be strictly separated, then F_1 and F_2 can be strictly separated.

Webster[5] gives an elegant proof of Jung's theorem also based on Helly's theorem.

Theorem (Jung). Every set A in $\mathbb{R}^n$ with diameter 1 is contained in a closed ball of radius no more than $\sqrt{n/(2n + 2)}$.

Finally, we give Krasnosel'skiĭ's Theorem (sometimes called the "Art Gallery" Theorem).

Theorem (Krasnosel'skiĭ). Let A be a compact subset of $\mathbb{R}^n$. If, for every $(n + 1)$ points, $a_1, a_2, \ldots, a_{n+1}$ of A there is a point x of A such that the line segments $[x, a_i]$ all lie in A, then A is star shaped.

In other words, in an art gallery, if for every finite set of $(n + 1)$ pictures there is a point in the gallery from which one can see all $(n + 1)$, then there is a point from which one can see the whole art gallery.

All of these theorems have been much generalized. For one collection of such results, see the article by J. Eckhoff in the *Handbook of Convex Geometry*.[2]

VIII. VOLUMES AND MIXED VOLUMES

For many of the more interesting properties of convex sets it is necessary to measure them in some way. Such measurement requires more advanced ideas than were presented in Section I. The basic concept is that of volume in an n-dimensional space. There are several approaches which coincide for compact convex sets but may not for more general types of sets.

The most straightforward approach is Eggleston's,[1] which says that the volume of a convex set in $\mathbb{R}^n$ is its n-dimensional Lebesgue measure. The volume of a set in $\mathbb{R}^n$ will be denoted by $V_n(K)$. One-dimensional volume is usually called *length*, and V_2 is usually called *area*. In

n-dimensional space, V_{n-1} is also frequently referred to as area. No confusion should arise.

The volume functional takes values in the extended real numbers and has a number of very important properties:

1. It is non-negative—$V_n(K) \geq 0$ for all convex sets K in $\mathbb{R}^n$—and is strictly positive if K has a non-empty interior (has dimension n).
2. It is countably additive, in the sense that, if $\{K_m: m = 1, 2, \ldots, \}$ is a sequence of disjoint convex sets, then $V_n(\cup_m K_m) = \sum_m V_n(K_m)$.
3. It is finite for compact sets; $V_n(K) < \infty$ if $K \subseteq \mathbb{R}^n$ is compact.
4. It is continuous with respect to the Hausdorff metric.
5. It is monotonic; if $K \subseteq L$, then $V_n(K) \leq V_n(L)$ (this follows from (property 2)).
6. It is translation invariant; $V_n(K + x) = V_n(K)$.
7. For linear transformations T, it has the property that $V_n(T(K)) = \det T\ V_n(K)$.
   In particular:
8. $V_n(\lambda K) = \lambda^n V_n(K)$ and also V_n is invariant under rigid motions.

The important fact from the theory of Haar measure is that, up to a scalar factor, there is only one functional that has properties 1, 2, 3, and 6.

However, if we relax property 2 a little, then a number of important functions are relevant to the theory of convex sets.

A real-valued function v defined on a collection of sets $\mathcal{S}$ is said to be a valuation if:

$$v(K \cup L) + v(K \cap L) = v(K) + v(L)$$

whenever $K,\ L,\ K \cup L$, and $K \cap L$ are all in $\mathcal{S}$.

It follows that for finitely many disjoint sets, a valuation is additive, hence, if its values are non-negative, it is monotonic. Volume is a valuation.

The n-dimensional volume of a convex set may be calculated by integrating the $(n-1)$-dimensional volumes of its cross sections in some particular direction. Thus, if f is a linear functional and if K is a convex set and if we let $K_\alpha := K \cap H_f^\alpha$, then:

$$V_n(K) = \int_{-\infty}^{\infty} V_{n-1}(K_\alpha) d\alpha.$$

Examples. The above formula can be used inductively to prove that:

$$V_n(B) = \frac{\pi^{n/2}}{\Gamma(1 + n/2)}.$$

This number is abbreviated ϵ_n (many authors use κ_n).

There is a more general formula for the n-dimensional p-ball $B(p)$:

$$V_n(B(p)) = \frac{2^n \Gamma(1 + 1/p)^n}{\Gamma(1 + n/p)}.$$

The volume of the standard cube C_n is $V_n(C_n) = 2^n$.

The volume of the standard simplex S_n is $V_n(S_n) = 1/n!$.

The volume of the standard cross-polytope $\mathcal{O}_n$ is $V_n(\mathcal{O}_n) = 2^n/n!$.

If K is a convex set such that $K \subseteq H_f^\alpha$ with $\alpha \neq 0$ and $\|f\| = 1$, then the pyramid with vertex at 0 and base K is the convex hull of K and 0. The volume of this pyramid is $\alpha V_{n-1}(K)/n$.

The previous formula can be generalized. If P is a polytope with 0 as an interior point and if $F_i,\ i = 1, 2, \ldots, m$, denote the facets of P and if α_i is the perpendicular distance from 0 to the hyperplane containing F_i, then $V_n(P) = (1/n) \sum_1^m \alpha_i V_{n-1}(F_i)$.

With a notion of surface area, this can be further generalized to arbitrary convex sets:

$$V_n(K) = \frac{1}{n} \int_{S^{n-1}} h_K(f) d\sigma_K(f)$$

where σ_K is the "surface area measure" induced on the surface of the dual unit ball S^{n-1} by K (see Schneider[4] for details).

Properties (7) and (8) show how V_n behaves for scalar multiples and under linear maps. The basic question, first considered by Brunn and Minkowski toward the end of the 19th century, is how does V_n behave with respect to the addition of sets; i.e., how is property 6 extended from singletons to general convex sets?

To see how this might work, consider the special case of $K + \lambda K$; then:

$$V_n(K + \lambda K) = V_n((1 + \lambda)K) = (1 + \lambda)^n V_n(K)$$
$$= \sum \binom{n}{i} \lambda^i V_n(K);$$

that is, we get a polynomial in λ whose coefficients are multiples of $V_n(K)$. In fact, we always get a polynomial.

Theorem. If K and L are convex bodies, then $V_n(K + \lambda L)$ is a polynomial in λ of degree n whose coefficients are written in the following way:

$$V_n(K + \lambda L) = \sum_0^n \binom{n}{i} V(K, n - i; L, i) \lambda^i.$$

The numbers $V(K, n - i; L, i)$ are called the *mixed volumes* of K and L. The numbers $n - i$ and i are inserted in this notation because the mixed volumes are functions of n variables and here K and L occur $n - i$ times and i times, respectively. Since this is true for a summand with two terms, it can be extended inductively to sums with an arbitrary number of terms. The essential feature

is that the volume of a linear combination $(\sum \lambda_i K_i)$ is a homogeneous polynomial in the λ_i's of degree n whose coefficients are functions of precisely n of the K_i's.

Example. Referring again to Figure 5, one sees that

$$V_3(C_3 + \lambda B) = V_3(C_3) + A(C_3)\lambda + (3\pi)L(C_3)\lambda^2 + V_3(B)\lambda^3$$

where $A(C_3)$ is the sum of the areas of the facets of the cube and $L(C_3)$ is the length of an edge of the cube and measures the width of the cube. Thus, in this instance, we have $V(C_3, C_3, C_3) = V_3(C_3)$, $V(C_3, C_3, B) = A(C_3)/3$, $V(C_3, B, B) = \pi L(C_3)$, and $V(B, B, B) = V_3(B)$.

Mixed volumes have the following properties:

1. They are non-negative.
2. They are monotonic in each variable.
3. They are homogeneous (for non-negative scalars) in each variable. Recall that with the notation $V(K, n-i; L, i)$ the variable K occurs $n-i$ times.
4. They are additive in each variable.
5. They are translation invariant (in each variable).
6. They are continuous with respect to the Hausdorff metric.
7. $V(K, K, \ldots, K) = V_n(K)$.

The example shows that mixed volumes of the form $V(K, n-i; B, i)$ are closely related to the geometry of K. For this, the notation is modified.

Definition. The *quermassintegrals* or *cross-sectional measures* or *Minkowski functionals* of K are written $W_i(K)$ and are defined by $W_i(K) := V(K, i; B, n-i)$ (note the change of place of i).

Then, $W_0(K) = V_n(K)$, $W_1(K) = A(K)/n$, and $W_{n-1}(K) = \epsilon_n b(K)/2$, where $A(K)$ is the surface area of K and $b(K)$ is the mean width of K. These last two equations may either be taken as the definition of these quantities or, if they are defined in other ways, then as theorems. Perhaps surprisingly, even $W_n(K)$ has relevance to the set K: $W_n(K) = \epsilon_n \chi(K)$, where $\chi(K)$ is the Euler characteristic of K which (for convex sets) is always 1.

Since W_1 is the coefficient of λ in a certain polynomial, it (and hence $A(K)$) can be obtained via differentiation in the usual way.

Theorem. The surface area $A(K)$ of K is obtained by the formula:

$$A(K) = \lim_{\lambda \to 0} \frac{V_n(K + \lambda B) - V_n(K)}{\lambda}.$$

There are also integral formulas for both $A(K)$ and $b(K)$:

Theorem (Cauchy). If K is a convex body in $\mathbb{R}^n$, then:

$$A(K) = \frac{1}{n\epsilon_n} \int_{S^{n-1}} V_{n-1}(p_u(K)) d\omega(u)$$

where S^{n-1} is the surface of the unit ball, $d\omega$ is surface area measure (Lebesgue measure) on S^{n-1}, u is a unit vector, and $p_u(K)$ is the projection of K onto the subspace orthogonal to u.

Theorem. If K is a convex body in $\mathbb{R}^n$, then:

$$b(K) = \frac{2}{n\epsilon_n} \int_{S^{n-1}} h_K(u) d\omega(u).$$

(This is often taken as the definition of $b(K)$).

There are various relationships between the functionals W_i expressed as integral formulas.

All the functionals W_i are valuations and, in a certain precise sense, these are *all* the valuations.

Theorem (Hadwiger). If v is a valuation on the collection of convex bodies that is invariant under rigid motions and is either continuous or monotonic, then:

$$v(K) = \sum_0^n \alpha_i W_i(K)$$

where the coefficients α_i are real if v is continuous and non-negative if v is monotonic. The first straightforward proof of this theorem was given in 1995 by Dan Klain.

IX. INEQUALITIES

There is a huge variety of inequalities that relate to convex sets in one way or another. We present a brief sample. The reader is referred to the work of Erwin Lutwak and, in particular, his article in the *Handbook of Convex Geometry*.[2] We do not always give the most general result (which may need more notation or definitions); many extensions can be found in Schneider.[4] The word "homothetic" often occurs in the conditions for equality. The sets A and B are homothetic if $A = \lambda B + a$ $(\lambda > 0)$; i.e., A is the image of B under a translation and a (positive) dilation.

We begin with the fact that the nth root of the volume is a concave function.

Theorem (Brunn-Minkowski). If K and L are convex bodies with interior in $\mathbb{R}^n$ and if $0 \le \alpha \le 1$, then:

$$V_n^{1/n}(\alpha K + (1-\alpha)L) \ge \alpha V_n^{1/n}(K) + (1-\alpha)V_n^{1/n}(L)$$

with equality if and only if K and L are homothetic.

Theorem (Minkowski inequality for mixed volumes). If K and L are convex bodies in $\mathbb{R}^n$, then:

$$V(K, n-1; L, 1)^n \geq V_n(K)^{n-1} V_n(L)$$

with equality if and only if K and L are homothetic.

The power of this inequality is shown by the fact that, substituting the unit ball for L, one immediately gets the isoperimetric theorem for convex sets.

Theorem (Isoperimetric). If K is a convex body in $\mathbb{R}^n$ with prescribed volume v, then $A(K) \geq A(B_v)$, where B_v is the dilation of B with volume v. Moreover, equality holds if and only if K is a translate of B_v.

Corollary (Isoperimetric inequality). If K is a convex body in $\mathbb{R}^n$, then:

$$\frac{A(K)^n}{V_n(K)^{n-1}} \geq \frac{A(B)^n}{V_n(B)^{n-1}} = n^n \epsilon_n.$$

Theorem (Urysohn). If K is a convex body in $\mathbb{R}^n$, then:

$$\frac{b(K)^n}{V_n(K)} \geq \frac{b(B)^n}{V_n(B)} = 2^n/\epsilon_n.$$

Since $b(K) \leq D(K)$ (the diameter of K), we get the following corollary.

Corollary (Isodiametric inequality). If K is a convex body in $\mathbb{R}^n$, then:

$$\frac{D(K)^n}{V_n(K)} \geq \frac{D(B)^n}{V_n(B)} = 2^n/\epsilon_n.$$

In the last three inequalities, equality holds if and only if K is a ball.

The next set of inequalities relate the volume functional to the operations given in Section V. However, they are also affine inequalities because the quantities involved are unchanged by affine transformations.

Theorem (Blaschke-Santaló). If K is a convex body in $\mathbb{R}^n$, then:

$$V_n(K) V_n(K^\circ) \leq V_n(B) V_n(B^\circ) = \epsilon_n^2$$

with equality if and only if K is an ellipsoid.

The quantity $V_n(K)V_n(K^\circ)$ is called the *volume product* of K and is an affine invariant. There is a conjecture of Mahler on the lower bound for the volume product. It has been proved for zonoids (the closure, in the Hausdorff metric, of the set of zonotopes).

Theorem (Reisner). If K is a zonoid in $\mathbb{R}^n$, then:

$$V_n(K)V_n(K^\circ) \geq V_n(C_n)V_n(C_n^\circ) = V_n(\mathcal{O}_n)V_n(\mathcal{O}_n^\circ) = 4^n/n!$$

with equality if and only if $K = C_n$ ($\mathcal{O}_n$ is not a zonoid).

Theorem (Rogers-Shephard). If K is a convex body in $\mathbb{R}^n$ then

$$2^n V_n(K) \leq V_n(D(K)) \leq \binom{2n}{n} V_n(K)$$

with equality on the left if and only if K is symmetric and on the right if and only if K is a simplex. (The left-hand inequality is trivial; it is the other one that is due to Rogers and Shephard.)

Theorem (Busemann intersection inequality). If K is a convex body in $\mathbb{R}^n$ with 0 as an interior point, then:

$$\frac{V_n(I(K))}{V_n(K)^{n-1}} \leq \frac{V_n(I(B))}{V_n(B)^{n-1}} = \frac{\epsilon_{n-1}^n}{\epsilon_n^{n-2}}$$

with equality if and only if K is an ellipsoid.

Theorem (Petty projection inequality). If K is a convex body in $\mathbb{R}^n$, then:

$$V_n(K)^{n-1} V_n([\Pi(K)^\circ]) \leq V_n(B)^{n-1} V_n([\Pi(B)^\circ]) = (\epsilon_n/\epsilon_{n-1})^n$$

with equality if and only if K is an ellipsoid.

Here, we insert results dealing with two problems that generated a great deal of research in the latter part of the 20th century: the Busemann-Petty problem and its "dual," the Shephard problem. Both deal with convex bodies symmetric about 0 because otherwise it is easy to give a negative answer to both questions.

Question (Busemann-Petty). If K and L are two centrally symmetric convex bodies in $\mathbb{R}^n$ and if, for every linear functional f, we have:

$$V_{n-1}\big(K \cap H_f^0\big) \leq V_{n-1}\big(L \cap H_f^0\big)$$

does it follow that $V_n(K) \leq V_n(L)$?

The breakthrough on this problem came when Lutwak showed that the answer is "yes" if the body K is an intersection body (in a broader sense than our definition), otherwise, there will be a convex body L yielding a counter example. The work of many authors and finally of Zhang, Gardner, and Koldobsky showed that the answer to the problem is "no" in all dimensions ≥ 5 but that for $n = 2, 3, 4$ the answer is "yes."

Question (Shephard). If K and L are two centrally symmetric convex bodies in $\mathbb{R}^n$ and if, for every direction u, we have:

$$V_{n-1}(p_u(K)) \leq V_{n-1}(p_u(L))$$

does it follow that $V_n(K) \leq V_n(L)$?

Petty and Schneider showed that the answer is "no" in general but "yes" if the body L is a projection body. Since all symmetric convex bodies in $\mathbb{R}^2$ are projection bodies, the answer is "yes" in $\mathbb{R}^2$ but "no" in all higher dimensions.

Finally we mention two well-known inequalities that relate to the finite (and infinite) dimensional ℓ_p-spaces.

Theorem (Hölder's inequality). If the numbers $p \geq 1$ and q are related by the equation $p^{-1} + q^{-1} = 1$ and if $x = (\xi_i)$ and $y = (\eta_i)$ are two vectors in $\mathbb{R}^n$, then:

$$\left| \sum \xi_i \eta_i \right| \leq \left(\sum |\xi_i|^p \right)^{1/p} \left(\sum |\eta_i|^q \right)^{1/q}.$$

A consequence of this inequality is that of Minkowski.

Theorem (Minkowski's inequality). If x and y are vectors in $\mathbb{R}^n$ and if $\|x\|_p := (\sum |\xi_i|^p)^{1/p}$, then

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p.$$

Therefore, the functional $\|x\|_p$ is a norm and the set $B(p)$ defined in Example 9 in Section III is convex.

X. SPECIAL CLASSES OF CONVEX SETS

A. Polytopes

In this section we deal with bounded polytopes. Recall that such a polytope is a convex body that may be regarded either as the convex hull of finitely many points or, dually, as the intersection of finitely many half-spaces. Polygons and polyhedra have been studied since the beginnings of mathematics. The existence theorem of Minkowski is very important.

Theorem (Minkowski). If $\{u_1, u_2, \ldots u_k\}$ is a set of dual unit vectors which do not all lie in a hyperplane and if $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ are positive real numbers such that:

$$\sum_i^k \alpha_i u_i = 0$$

then there is a polytope (unique up to translation) that has k facets whose areas are given by the α_i's and whose "normals" are given by the u_i's.

There is a generalization of this theorem to general convex bodies determined by general "surface area measures" (see Schneider[4]).

By the facial structure of a polytope, we mean the collection of all of its faces classified by their dimension. If P is an n-dimensional polytope, then there is precisely one n-dimensional face, P itself. The empty set is usually included as the unique face of dimension -1. The 0-dimensional faces are the vertices, the one-dimensional faces are the edges, and the $(n-1)$-dimensional faces are the facets. The f-vector of P is the vector $(f_0, f_1, f_2, \ldots, f_{n-1})$ where f_i is the number of faces of dimension i. The faces of P form a lattice under the inclusion relation (this is why P and $\emptyset$ are included as faces). The lattice of faces of P° forms the dual lattice.

A key combinatorial problem is to characterize those vectors that are f-vectors of some polytope. Only in three dimensions has this problem been solved. The primary necessary condition for a vector to be an f-vector is that it satisfies the so-called Euler relation.

Theorem. If f_i denotes the number of faces of a polytope of dimension i, then

$$\sum_0^n (-1)^i f_i = 1.$$

The number 1 is the Euler characteristic of P.

Corollary. In three-dimensional space, the number of vertices f_0 of edges f_1 and of facets f_2 satisfy the equation $f_2 - f_1 + f_0 = 2$.

Theorem (Steinitz). In three-dimensional space, (f_0, f_1, f_2) is the f-vector of a polyhedron if and only if it satisfies, in addition to the Euler relation, the following inequalities:

1. $4 \leq f_0 \leq 2f_2 - 4$
2. $4 \leq f_2 \leq 2f_0 - 4$

Such conditions are not completely known in higher dimensions but there are many partial results, especially for $n = 4$ (see the survey article by Bayer and Lee[2]). However, it is known that the Euler relation is the only affine relation satisfied by all f vectors.

Of particular appeal are the regular figures. If $n = 2$, then there is an infinite family of regular polygons. Up to rigid motions and dilations, there is precisely one for each number k of vertices (edges). If $n = 3$, then there are precisely five regular polyhedra (the Platonic solids). There are many proofs that there can be no more. We outline one. Suppose that the facets are p-gons and that q meet at each vertex. Then the Euler relation implies that $p^{-1} + q^{-1} > 1/2$. Since $p, q \geq 3$, only the values $(p, q) = (3, 3)$, $(3, 4)$, $(4, 3)$, $(3, 5)$ and $(5, 3)$ are possible.

In all dimensions it is possible to construct a regular cube C_n, a regular cross-polytope $\mathcal{O}_n$, and a regular simplex (a linear image of S_n). For $n \geq 5$, these are the only regular polytopes. When $n = 4$, there are three more with, respectively, 24 octahedral facets, 120 dodecahedral facets, and 600 tetrahedral facets.

A great variety of polytopes have some degree of regularity; for example, the cuboctahedron has six square facets and eight that are equilateral triangles and whose vertices are all alike. Its dual is the rhombic dodecahedron, which has all its facets alike (they are all rhombi).

B. Zonotopes

This special class of polytopes was first brought to attention by Federov in connection with crystallography. They are centrally symmetric but, more than that, they are characterized by having all their two-dimensional faces centrally symmetric. A theorem of Alexandrov implies that *all* of the faces are centrally symmetric. Thus, when $n = 3$, the zonohedra are those polyhedra with centrally symmetric facets. For $n > 3$, there are polytopes whose facets are centrally symmetric but are not zonotopes.

From the point of view of the Brunn-Minkowski theory based on sums and scalar multiples of convex sets, zonotopes are the simplest objects because they are the ones that can be expressed as sums of line segments. To simplify matters, the line segments can be centered at 0 so that the zonotope is also symmetric about 0. The support function of a line segment $[-x, x]$ is $h_{[-x,x]}(f) = |f(x)|$ and, since support functions respect addition, if $Z := \sum_i [-x_i, x_i]$, then $h_Z(f) = \sum_i |f(x_i)|$.

In addition to crystallography, these objects are important in several areas of mathematics. First, if we restrict the domain of the projection body operator Π to the polytopes, its range is the set of zonotopes. Moreover, if the domain is restricted to the centrally symmetric polytopes, then Π is one–one.

If P is a polytope, then the action of Π on P is described in the following way. Let the facets of P be denoted by $F_1, F_2, \ldots, F_m$. Each F_i is contained in a hyperplane $H_{f_i}^{\alpha_i}$ with $\|f_i\| = 1$. The functional f_i is called the unit *normal* to the facet F_i; the sign is usually chosen so that it is directed outwards but that is immaterial here. The f_i's are in the directions of the vertices of P°. If μ_i denotes the area of the facet F_i, then:

$$\Pi(P) = 1/2 \sum_i^m \mu_i / 2[-f_i, f_i].$$

One factor of $1/2$ is to make the line segments be of length μ_i, the other is because the projection of P in some direction is covered twice by the projections of the facets; if P is symmetric, the sum can be taken over half the facets, one from each pair. This construction reveals that Π maps objects in X to objects in the dual space X^*.

Examples. For the standard cube C_3, the normals are the usual basis vectors (and their negatives). The areas are all 4, hence, $\Pi(C_3) = \sum 2[-e_i, e_i] = 2C_3$.

For the standard octahedron $\mathcal{O}_3$, the normals are

$$3^{-1/2}(1, 1, 1),\ 3^{-1/2}(1, -1, 1)\ 3^{-1/2}(-1, -1, 1)$$
$$3^{-1/2}(-1, 1, 1)$$

and the areas are all $\sqrt{3}/2$. Another way to find the sum of the corresponding line segments is as the convex hull of the vectors:

$$1/4(\pm(1, 1, 1) \pm (1, -1, 1) \pm (-1, -1, 1) \pm (-1, 1, 1))$$

where all (16) possible choices of signs are taken. Of these, 14 are vertices and the other two are interior points (they are 0). These vertices are those of the rhombic dodecahedron, as stated in Section V (with an extra factor of $1/2$).

For the standard simplex, $\Pi(S_3)$ is also a sum of four line segments and is combinatorially the same as $\Pi(\mathcal{O}_3)$ but it has different proportions. However, the two are affinely equivalent showing that Π is not one to one on the set of all polytopes.

There are two other useful characterizations of zonotopes. If P is a centrally symmetric polytope with k pairs of opposite vertices, then P is a projection of $\mathcal{O}_k$. Similarly, if a zonotope is a sum of k line segments, then it is a projection of C_k. The dual characterization is that the dual of a zonotope is a central cross section of $\mathcal{O}_k$ for some suitable k.

A problem of great antiquity is that of characterizing and classifying the tilings of space. A more tractable problem is to ask what objects can be used as a single tile that will tile space by translation. In $\mathbb{R}^3$ these are known. There are five such objects and all are zonohedra. In higher dimensions, it is known (Venkov) that such objects are centrally symmetric polytopes with centrally symmetric facets. However, for $n > 3$ it was noted above that these need not be zonotopes. For results on tilings, see the article of that title by Schulte.[2]

C. Zonoids

Whereas the polytopes form a dense set among the convex bodies in $\mathbb{R}^n$, this is far from true of the zonotopes. The class of zonoids is defined as the closure of the set of zonotopes in the Hausdorff metric (in some fixed vector space). Thus, every zonoid is the limit of a sequence of zonotopes. Since the map from convex body to support function is continuous, it follows that the support function of a zonoid is the limit of a sequence of functions of the form $\sum_i |f(x_i)|$. Such limits can be written as an integral. Thus, a convex body Z is a zonoid if and only if its support function has the form:

$$h_Z(f) = \int |f(x)| d\rho(x)$$

where the integral is over the surface of the unit ball and ρ is an even measure on that surface.

The set of zonoids is also the range of the projection body operator Π when its domain is the set of all convex

bodies. This operator is one to one on the set of centrally symmetric bodies. An important context in which zonoids arise is that of vector measures. The range of any non-atomic vector measure is a zonoid.

In finite-dimensional normed spaces, there are several ways in which area can be defined. For that due to Holmes and Thompson the solution to the isoperimetric problem is always a zonoid. This fact makes that definition especially suitable for integral geometry. An interesting open question is whether there are non-Euclidean normed spaces for which the solution to the isoperimetric problem is the unit ball. For further results about zonoids, consult the article by Goodey and Weil.[2]

D. Ellipsoids

An ellipsoid in $\mathbb{R}^n$ was defined in Section III to be an affine image of the unit ball. Here, we restrict attention to ellipsoids with the origin as center and as an interior point; that is, the image of the unit ball under an invertible linear map T. Alternatively, if S is a positive definite symmetric matrix, the ellipsoid E_S is defined (using the inner product) by:

$$E_S := \{x : \langle Sx, x\rangle \leq 1\}.$$

Thus, the unit ball B is E_I where I is the identity matrix. The two approaches are connected by the fact that if T is an invertible matrix then $S := T^{-1*}T^{-1}$ is positive definite symmetric and $T(B) = T(E_I) = E_S$. Conversely, a positive definite matrix S has a positive definite square root $S^{1/2}$ and

$$E_S = S^{-1/2}(E_I) = S^{-1/2}(B).$$

Many properties of ellipsoids may be regarded as facts about positive definite symmetric matrices and conversely. If an ellipsoid E is given in the form $E = T(B)$, then $V_n(E) = \det T\, \epsilon_n$. Similarly, $V_n(E_S) = \epsilon_n/\sqrt{\det S}$. Useful here is the fact that the determinant of a matrix is the product of its eigenvalues.

The gauge function of a centrally symmetric convex body with 0 as an interior point is a norm. Ellipsoids are precisely those bodies whose corresponding norms come from an inner product—a Euclidean norm and (in infinite dimensions) a Hilbert space norm. Therefore, theorems that characterize ellipsoids among convex bodies also characterize inner product spaces among normed spaces (and conversely). Theorems of this type are extremely numerous in the literature. Here we give a small sample.

The most well-known characterization of inner product spaces is via the parallelogram law and is due to Jordan and von Neumann: a norm $\|\,.\,\|$ comes from an inner product if and only if:

$$\|x + y\|^2 + \|x - y\|^2 = \|x\|^2 + \|y\|^2$$

for all vectors x and y. Thus, a convex body is an ellipsoid if and only if its gauge function satisfies the above equation. Another characterization of inner product spaces is that of Kakutani and is closely related to the Blaschke result below.

Theorem (Kakutani). If $(X, \|\,.\,\|)$ is a normed space of dimension $n \geq 3$ and if for every two-dimensional subspace H of X there is a projection of X onto H which does not increase the norm of any vector, then the norm comes from an inner product.

Ellipsoids are characterized by the equality case in a number of the inequalities in Section IX. A few more geometric results are the following.

Theorem (Bertrand, Brunn). Let K be a convex body. For every vector x, consider the chords of K that are parallel to x. The midpoints of these chords all lie on a hyperplane if and only if K is an ellipsoid with center on that hyperplane.

Definition. If K is a convex body, if $x \neq 0$ is a vector, and if L_x is the line spanned by x, then the shadow boundary of K in the direction x is the set $\mathcal{S}_x := \partial(K + L_x) \cap \partial K$.

The image here is of a beam of light shining on K in the direction x. Part of the body is illuminated and part is in shadow. The set $\mathcal{S}_x$ is the "edge" of the shadow. If K has some flat parts to its boundary, the edge may be broad.

Theorem (Blaschke). If $n \geq 3$ and if K is a convex body, then the shadow boundary $\mathcal{S}_x$ lies in a hyperplane for all x if and only if K is an ellipsoid.

Theorem (Shaĭdenko, Goodey). If $\lambda \geq 3$ and if K and L are two convex bodies such that for all translates $L + x$ of L (except if $L + x = K$) the set $\partial K \cap \partial(L + x)$ is contained in a hyperplane, then K and L are homothetic ellipsoids.

Lastly, we recall the Löwner-John ellipsoids (see Section VI). For every convex body K, there is a *unique* ellipsoid of maximal volume inscribed to K and a *unique* ellipsoid of minimal volume circumscribed to K.

E. Simplices

In many respects, among convex bodies simplices stand at the opposite extreme from ellipsoids. Whereas ellipsoids

are among the most symmetric of bodies, simplices are among the most asymmetric. In some of the inequalities in Section IX, ellipsoids are precisely at one extreme and simplices at the other.

The place where many students and users of mathematics meet the word "simplex" is in the terms *simplex method* or *simplex algorithm*; see Webster.[5]

Choquet observed that if K is a convex body such that its intersection with any homothetic image is again a homothetic image then K is a simplex. He used this idea to define a simplex in infinite-dimensional spaces. Moreover, if K is of dimension $(n-1)$ and is contained in a hyperplane H_f^1 in $\mathbb{R}^n$ (*not* a hyperplane through 0), then one can construct a cone $C_K := \{\lambda x : x \in K,\ \lambda \geq 0\}$ which induces an order relation on $\mathbb{R}^n$ by:

$$x \leq y \text{ if and only if } y - x \in C_K.$$

This order relation makes $\mathbb{R}^n$ into a lattice if and only if K is a simplex. This is the starting point of a far-reaching theory of infinite-dimensional spaces initiated by Choquet.

Finally, we give a characterization due to Martini that nicely connects three of the operations of Section V.

Theorem (Martini). A polytope is a simplex if and only if $D(K) = \lambda \Pi(K)^\circ$ for some $\lambda > 0$.

Other characterizations of (and information about) simplices can be found in the article by Heil and Martini in the *Handbook of Convex Geometry*.[2]

F. Sets of Constant Width

The width of a convex set in a direction f, $w_K(f)$, was defined in Section V and sets of constant width were also introduced there. In Section V, it was shown that K is of constant width if and only if $D(K)$ is a ball. If K_1 and K_2 are each of constant width, then so is $K_1 + K_2$.

For each point x in ∂K, if we define the diameter of K at x to be $d_K(x) := \sup\{\|x - y\| : y \in K\}$, then K is of constant width if and only if $d_K(x)$ is constant.

If the convex body K has diameter d, we may define the spherical hull of K, $\mathrm{sh}(K)$, to be the intersection of all balls of radius d with center in K:

$$\mathrm{sh}(K) := \bigcap \{B[x, d] : x \in K\}.$$

Then $K \subseteq \mathrm{sh}(K)$ and equality occurs if and only if K is of constant width. Sets of constant width are precisely those sets of diameter d such that if $y \notin K$ then $\mathrm{diam}(K \cup \{y\}) > d$. Moreover, every set K of diameter d is contained in a set of constant width d.

The inradius of a convex body K is the maximal radius of a ball contained in K, and the circumradius is the minimal radius of a ball that contains K. If K has inradius r and circumradius R and is of constant width w, then there is a unique inscribed ball of radius r and a unique circumscribed ball of radius R, these balls are concentric, and $r + R = w$.

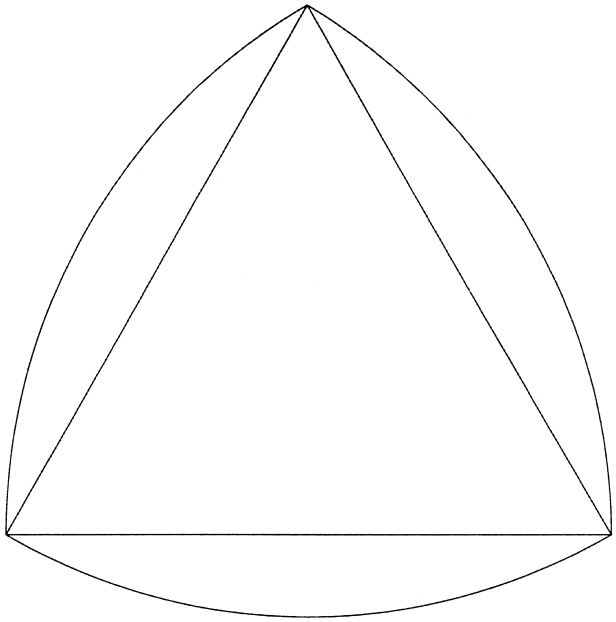


FIGURE 8

There is an inequality relating the volume and surface area of bodies of constant width w in R^n:

$$V_n(K)/V_{n-1}(\partial K) \leq w/2n.$$

Two-dimensional sets of constant width have been much studied. A Reuleaux triangle of width w is the intersection of three discs of radius w centered at the vertices of an equilateral triangle (see Fig. 8). Reuleaux polygons with any odd number of sides may be constructed similarly. Discs $B[x, w/2]$ are the sets of constant width w whose area is maximal.

Theorem (Blaschke, Lebesgue). If K is a two-dimensional set of constant width w, then:

$$V_2(K) \geq \left(\pi - \sqrt{3}\right) w^2/2$$

with equality if and only if K is a Reuleaux triangle.

The corresponding problem in higher dimensions is unsolved. Constructions analogous to the Reuleaux polygons can be made in higher dimensions but not so easily.

Cauchy's formula for surface area (Section IX) yields an elegant result for bodies of constant width in $\mathbb{R}^2$. The length of the perimeter of such a body is πw where w is the width.

SEE ALSO THE FOLLOWING ARTICLES

FUNCTIONAL ANALYSIS • LINEAR OPTIMIZATION • SET THEORY • TOPOLOGY, GENERAL

BIBLIOGRAPHY

Eggleston, H. G. (1958). "Cambridge Tracts in Mathematics and Physics," Vol. 47 "Convexity," Cambridge University Press, Cambridge, U.K.

Gruber, P. M., and Wills, J. M. (eds.) (1993). "Handbook of Convex Geometry" (two vols.), North-Holland, Amsterdam.

Minkowski, H. (1911). "*Gesammelte Abhandlungen*," Vol. II, Teubner, Leipzig.

Schneider, R. (1993). "Encyclopedia of Math. and Its Applications," Vol. 44, "Convex bodies: the Brunn-Minkowski Theory," Cambridge University Press, Cambridge, U.K.

Webster, R. J. (1994). "Convexity," Oxford University Press, London.



Data Mining, Statistics

David L. Banks
U.S. Department of Transportation

Stephen E. Fienberg
Carnegie Mellon University

I. Two Problems
II. Classification
III. Cluster Analysis
IV. Nonparametric Regression

GLOSSARY

Bagging A method that averages predictions from several models to achieve better predictive accuracy.
Boosting A method that uses outcome-based weighting to achieve improved predictive accuracy in a classification rule.
Classification The statistical problem of assigning new cases into categories based on observed values of explanatory variables, using previous information on a sample of cases for which the categories are known.
Clustering The operation of finding groups within multivariate data. One wants the cases within a group to be similar and cases in different groups to be clearly distinct.
Curse of dimensionality The unfortunate fact that statistical inference becomes increasingly difficult as the number of explanatory variables becomes large.
Nonparametric regression Methods that attempt to find a functional relationship between a response variable and one or more explanatory variables without making strong assumptions (such as linearity) about the form of the relationship.
Principle of parsimony This asserts that a simpler model usually has better predictive accuracy than a complex model, even if the complex model fits the original sample of data somewhat better.
Projection pursuit An exploratory technique in which one looks for low-dimensional projections of high-dimensional data that reveal interesting structure or define useful new explanatory variables.
Recursive partitioning An approach to regression or classification in which one fits different models in different regions of the space of explanatory variables and the data are used to identify the regions that require distinct models.
Smoothing Any operation that does local averaging or local fitting of the response values in order to estimate a functional relationship.

DATA MINING is an emerging analytical area of research activity that stands at the intellectual intersection of

statistics, computer science, and database management. It deals with very large datasets, tries to make fewer theoretical assumptions than has traditionally been done in statistics, and typically focuses on problems of classification, clustering, and regression. In such domains, data mining often uses decision trees or neural networks as models and frequently fits them using some combination of techniques such as bagging, boosting/arcing, and racing. These domains and techniques are the primary focus of the present article. Other activities in data mining focus on issues such as causation in large-scale systems (e.g., Spirtes *et al.* 2001; Pearl, 2000), and this effort often involves elaborate statistical models and, quite frequently, Bayesian methodology and related computational techniques (e.g., Cowell *et al.*, 1999, Jordan, 1998). For an introductory discussion of this dimension of data mining see Glymour *et al.* (1997).

The subject area of data mining began to coalesce in 1990, as researchers from the three parent fields discovered common problems and complementary strengths. The first KDD workshop (on Knowledge Discovery and Data Mining) was held in 1989. Subsequent workshops were held in 1991, 1993, and 1994, and these were then reorganized as an annual international conference in 1995. In 1997 the *Journal of Data Mining and Knowledge Discovery* was established under the editorship of Usama Fayyad. It has published special issues on electronic commerce, scalable and parallel computing, inductive logic programming, and applications to atmospheric sciences. Interest in data mining continues to grow, especially among businesses and federal statistical agencies; both of these communities gather large amounts of complex data and want to learn as much from their collections as is possible.

The needs of the business and federal communities have helped to direct the growth of data mining research. For example, typical applications in data mining include the following:

- Use of historical financial records on bank customers to predict good and bad credit risks.
- Use of sales records and customer demographics to perform market segmentation.
- Use of previous income tax returns and current returns to predict the amount of tax a person owes.

These three examples imply large datasets without explicit model structure, and the analyses are driven more by management needs than by scientific theory. The first example is a classification problem, the second requires clustering, and the third would employ a kind of regression analysis.

In these examples the raw data are numerical or categorical values, but sometimes data mining treats more exotic situations. For example, the data might be satellite photographs, and the investigator would use data mining techniques to study climate change over time. Or they can be other forms of images such as those arising from functional magnetic resonance imaging in medicine, or robot sensors. Or the data might be continuous functions, such as spectra from astronomical studies of stars. This review focuses upon the most typical applications, but the reader can obtain a fuller sense of the scope by examining the repository of benchmark data maintained at the University of California at Irvine (Bay, 1999). This archive is resource for the entire machine learning community, who use it to test and tune new algorithms. It contains, among other datasets, all those used in the annual KDD Cup contest, a competitive comparison of data mining algorithms run by the organizers of the KDD conference.

It is unclear whether data mining will continue to gain intellectual credibility among academic researchers as a separate discipline or whether it will be subsumed as a branch of statistics or computer science. Its commercial success has created a certain sense of distance from traditional research universities, in part because so much of the work is being done under corporate auspices. But there is broad agreement that the content of the field has true scientific importance, and it seems certain that under one label or another, the body of theory and heuristics that constitutes the core of data mining will persist and grow.

The remainder of this article describes the two practical problems whose tension defines the ambit of data mining, then reviews the three basic kinds of data mining applications: classification, clustering, and regression. In parallel, we describe three relatively recent ideas that are characteristic of the intellectual cross-fertilization that occurs in data mining, although they represent just a small part of the work going on in this domain. Boosting is discussed in the context of classification, racing is discussed in the context of clustering, and bagging is discussed in the context of regression. Racing and bagging have broad applicability, but boosting is (so far) limited in application to classification problems. Other techniques for these problems of widespread interest not discussed here include support vector machines and kernel smoothing methods (e.g., Vapnik, 2000).

I. TWO PROBLEMS

Data mining exists because modern methods of data collection and database management have created sample sizes large enough to overcome (or partially overcome) the limitations imposed by the curse of dimensionality. However, this freedom comes at a price—the size of the datasets severely restricts the complexity of the calculations that can be made. These two issues are described in the following subsections.

A. The Curse of Dimensionality

The curse of dimensionality (COD) was first described by Richard Bellman, a mathematician, in the context of approximation theory. In data analysis, the term refers to the difficulty of finding hidden structure when the number of variables is large.

For all problems in data mining, one can show that as the number of explanatory variables increases, the problem of structure discovery becomes harder. This is closely related to the problem of variable selection in model fitting. Classical statistical methods, such as discriminant analysis and multiple linear regression, avoided this problem by making the strong model assumption that the mathematical relationship between the response and explanatory variables was linear. But data miners prefer alternative analytic approaches, since the linearity assumption is usually wrong for the kinds of applications they encounter.

For specificity, consider the case of regression analysis; here one looks for structure that predicts the value of the response variable Y from explanatory variables $\boldsymbol{X} \in \mathbb{R}^p$. Thus one might want to predict the true amount of tax that an individual owes from information reported in the previous year, information declared on the current form, and ancillary demographic or official information. The COD arises when p is large.

There are three essentially equivalent descriptions of the COD, but each gives a usefully different perspective on the problem:

1. For large p nearly all datasets are too sparse.
2. The number of regression functions to consider grows quickly (faster than exponentially) with p, the dimension of the space of explanatory variables.
3. For large p, nearly all datasets are multicollinear (or concurve, the nonparametric generalization of multicollinearity).

These problems are minimized if data can be collected using an appropriate statistical design, such as Latin hypercube sampling (Stein, 1987). However, aside from simulation experiments, this is difficult to achieve.

The first version of the curse of dimensionality is most easily understood. If one has five points at random in the unit interval, they tend to be close together, but five random points in the unit square and the unit cube tend to be increasingly dispersed. As the dimension increases, a sample of fixed size provides less information on local structure in the data.

To quantify this heuristic about increasing data sparsity, one can calculate the side length of a p-dimensional subcube that is expected to contain half of the (uniformly random) data in the p-dimensional unit cube. This side-length is $(0.5)^{1/p}$, which increases to 1 as p gets large. Therefore the expected number of observations in a fixed volume in $\mathbb{R}^p$ goes to zero, which implies that the data can provide little information on the local relationship between $\boldsymbol{X}$ and Y. Thus it requires very large sample sizes to find local structure in high dimensions.

The second version of the COD is related to complexity theory. When p is large there are many possible models that one might fit, making it difficult for a finite sample properly to choose among the alternative models. For example, in predicting the amount of tax owed, competing models would include a simple one based just on the previous year's payment, a more complicated model that modified the previous year's payment for wage inflation, and a truly complex model that took additional account of profession, location, age, and so forth.

To illustrate the explosion in the number of possible models, suppose one decides to use only a polynomial model of degree 2 or less. When $p = 1$ (i.e., a single explanatory variable), there are seven possible regression models:

$$
\begin{aligned}
Y &= \beta_0 + \epsilon, \\
Y &= \beta_0 + \beta_1 X_1 + \epsilon, \\
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon, \\
Y &= \beta_1 X_1 + \epsilon, \\
Y &= \beta_0 + \beta_2 X_1^2 + \epsilon, \\
Y &= \beta_1 X_1 + \beta_2 X_1^2 + \epsilon, \\
Y &= \beta_0 + \beta_2 X_1^2 + \epsilon,
\end{aligned}
$$

where ϵ denotes noise in the observation of Y. For $p = 2$ there are 63 models to consider (including interaction terms of the form $X_1 X_2$), and simple combinatorics shows that, for general p, there are $2^{1+p+p(p+1)/2} - 1$ models of degree 2 or less. Since real-world applications usually explore much more complicated functional relationships than low-degree polynomials, data miners need vast quantities of data to discriminate among the many possibilities.

Of course, it is tempting to say that one should include all the explanatory variables, and in the early days of data mining many naive computer scientists did so. But statisticians knew that this violated the principle of parsimony and led to inaccurate prediction. If one does not severely limit the number of variables (and the number of transformations of the variables, and the number of interaction terms between variables) in the final model, then one ends up *overfitting* the data. Overfit happens when the chosen model describes the accidental noise as well as the true signal. For example, in the income tax problem it

might by chance happen that people whose social security numbers end in 2338 tend to have higher incomes. If the model fitting process allowed such frivolous terms, then this chance relationship would be used in predicting tax obligation, and such predictions would be less accurate than those obtained from a more parsimonious model that excluded misinformation. It does no good to hope that one's dataset lacks spurious structure; when p is large, it becomes mathematically certain that chance patterns exist.

The third formulation of the COD is more subtle. In standard multiple regression, multicollinearity arises when two or more of the explanatory variables are highly correlated, so that the the data lie mostly inside an affine subspace of $\mathbb{R}^p$ (e.g., close to a line or a plane within the p-dimensional volume). If this happens, there are an uncountable number of models that fit the data about equally well, but these models are dramatically different with respect to predictions for future responses whose explanatory values lie outside the subspace. As p gets large, the number of possible subspaces increases rapidly, and just by chance a finite dataset will tend to concentrate in one of them.

The problem of multicollinearity is aggravated in nonparametric regression, which allows nonlinear relationships in the model. This is the kind of regression most frequently used in data mining. Here the analogue of multicollinearity arises when predictors concentrate on a smooth manifold within $\mathbb{R}^p$, such as a curved line or sheet inside the p-volume. Since there are many more manifolds than affine subspaces, the problem of concurvity in nonparametric regression distorts prediction even more than does multicollinearity in linear regression.

When one is interested only in prediction, the COD is less of a problem for future data whose explanatory variables have values close to those observed in the past. But an unobvious consequence of large p is that nearly all new observation vectors tend to be far from those previously seen. Furthermore, if one needs to go beyond simple prediction and develop interpretable models, then the COD can be an insurmountable obstacle. Usually the most one can achieve is local interpretability, and that happens only where data are locally dense. For more detailed discussions of the COD, the reader should consult Hastie and Tibshirani (1990) and Scott and Wand (1991).

The classical statistical or psychometrical approach to dimensionality reduction typically involves some form of principal component analysis or multidimensional scaling. Roweis and Saul (2000) and Tenenbaum *et al.* (2000) suggest some novel ways to approach the problem of nonlinear dimensionality reduction in very high dimensional problems such as those involving images.

B. Massive Datasets

Massive datasets pose special problems in analysis. To provide some benchmarks of difficulty, consider the following taxonomy by dataset size proposed by Huber (1995):

Type	Size (bytes)
Tiny	10^2
Small	10^4
Medium	10^6
Large	10^8
Huge	10^{10}
Monster	10^{12}

Huber argues that the category steps, which are factors of 100, correspond to reasonable divisions at which the quantitative increase in sample size compels a qualitative change in the kind of analysis.

For tiny datasets, one can view all the data on a single page. Computational time and storage are not an issue. This is the realm in which classical statistics began, and most problems could be resolved by intelligent inspection of the data tables.

Small datasets could well defeat tabular examination by humans, but they invite graphical techniques. Scatterplots, histograms, and other visualization techniques are quite capable of structure discovery in this realm, and modern computers permit almost any level of analytic complexity that is wanted. It is usually possible for the analyst to proceed adaptively, looking at output from one trial in order to plan the next modeling effort.

Medium datasets begin to require serious thought. They will contain many outliers, and the analyst must develop automatic rules for detecting and handling them. Visualization is still a viable way to do structure discovery, but when the data are multivariate there will typically be more scatterplots and histograms than one has time to examine individually. In some sense, this is the first level at which the entire analytical strategy must be automated, which limits the flexibility of the study. This is the lowest order in the taxonomy in which data mining applications are common.

Large datasets are difficult to visualize; for example, it is possible that the point density is so great that a scatterplot is completely black. Many common statistical procedures, such as cluster analysis or regression and classification based on smoothing, become resource-intensive or impossible.

Huge and monster datasets put data processing issues at the fore, and analytical methods become secondary. Simple operations such as averaging require only that data be read once, and these are feasible at even the largest

size, but analyses that require more than $\mathcal{O}(n)$ or perhaps $\mathcal{O}(n \log(n))$ operations, for n the sample size, are impossible.

The last three taxa are a mixed blessing for data mining. The good news is that the large sample sizes mitigate the curse of dimensionality, and thus it is possible, in principle, to discover complex and interesting structure. The bad news is that many of the methods and much of the theory developed during a century of explosive growth in statistical research are impractical, given current data processing limitations.

From a computational standpoint, the three major issues in data mining are processor speed, memory size, and algorithmic complexity. Speed is proportional to the number of floating point operations (flops) that must be performed. Using PCs available in 2000, one can undertake analyses requiring about about 10^{13} flops, and perhaps up to 10^{16} flops on a supercomputer. Regarding memory, a single-processor machine needs sufficient memory for about four copies of the largest array required in the analysis, and the backup storage (disk) should be able to hold an amount equal to about 10 copies of the raw data (otherwise one has trouble storing derived calculations and exploring alternative analyses). The algorithmic complexity of the analysis is harder to quantify; at a high level it depends on the number of logical branches that the analyst wants to explore when planning the study, and at a lower level it depends on the specific numerical calculations employed by particular model fitting procedures.

Datasets based on extracting information from the World Wide Web or involving all transactions from banks or grocery stores, or collections of fMRI images can rapidly fill up terabytes of disk storage, and all of these issues become relevant. The last issue, regarding algorithmic complexity, is especially important added on top of the sheer size of some datasets. It implies that one cannot search too widely in the space of possible models, and that one cannot redirect the analysis in midstream to respond to an insight found at an intermediate step.

For more information on the issues involved in preparing superlarge datasets for analysis, see Banks and Parmigiani (1992). For additional discussion of the special problems proposed in the analysis of superlarge datasets, see the workshop proceedings on massive datasets produced by the National Research Council (1997).

II. CLASSIFICATION

Classification problems arise when one has a training sample of cases in known categories and their corresponding explanatory variables. One wants to build a decision rule that uses the explanatory variables to predict category membership for future observations. For example, a common data mining application is to use the historical records on loan applicants to classify new applicants as good or bad credit risks.

A. Methods

Classical classification began in 1935, when Sir Ronald Fisher set his children to gather iris specimens. These specimens were then classified into three species by a botanist, and numerical measurements were made on each flower. Fisher (1936) then derived mathematical formulas which used the numerical measurements to find hyperplanes that best separated the three different species.

In modern classification, the basic problem is the same as Fisher faced. One has a training sample, in which the correct classification is known (or assumed to be accurate with high probability). For each case in the training sample, one has additional information, either numerical or categorical, that may be used to predict the unknown classifications of future cases. The goal is use the information in the learning sample to build a decision function that can reliably classify future cases.

Most applications in data mining use either logistic regression, neural nets, or recursive partitioning to build the decision functions. The next three subsections describe these approaches. For information on the more traditional discriminant analysis techniques, see Press (1982).

1. Logistic Regression

Logistic regression is useful when the response variable is binary but the explanatory variables are continuous. This would be the case if one were predicting whether or not an customer is a good credit risk, using information on their income, years of employment, age, education, and other continuous variables.

In such applications one uses the model

$$P[Y = 1] = \frac{\exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})}{1 + \exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})}, \tag{1}$$

where $Y = 1$ if the customer is a good risk, $\boldsymbol{X}$ is the vector of explanatory variables for that customer, and $\boldsymbol{\theta}$ are the unknown parameters to be estimated from the data. This model is advantageous because, under the transformation

$$p = \ln \frac{P[Y = 1]}{1 - P[Y = 1]}$$

one obtains the linear model $p = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}$. Thus all the usual machinery of multiple linear regression will apply.

Logistic regression can be modified to handle categorical explanatory variables through definition of dummy

variables, but this becomes impractical if there are many categories. Similarly, one can extend the approach to cases in which the response variable is polytomous (i.e., takes more than two categorical values). Also, logistic regression can incorporate product interactions by defining new explanatory variables from the original set, but this, too, becomes impractical if there are many potential interactions. Logistic regression is relatively fast to implement, which is attractive in data mining applications that have large datasets. Perhaps the chief value of logistic regression is that it provides an important theoretical window on the behavior of more complex classification methodologies (Friedman *et al.*, 2000).

2. Neural Nets

Neural nets are a classification strategy that employs an algorithm whose architecture is intended to mimic that of a brain, a strategy that was casually proposed by von Neumann. Usually, the calculations are distributed across multiple nodes whose behavior is analogous to neurons, their outputs are fed forward to other nodes, and these results are eventually accumulated into a final prediction of category membership.

To make this concrete, consider a simple perceptron model in which one is attempting to use a training set of historical data to teach the network to identify customers who are poor credit risks. Figure 1 shows a hypothetical situation; the inputs are the values of the explanatory variables and the output is a prediction of either 1 (for a good risk) or −1 (for a bad risk). The weights in the nodes are developed as the net is trained on the historical sample,

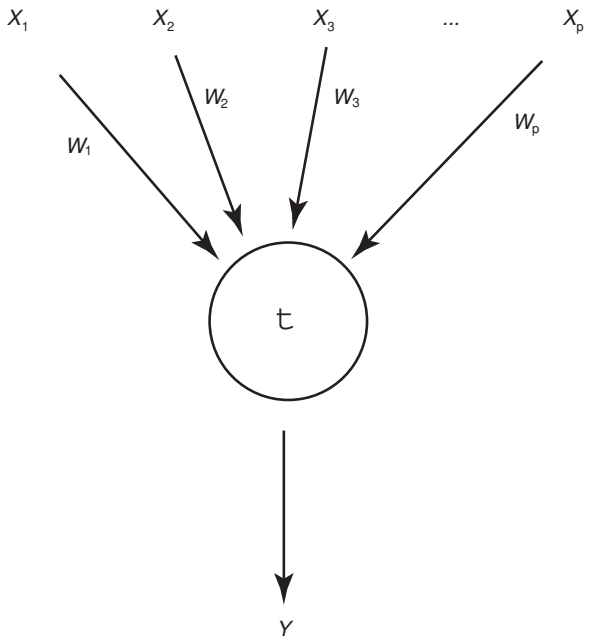


FIGURE 1 Simple perceptron.

and may be thought of as estimated values of model parameters. As diagrammed in Fig. 1, the simple perceptron fits the model

$$y = \text{signum}\left\{\sum_{i=1}^{p} w_i x_i + \tau\right\},$$

where the weights w_i and the threshhold parameter τ are estimated from the training sample.

Unlike logistic regression, it is easy for the simple perceptron to include categorical explanatory variables such as profession or whether the applicant has previously declared bankruptcy, and there is much less technical difficulty in extending the perceptron to the prediction of polytomous outcomes. There is no natural way, however, to include product interaction terms automatically; these still require hand-tuning.

The simple perceptron has serious flaws, and these have been addressed in a number of ways. The result is that the field of neural networks has become complex and diverse. The method shown in Fig. 1 is too primitive to be commonly used in modern data mining, but it serves to illustrate the basic ideas.

As a strategy, neural networks go back to the pioneering work of McCulloch and Pitts (1943). The computational obstacles in training a net went beyond the technology then available. There was a resurgence of attention when Rosenblatt (1958) introduced the perceptron, an early neural net whose properties were widely studied in the 1960s; fundamental flaws with the perceptron design were pointed out by Minsky and Papert (1969). The area languished until the early 1980s, when the Hopfield net (Hopfield, 1982) and the discovery of the backpropagation algorithm [Rumelhart *et al.* (1985) were among several independent inventors] led to networks that could be used in practical applications. Additional information on the history and development of these ideas can be found in Ripley (1996).

The three major drawbacks in using neural nets for data mining are as follows:

1. Neural nets require such computer-intensive that fitting even Huber's large datasets can be infeasible. In practice, analysts who are intent on applying neural nets to their problems typically train the net on only a small fraction of the learning sample and then use the remainder to estimate the accuracy of their classifications.
2. They are difficult to interpret. It is hard to examine a trained network and discover which variables are most influential and what are the functional relationships between the variables. This impedes the kind of scientific insight that is especially important to data miners.

3. Neural nets do not automatically provide statements of uncertainty. At the price of greatly increased computation one can use statistical techniques such as cross-validation, bootstrapping, or jackknifing to get approximate misclassification rates or standard errors or confidence intervals.

Nonetheless, neural nets are one of the most popular tools in the data mining community, in part because of their deep roots in computer science.

Subsection IV.A.4 revisits neural nets in the context of regression rather than classification. Instead of describing the neural net methodology in terms of the nodes and connections which mimic the brain, it develops an equivalent but alternative representation of neural nets as a procedure for fitting a mathematical model of a particular form. This latter perspective is the viewpoint embraced by most current researchers in the field.

3. Recursive Partitioning

Data miners use recursive partitioning to produce decision trees. This is one of the most popular and versatile of the modern classification methodologies. In such applications, the method employs the training sample to recursively partition the set of possible explanatory measurements. The resulting classification rule can displayed as a decision tree, and this is generally viewed as an attractive and interpretable rule for inference.

Formally, recursive partitioning splits the training sample into increasingly homogeneous groups, thus inducing a partition on the space of explanatory variables. At each step, the algorithm considers three possible kinds of splits using the vector of explanatory values $\boldsymbol{X}$:

1. Is $X_i \leq t$ (univariate split)?
2. Is $\sum_{i=1}^{p} w_i x_i \leq t$ (linear combination split)?
3. Does $x_i \in S$ (categorical split, used if x_i is a categorical variable)?

The algorithm searches over all possible values of t, all coefficients $\{w_i\}$, and all possible subsets S of the category values to find the split that best separates the cases in the training sample into two groups with maximum increase in overall homogeneity.

Different partitioning algorithms use different methods for assessing improvement in homogeneity. Some seek to minimize Gini's index of diversity, others use a "twoing rule," and hybrid methods can switch criteria as they move down the decision tree. Similarly, some methods seek to find the greatest improvement on both sides of the split, whereas other methods choose the split that achieves maximum homogeneity on one side or the other. Some methods grow elaborate trees and then prune back to improve predictive accuracy outside the training sample (this is a partial response to the kinds of overfit concerns that arise from the curse of dimensionality).

However it is done, the result is a decision tree. Figure 2 shows a hypothetical decision tree that might be built from credit applicant data. The first split is on income, a continuous variable. To the left-hand side, corresponding to applicants with incomes less than $25,000, the next split is categorical, and divides according to whether the applicant has previously declared bankruptcy. Going back the the top of the tree, the right-hand side splits on a linear combination; the applicant is considered a good risk if a linear combination of their age and income exceeds a threshold. This is a simplistic example, but it illustrates the interpretability of such trees, the fact that the same variable may be used more than once, and the different kinds of splits that can be made.

Recursive partitioning methods were first proposed by Morgan and Sonquist (1963). The method became widely popular with the advent of CART, a statistically sophisticated implementation and theoretical evaluation developed by Breiman *et al.* (1984). Computer scientists have also contributed to this area; prominent implementations of decision trees include ID3 and C4.5 (Quinlan, 1992). A treatment of the topic from a statistical perspective is given by Zhang and Singer (1999). The methodology extends to regression problems, and this is described in Section IV.A.5 from a model-fitting perspective.

B. Boosting

Boosting is a method invented by computer scientists to improve weak classification rules. The idea is that if one has a classification procedure that does slightly better than chance at predicting the true categories, then one can apply this procedure to the portions of the training sample that are misclassified to produce new rules and then weight all the rules together to achieve better predictive accuracy. Essentially, each rule has a weighted vote on the final classification of a case.

The procedure was proposed by Schapire (1990) and improved by Freund and Schapire (1996) under the name AdaBoost. There have been many refinements since, but the core algorithm for binary classification assumes one has a weak rule $g_1(\mathbf{X})$ that takes values in the set $\{1, -1\}$ according to the category. Then AdaBoost starts by putting equal weight $w_i = n^{-1}$ on each of the n cases in the training sample. Next, the algorithm repeats the following steps K times:

1. Apply the procedure g_k to the training sample with weights $w_1, \ldots, w_n$.

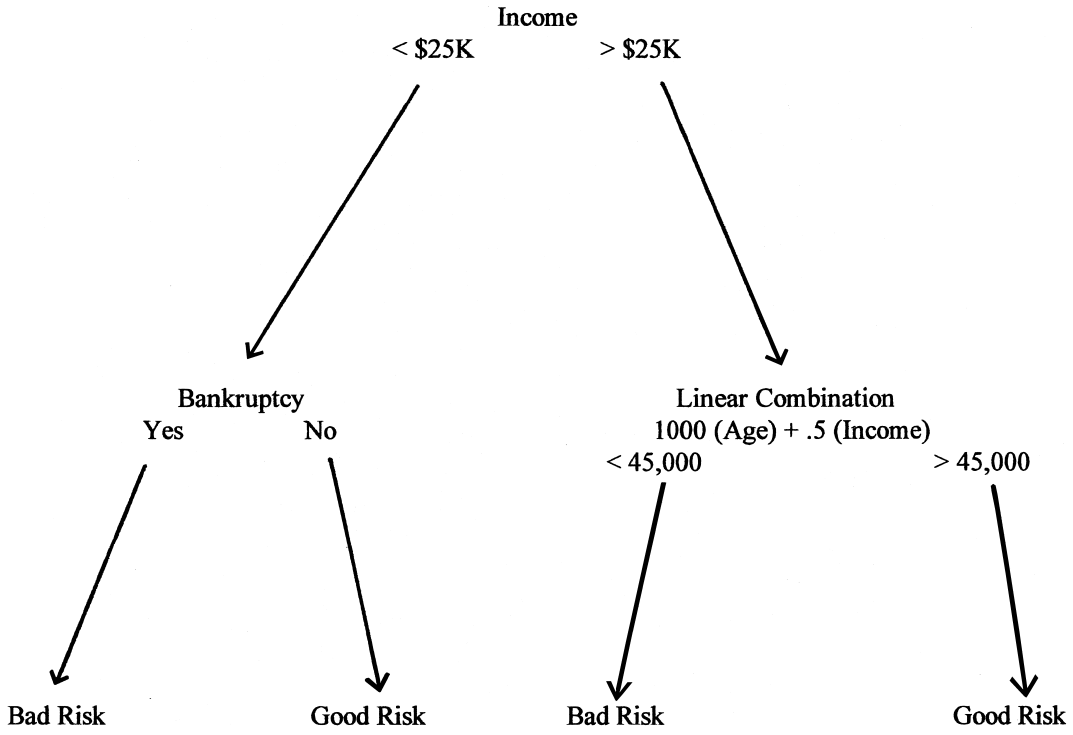


FIGURE 2 Decision tree.

2. Find the empirical probability p_w of misclassification under these weightings.
3. Calculate $c_k = \ln[(1 - p_w)/p_w]$.
4. If case i is misclassified, replace w_i by $w_i \exp c_k$. Then renormalize so that $\sum_i w_i = 1$ and go to step 1.

The final inference is the sign of $\sum_{k=1}^{K} c_k g_k(\boldsymbol{X})$, which is a weighted sum of the determinations made by each of the K rules formed from the original rule g_1.

This rule has several remarkable properties. Besides provably improving classification, it is also resistant to overfit, which arises when K is large. The procedure allows quick computation, and thus can be made practical even for huge datasets, and it can be generalized to handle more than two categories. Boosting therefore provides an automatic and effective way to increase the capability of almost any classification technique. As a new method, it is the object of active research; Friedman *et al.* (2000) describe the current thinking in this area, linking it to the formal role of statistical models such as logistic regression and generalized additive models.

III. CLUSTER ANALYSIS

Cluster analysis is the term that describes a collection of data mining methods which take observations and form groups that are usefully similar. One common application is in market segmentation, where a merchant has data on the purchases made by a large number of customers, together with demographic information on the customers. The merchant would like to identify clusters of customers who make similar purchases, so as to better target advertising or forecast the effect of changes in product lines.

A. Clustering Strategies

The classical method for grouping observations is hierarchical agglomerative clustering. This produces a cluster tree; the top is a list of all the observations, and these are then joined to form subclusters as one moves down the tree until all cases are merged in a single large cluster. For most applications a single large cluster is not informative, however, and so data miners require a rule to stop the agglomeration algorithm before complete merging occurs. The algorithm also requires a rule to identify which subcluster should be merged next for each stage of the tree-building process.

Statisticians have not found a universally reliable rule to determine when to stop a clustering algorithm, but many have been suggested. Milligan and Cooper (1985) described a large simulation study that included a range of realistic situations. They found that no rule dominated all the others, but that the cubic clustering criterion was rarely

bad and often quite good. However, in practice, most analysts create the entire tree and then inspect it to find a point at which further linkages do not seem to further their purpose. For the market segmentation example, a merchant might be pleased to find clusters that can be interpreted as families with children, senior citizens, yuppies, and so forth, and would want to stop the clustering algorithm when further linkage would conflate these descriptive categories.

Similar diversity exists when choosing a subcluster merging rule. For example, if the point clouds associated with the final interpretable clusters appear ellipsoidal with similar shape and orientation, then one should probably have used a joining rule that connects the two subclusters whose centers have minimum Mahalanobis (1936) distance. Alternatively, if the final interpretable clusters are nonconvex point clouds, then one will have had to discover those by using some kind of nearest neighbor joining rule. Statisticians have devised many such rules and can show that no single approach can solve all clustering problems (Van Ness, 1999).

In data mining applications, most of the joining rules developed by statisticians require infeasible amounts of computation and make unreasonable assumptions about homogeneous patterns in the data. Specifically, large, complex data sets do not generally have cluster structure that is well described by sets of similar ellipsoids. Instead, data miners expect to find structures that look like sheets, ellipsoids, strings, and so forth (rather as astronomers see when looking at the large-scale structure of the universe).

Therefore, among agglomerative clustering schemes, data miners almost always use nearest neighbor clustering. This is one of the fastest clustering algorithms available, and is basically equivalent to finding a minimum spanning tree on the data. Using the Prim (1957) algorithm for spanning trees, the computation takes $\mathcal{O}(n^2)$ comparisons (where n is the number of observations). This is feasible for medium datasets in Huber's taxonomy. Furthermore, nearest neighbor methods are fairly robust at finding the diverse kinds of structure that one anticipates.

As an alternative to hierarchical agglomerative clustering, some data miners use k-means cluster analysis, which depends upon a strategy pioneered by MacQueen (1967). Starting with the assumption that the data contain a prespecified number k of clusters, this method iteratively finds k cluster centers that maximize between-cluster distances and minimize within-cluster distances, where the distance metric is chosen by the user (e.g., Euclidean, Mahalanobis, sup norm, etc.). The method is useful when one has prior beliefs about the likely number of clusters in the data. It also can be a useful exploratory tool. In the computer science community, k-means clustering is known as the quantization problem and is closely related to Voronoi tesselation.

The primary drawback to using k-means clustering in data mining applications is that exact solution requires extensive computation. Approximate solutions can be obtained much more rapidly, and this is the direction in which many researchers have gone. However, the quality of the approximation can be problematic.

Both k-means and agglomerative cluster analysis are strongly susceptible to the curse of dimensionality. For the market segmentation example, it is easy to see that customers could form tight clusters based upon their first names, or where they went to elementary school, or other misleading features that would not provide commercial insight. Therefore it is useful to do variable selection, so that clustering is done only upon features that lead to useful divisions. However, this requires input from the user on which clusters are interpretable and which are not, and encoding such information is usually impractical. An alternative is to use robust clustering methods that attempt to find a small number of variables which produce well-separated clusters. Kaufman and Rousseeuw (1990) review many ideas in this area and pay useful attention to computational issues.

For medium datasets in Huber's taxonomy, it is possible to use visualization to obtain insight into the kinds of cluster structure that may exist. This can provide guidance on the clustering algorithms one should employ. Swayne *et al.* (1997) describe software that enables data miners to see and navigate across three-dimensional projections of high-dimensional datasets. Often one can see groups of points or outliers that are important in the application.

B. Racing

One of the advances that came from the data mining synergy between statisticians and computer scientists is a technique called 'racing' (Maron and Moore, 1993). This enables analysts to do much larger searches of the space of models than was previously possible.

In the context of cluster analysis, suppose one wanted to do variable selection to discover which set of demographic features led to, say, 10 consumer clusters that had small intracluster variation and large intercluster variation. One approach would be to consider each possible subset of the demographic features, run the clustering algorithm, and then decide which set of results had the best cluster separation. Obviously, this would entail much computation.

An alternative approach is to consider many feature subsets simultaneously and run the clustering algorithm for perhaps 1% of the data on each subset. Then one compares the results to see if some subsets lead to less well-defined clusters than others. If so, those subsets are eliminated and the remaining subsets are then tested against each other on a larger fraction of the data. In this way one can weed out poor feature subset choices with minimal computation

and reserve resources for the evaluation of the very best candidates.

Racing can be done with respect to a fixed fraction of the data or a fixed amount of runtime. Obviously, this admits the possibility of errors. By chance a good method might appear poor when tested with only a fraction of the data or for only a fixed amount of computer time, but the probability of such errors can be controlled statistically, and the benefit far outweighs the risk. If one is using racing on data fractions, the problem of chance error makes it important that the data be presented to the algorithm in random order, otherwise the outcome of the race might depend upon subsamples that are not representative of the entire dataset.

Racing has much broader application than cluster analysis. For classification problems, competing classifiers can be tested against each other, and those with high misclassification rates are quickly eliminated. Similarly, for regression problems, competing regression models are raced to quickly eliminate those that show poor fit. In both applications one can easily obtain a 100-fold increase in the size of the model space that is searched, and this leads to the discovery of better classifiers and better regression functions.

IV. NONPARAMETRIC REGRESSION

Regression is a key problem area in data mining and has attracted a substantial amount of research attention. Among dozens of new techniques for nonparametric regression that have been invented over the last 15 years, we detail seven that are widely used. Section IV.A describes the additive model (AM), alternating conditional expectation (ACE), projection pursuit regression (PPR), neural nets (NN), recursive partitioning regression (RPR), multivariate adaptive regression splines (MARS), and locally weighted regression (LOESS). Section IV.B compares these techniques in terms of performance and computation, and Section IV.C describes how bagging can be used to improve predictive accuracy.

A. Seven Methods

The following seven methods may seem very different, but they employ at most two distinct strategies for addressing the curse of dimensionality. One strategy fits purely local models; this is done by RPR and LOESS. The other strategy uses low-dimensional smoothing to achieve flexibility in fitting specific model forms; this is done by AM, ACE, NN, and PPR. MARS combines both strategies.

1. The Additive Model (AM)

Many researchers have developed the AM; Buja *et al.* (1989) describe the early development in this area. The simplest AM says that the expected value of an observation Y_i can be written

$$\mathrm{E}[Y_i] = \theta_0 + \sum_{j=1}^{p} f_j(X_{ij}), \tag{2}$$

where the functions f_j are unspecified but have mean zero.

Since the functions f_j are estimated from the data, the AM avoids the conventional statistical assumption of linearity in the explanatory variables; however, the effects of the explanatory variables are still additive. Thus response is modeled as a sum of arbitrary smooth univariate functions of explanatory variables. One needs about 100 data points to estimate each f_j, but under the model given in (1), the requirement for data grows only linearly in p.

The backfitting algorithm is the essential tool used in estimating an additive model. This algorithm requires some smoothing operation (e.g., kernel smoothing or nearest neighbor averages; Hastie and Tibshirani, 1990) which we denote by $Sm(\cdot|\cdot)$. For a large classes of smoothing operations, the backfitting algorithm converges uniquely.

The backfitting algorithm works as follows:

1. At initialization, define functions $f_j^{(0)} \equiv 0$ and set $\theta_0 = \bar{Y}$.
2. At the ith iteration, estimate $f_j^{(i+1)}$ by
$$f_j^{(i+1)} = Sm\left(\boldsymbol{Y} - \theta_0 - \sum_{k \neq j} f_k^i \middle| X_{1j}, \ldots, X_{nj}\right)$$
for $j = 1, \ldots, p$.
3. Check whether $|f_j^{(i+1)} - f_j^{(i)}| < \delta$ for all $j = 1, \ldots, p$, for δ the prespecified convergence tolerance. If not, go back to step 2; otherwise, take the current $f_j^{(i)}$ as the additive function estimate of f_j in the model.

This algorithm is easy to code. Its speed is chiefly determined by the complexity of the smoothing function.

One can generalize the AM by permitting it to add a few multivariate functions that depend on prespecified explanatory variables. Fitting these would require bivariate smoothing operations. For example, if one felt that prediction of tax owed in the AM would improved by including a function that depended upon both a person's previous year's declaration and their current marital status, then this bivariate smoother could be used in the second step of the backfitting algorithm.

2. Alternating Conditional Expectations (ACE)

A generalization of the AM allows a smoothing transformation of the response variable as well the smoothing of the p explanatory variables. This uses the ACE algorithm, as developed by Breiman and Friedman (1985), and it fits the model

$$\mathrm{E}[g(Y_i)] = \theta_0 + \sum_{j=1}^{p} f_j(X_{ij}). \tag{3}$$

Here all conditions are as stated before for (1), except that g is an arbitrary smooth function scaled to ensure the technically necessary requirement that $\mathrm{var}[g(Y)] = 1$ (if not for this constraint, one could get a perfect fit by setting all functions to be identically zero).

Given data Y_i and $\boldsymbol{X}_i$, one wants to find g, θ_0, and $f_1, \ldots, f_p$ such that $\mathrm{E}[g(Y_i) \mid \boldsymbol{X}_i] - \theta_0 - \sum_{j=1}^{p} f_j(X_{ij})$ is well described as independent error. Thus one solves

$$(\hat{g}, \hat{f}_1, \ldots, \hat{f}_p) = \underset{(g, f_1, \ldots, f_p)}{\operatorname{argmin}} \times \left\{ \sum_{i=1}^{n} \left[g(Y_i) - \sum_{j=1}^{p} f_j(X_{ij}) \right]^2 \right\},$$

where $\hat{g}$ is constrained to satisfy the unit-variance requirement. The algorithm for achieving this is described by Breiman and Friedman (1985); they modify the backfitting algorithm to provide a step that smoothes the left-hand side while maintaining the variance constraint.

ACE analysis returns sets of functions that maximize the linear correlation between the sum of the smoothed explanatory variables and the smoothed response variable. Therefore ACE is more similar in spirit to the multiple correlation coefficient than to multiple regression. Because ACE does not directly attempt a regression analysis, it has certain undesirable features; for example, small changes in the data can lead to very different solutions (Buja and Kass, 1985), it need not reproduce model transformations, and, unlike regression, it treats the explanatory and response variables symmetrically.

To redress some of the drawbacks in ACE, Tibshirani (1988) devised a modification called AVAS, which uses a variance-stabilizing transformation in the backfitting loop when fitting the explanatory variables. This modification is somewhat technical, but in theory it leads to improved properties when treating regression applications.

3. Projection Pursuit Regression (PPR)

The AM uses sums of functions whose arguments are the natural coordinates for the space $\mathbb{R}^p$ of explanatory variables. But when the true regression function is additive with respect to pseudovariables that are linear combinations of the explanatory variables, then the AM is inappropriate. PPR was developed by Friedman and Stuetzle (1981) to address such situations.

Heuristically, imagine there are two explanatory variables and suppose the regression surface is shaped like a sheet of corrugated aluminum. If that sheet is oriented to make the corrugations parallel to the axis of the first explanatory variable (and thus perpendicular to the second), then AM works well. When the aluminum sheet is rotated slightly so that the corrugations do not parallel a natural axis, however, AM fails because the true function is a nonadditive function of the explanatory variables. PPR would succeed, however, because the true function can be written as an additive model whose functions have arguments that are linear combinations of the explanatory variables.

PPR combines the backfitting algorithm with a numerical search routine, such as Gauss–Newton, to fit models of the form

$$\mathrm{E}[Y_i] = \sum_{k=1}^{r} f_k(\boldsymbol{\alpha}_k' \boldsymbol{X}_i). \tag{4}$$

Here the $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r$ are unit vectors that define a set of r linear combinations of explanatory variables. The linear combinations are similar to those used for principal components analysis (Flury, 1988). These vectors need not be orthogonal, and are chosen to maximize predictive accuracy in the model as estimated by cross-validation.

Operationally, PPR alternates calls to two routines. The first routine conditions on a set of pseudovariables given by linear combinations of original variables; these are fed into the backfitting algorithm to obtain an AM in the pseudovariables. The other routine conditions on the estimated AM functions from the previous step and then searches to find linear combinations of the original variables which maximize the fit of those functions. By alternating iterations of these routines, the result converges to a unique solution.

PPR is often hard to interpret for $r > 1$ in (3). When r is allowed to increase without bound, PPR is consistent, meaning that as the sample size grows, the estimated regression function converges to the true function.

Another improvement that PPR offers over AM is this it is invariant to affine transformations of the data; this is often desirable when the explanatory variables are measured in the same units and have similar scientific justifications. For example, PPR might be sensibly used to predict tax that is owed when the explanatory variables are shares of stock owned in various companies. Here it makes sense that linear combinations of shares across commercial sectors would provide better prediction of portfolio appreciation than could be easily obtained from the raw explanatory variables.

4. Neural Nets (NN)

Many neural net techniques exist, but from a statistical regression standpoint (Barron and Barron, 1988), nearly all variants fit models that are weighted sums of sigmoidal

functions whose arguments involve linear combinations of the data. A typical feedforward network uses a model of the form

$$\mathrm{E}[Y] = \beta_0 + \sum_{i=1}^{m} \beta_i f\left(\boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{x} + \gamma_{i0}\right),$$

where $f(\cdot)$ is a logistic function and the β_0, γ_{i0}, and $\boldsymbol{\alpha}_i$ are estimated from the data. Formally, this approach is similar to that in PPR. The choice of m determines the number of hidden nodes in the network and affects the smoothness of the fit; in most cases the user determines this parameter, but it is also possible to use statistical techniques, such as cross-validation to assess model fit, that allow m to be estimated from the data.

Neural nets are widely used, although their performance properties, compared to alternative regression methods, have not been thoroughly studied. Ripley (1996) describes one assessment which finds that neural net methods are not generally competitive. Schwarzer *et al.* (1986) review the use of neural nets for prognostic and diagnostic classification in clinical medicine and reach similar conclusions. Another difficulty with neural nets is that the resulting model is hard to interpret. The Bayesian formulation of neural net methods by Neal (1996) provides a some remedy for this difficulty.

PPR is very similar to neural net methods. The primary difference is that neural net techniques usually assume that the functions f_k are sigmoidal, whereas PPR allows more flexibility. Zhao and Atkeson (1992) show that PPR has similar asymptotic properties to standard neural net techniques.

5. Recursive Partitioning Regression (RPR)

RPR has become popular since the release of CART (Classification and Regression Tree) software developed by Breiman *et al.* (1984). This technique has already been described in the context of classification, so this subsection focuses upon its application to regression. The RPR algorithm fits the model

$$\mathrm{E}[Y_i] = \sum_{j=1}^{M} \theta_j I_{R_j}(\boldsymbol{X}_i),$$

where the $R_1, \ldots, R_M$ are rectangular regions which partition $\mathbb{R}^p$, and $I_{R_j}(\boldsymbol{X}_i)$ denotes an indicator function that takes the value 1 if and only if $\boldsymbol{X}_i \in R_j$ and is otherwise zero. Here θ_j is the estimated numerical value of all responses with explanatory variables in R_j.

RPR was intended to be good at discovering local low-dimensional structure in functions with high-dimensional global dependence. RPR is consistent; also, it has an attractive graphic representation as a decision tree, as illustrated for classification in Fig. 2. Many common functions are difficult for RPR, however; for example, it approximates a straight line by a stairstep function. In high dimensions it can be difficult to discover when the RPR piecewise constant model closely approximates a simple smooth function.

To be concrete, suppose one used RPR to predict tax obligation. The algorithm would first search all possible splits in the training sample observations and perhaps divide on whether or not the declared income is greater than \$25,000. For people with lower incomes, the next search might split the data on marital status. For those with higher incomes, subsequent split might depend on whether the declared income exceeds \$75,000. Further splits might depend on the number of children, the age and profession of the declarer, and so forth. The search process repeats in every subset of the training data defined by previous divisions, and eventually there is no potential split that sufficiently reduces variability to justify further partitions. At this point RPR fits the averages of all the training cases within the most refined subsets as the estimates θ_j and shows the sequence of chosen divisions in a decision tree. (Note: RPR algorithms can be more complex; e.g., CART "prunes back" the final tree by removing splits to achieve better balance of observed fit in the training sample with future predictive error.)

6. Multivariate Adaptive Regression Splines (MARS)

Friedman (1991) proposed a data mining method that combines PPR with RPR through use of multivariate adaptive regression splines. It fits a model formed as a weighted sum of multivariate spline basis functions (tensor-spline basis functions) and can be written as

$$\mathrm{E}[Y_i] = \sum_{k=0}^{q} a_k B_k(\boldsymbol{X}_i),$$

where the coefficients a_k are estimated by (generalized) cross-validation fitting. The constant term is obtained by setting $B_0(X_1, \ldots, X_n) \equiv 1$, and the multivariate spline terms are products of univariate spline basis functions:

$$B_k(x_1, \ldots, x_n) = \prod_{s=1}^{r_k} b\left(x_{i(s,k)} \,|\, t_{s,k}\right), \qquad 1 \le k \le r.$$

The subscript $i(s,k)$ identifies a particular explanatory variable, and the basis spline for that variable puts a knot at $t_{s,k}$. The values of q, the $r_1, \ldots, r_q$, the knot locations, and the explanatory variables selected for inclusion are determined from the data adaptively.

MARS output can be represented and interpreted in a decomposition similar to that given by analysis of vari-

ance. It is constructed to work well if the true function has local dependence on only a few variables. MARS can automatically accommodate interactions between variables, and it handles variable selection in a natural way.

7. Locally Weighted Regression (LOESS)

Cleveland (1979) developed a technique based on locally weighted regression. Instead of simply taking a local average, LOESS fits the model $\mathrm{E}[Y] = \boldsymbol{\theta}(\boldsymbol{x})'\boldsymbol{x}$, where

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} w_i(\boldsymbol{x})(Y_i - \boldsymbol{\theta}'\boldsymbol{X}_i)^2 \qquad (5)$$

and w_i is a weight function that governs the influence of the ith datum according to the direction and distance of $\boldsymbol{X}_i$ from $\boldsymbol{x}$.

LOESS is a consistent estimator but may be inefficient at finding relatively simple structures in the data. Although not originally intended for high-dimensional regression, LOESS uses local information with advantageous flexibility. Cleveland and Devlin (1988) generalized LOESS to perform polynomial regression rather than the linear regression $\boldsymbol{\theta}'\boldsymbol{X}_i$ in (4), but from a data mining perspective, this increases the cost of computation with little improvement in overall predictive accuracy.

B. Comparisons

Nonparametric regression is an important area for testing the performance of data mining methods because statisticians have developed a rich theoretical understanding of the issues and obstacles. Some key sources of comparative information are as follows:

- Donoho and Johnstone (1989) make asymptotic comparisons among the more mathematically tractable techniques. They indicate that projection-based methods (PPR, MARS) perform better for radial functions but kernel-based methods (LOESS) are superior for harmonic functions. (Radial functions are constant on hyperspheres centered at $\mathbf{0}$, while harmonic functions vary sinusoidally on such hyperspheres.)
- Friedman (1991) describes simulation studies of MARS, and related work is cited by his discussants. Friedman focuses on several criteria; these include scaled versions of mean integrated squared error (MISE) and predictive-squared error (PSE). From the standpoint of data mining practice, the most useful conclusions are (1) for data that are pure noise in 5 and 10 dimensions, and sample sizes of 50, 100, and 200, AM and MARS are comparable and both are unlikely to find false structure, and (2) if test data are generated from the following additive function of five variables

$$Y = 0.1\exp(4X_1) + \frac{4}{1+\exp(-20X_2+10)} + 3X_3 + 2X_4 + X_5,$$

with five additional noise variables and sample sizes of 50, 100, and 200, MARS had a slight tendency to overfit, particularly for the smallest sample sizes.

All of these simulations, except for the pure noise condition, have large signal-to-noise ratios.

- Barron (1991, 1993) proved that in a narrow sense, the MISE for neural net estimates in a certain class of functions has order $\mathcal{O}(1/m) + \mathcal{O}(mp/n)\ln n$, where m is the number of nodes, p is the dimension, and n is the sample size. Since this is linear in dimension, it evades the COD; similar results have been obtained by Zhao and Atkeson (1992) for PPR, and it is probable that a similar result holds for MARS. These findings have limited practical value; Barron's class of functions consists of those whose Fourier transform $\tilde{g}$ satisfies $\int |\omega||\tilde{g}(\omega)|\,d\omega < c$ for some fixed c. This excludes such simple cases as hyperflats, and the class becomes smoother as dimension increases.
- De Veaux *et al.* (1993) tested MARS and a neural net on two functions; they found that MARS was faster and had better MISE.
- Zhang and Singer (1999) applied CART, MARS, multiple logistic regression, and other methods in several case studies and found that no method dominates the others. CART was often most interpretable, but logistic regression led more directly to estimation of uncertainty.

The general conclusions are (1) parsimonious modeling is increasingly important in high dimensions, (2) hierarchical models using sums of piecewise-linear functions are relatively good, and (3) for any method, there are datasets on which it succeeds and datasets on which it fails.

In practice, since it is hard to know what methods works best in a given situation, data miners usually hold out a part of the data, apply various regression techniques to the remaining data, and then use the models that are built to estimate the hold-out sample. The method that achieves minimum prediction error is likely to be the best for that application.

Most of the software detailed in this section is available from the StatLib archive at `http://lib.stat.cmu.edu`; this includes AM, ACE, LOESS, and PPR. Both MARS and CART are commercially available from Salford Systems, Inc., at `http://www.salford-systems.com`. Splus includes versions of RPR, the (generalized) AM, ACE, PPR, and LOESS; this

is commercially available from Mathsoft, Inc., at http://www.splus.mathsoft.com.

C. Bagging

Bagging is a strategy for improving predictive accuracy by model averaging. It was proposed by Breiman (1996), but has a natural pedigree in Bayesian work on variable selection, in which one often puts weights on different possible models and then lets the data update those weights.

Concretely, suppose one has a training sample and a nonparametric regression technique that takes the explanatory variables and produces an estimated response value. Then the simplest form of bagging proceeds by drawing K random samples (with replacement) from the training sample and applying the regression technique to each random sample to produce regression rules $T_1(\boldsymbol{X}), \ldots, T_K(\boldsymbol{X})$. For a new observation, say $\boldsymbol{X}^*$, the bagging predictor of the response Y^* is $K^{-1}\sum_{k=1}^{K} T_k(\boldsymbol{X}^*)$.

The idea behind bagging is that model fitting strategies usually have high variance but low bias. This means that small changes in the data can produce very different models but that there is no systematic tendency to produce models which err in particular directions. Under these circumstances, averaging the results of many models can reduce the error in the prediction that is associated with model instability while preserving low bias.

Model averaging strategies are moving beyond simple bagging. Some employ many different kinds of regression techniques rather than just a single method. Others modify the bagging algorithm in fairly complex ways, such as arcing (Brieman, 1998). A nice comparison of some of the recent ideas in this area is given by Dietterich (1998), and Hoeting *et al.* (1999) give an excellent tutorial on more systematic Bayesian methods for model averaging. Model averaging removes the analyst's ability to interpret parameters in the models used and can only be justified in terms of predictive properties.

SEE ALSO THE FOLLOWING ARTICLES

ARTIFICIAL NEURAL NETWORKS • DATABASES • DATA STRUCTURES • INFORMATION THEORY • STATISTICS, BAYESIAN • STATISTICS, MULTIVARIATE

BIBLIOGRAPHY

Banks, D. L., and Parmigiani, G. (1992). "Preanalysis of superlarge data sets," *J. Quality Technol.* **24,** 930–945.

Barron, A. R. (1991). "Complexity regularization with aapplications to artificial neural networks." *In* "Nonparametric Functional Estimation" (G. Roussas, ed.), pp. 561–576, Kluwer, Dordrecht.

Barron, A. R. (1993). "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Information Theory* **39,** 930–945.

Barron, A. R., and Barron, R. L. (1988). "Statistical learning networks: A unifying view," *Comput. Sci. Stat.*, **20,** 192–203.

Bay, S. D. (1999). "The UCI KDD Archive," http://kdd.ics.uci.edu. Department of Information and Computer Science, University of California, Irvine, CA.

Breiman, L. (1996). "Bagging predictors," *Machine Learning* **26,** 123–140.

Breiman, L. (1998). "Arcing classifiers," *Ann. Stat.* **26,** 801–824.

Breiman, L., and Friedman, J. (1985). "Estimating optimal transformations for multiple regression and correlation," *J. Am. Stat. Assoc.* **80,** 580–619.

Breiman, L., Friedman, J., Olshen, R.A., and Stone, C. (1984). "Classification and Regression Trees," Wadsworth, Belmont, CA.

Buja, A., and Kass, R. (1985). "Discussion of 'Estimating optimal transformations for multiple regression and correlation,' by Breiman and Friedman," *J. Am. Stat. Assoc.* **80,** 602–607.

Buja, A., Hastie, T. J., and Tibshirani, R. (1989). "Linear smoothers and additive models," *Ann. Stat.* **17,** 453–555.

Cleveland, W. (1979). "Robust locally weighted regression and smoothing scatterplots," *J. Am. Stat. Assoc.* **74,** 829–836.

Cleveland, W., and Devlin, S. (1988). "Locally weighted regression: An approach to regression analysis by local fitting," *J. Am. Stat. Assoc.* **83,** 596–610.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. G. (1999). "Probabilistic Networks and Expert Systems," Springer-Verlag, New York.

Deitterich, T. (1998). "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning* **28,** 1–22.

De Veaux, R. D., Psichogios, D. C., and Ungar, L. H. (1993). "A comparison of two nonparametric estimation schemes: MARS and neural networks," *Computers Chem. Eng.* **17,** 819–837.

Donoho, D. L., and Johnstone, I. (1989). "Projection based approximation and a duality with kernel methods," *Ann. Stat.* **17,** 58–106.

Fisher, R. A. (1936). "The use of multiple measurements in taxonomic problems," *Ann. Eugen.* **7,** 179–188.

Flury, B. (1988). "Common Principal Components and Related Multivariate Models," Wiley, New York.

Freund, Y., and Schapire, R. E. (1996). "Experiments with a new boosting algorithm." *In* "Machine Learning: Proceedings of the Thirteenth International Conference," pp. 148–156, Morgan Kaufmann, San Mateo, CA.

Friedman, J. H. (1991). "Multivariate additive regression splines," *Ann. Stat.* **19,** 1–66.

Friedman, J. H., and Stuetzle, W. (1981). "Projection pursuit regression," *J. Am. Stat. Assoc.* **76,** 817–23.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2000). "Additive logistic regression: A statistical view," *Ann. Stat.* **28,** 337–373.

Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. (1997). "Statistical themes and lessons for data mining," *Data Mining Knowledge Discovery* **1,** 11–28.

Hastie, T. J., and Tibshirani, R. J. (1990). "Generalized Additive Models," Chapman and Hall, New York.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian model averaging: A tutorial," *Stat. Sci.* **14,** 382–417.

Hopfield, J. J. (1982). "Neural networks and physical systems with emergent collective computational abilities," *Proc. Natl. Acad. Sci. USA* **79,** 2554–2558.

Huber, P. J. (1994). "Huge data sets." *In* "Proceedings of the 1994 COMPSTAT Meeting," (R. Dutter and W. Grossmann, eds.), pp. 221–239, Physica-Verlag, Heidelberg.

Jordan, M. I., (ed.). (1998). "Learning in Graphical Models," MIT Press, Cambridge, MA.

Kaufman, L., and Rousseeuw, P. J. (1990). "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley, New York.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations." *In* "Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability," pp. 281–297, University of California Press, Berkeley, CA.

Mahalanobis, P. C. (1936). "On the generalized distance in statistics," *Proc. Natl. Inst. Sci. India* **12,** 49–55.

Maron, O., and Moore, A. W. (1993). "Hoeffding races: Accelerating model selection search for classification and function approximation." *In* "Advances in Neural Information Processing Systems 6," pp. 38–53, Morgan Kaufmann, San Mateo, CA.

McCulloch, W. S., and Pitts, W. (1943). "A Logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.* **5,** 115–133.

Milligan, G. W., and Cooper, M. C. (1985). "An examination of procedures for determining the number of clusters in a dataset," *Psychometrika* **50,** 159–179.

Minsky, M., and Papert, S. A. (1969). "Perceptrons: An Introduction to Computational Geometry," MIT Press, Cambridge, MA.

Morgan, J. N., and Sonquist, J. A. (1963). "Problems in the analysis of survey data and a proposal," *J. Am. Stat. Assoc.* **58,** 415–434.

National Research Council. (1997). "Massive Data Sets: Proceedings of a Workshop," National Academy Press, Washington, DC.

Neal, R. (1996). "Bayesian Learning for neural networks," Springer-Verlag, New York.

Pearl J. (1982). "Causality," Cambridge University Press, Cambridge.

Pearl, J. (2000). "Causality: Models, Reasoning and Inference," Cambridge University Press, Cambridge.

Press, S. J. (1982). "Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference," 2nd ed., Krieger, Huntington, NY.

Prim, R. C. (1957). "Shortest connection networks and some generalizations." *Bell Syst. Tech. J.* **36,** 1389–1401.

Quinlan, J. R. (1992). "C4.5 : Programs for Machine Learning," Morgan Kaufmann, San Mateo, CA.

Ripley, B. D. (1996). "Pattern Recognition and Neural Networks," Cambridge University Press, Cambridge.

Rosenblatt, F. (1958). "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychol. Rev.* **65,** 386–408.

Roweis, S. T., and Saul, L. K. (2000). "Nonlinear dimensionality reduction by local linear embedding," *Science* **290,** 2323–2326.

Rumelhart, D., Hinton, G. E., and Williams, R. J. (1986). "Learning representations by back-propagating errors," *Nature* **323,** 533–536.

Schapire, R. E. (1990). "The strength of weak learnability," *Machine Learning* **5,** 197–227.

Schwarzer, G., Vach, W., and Schumacher, M. (1986). "On misuses of artificial neural networks for prognostic and diagnostic classification in oncology," *Stat. Med.* **19,** 541–561.

Scott, D. W., and Wand, M. P. (1991). "Feasibility of multivariate density estimates," *Biometrika* **78,** 197–206.

Spirtes, P., Glymour, C., and Scheines, R. (2001). "Causation, Prediction, and Search," 2nd ed., MIT Press, Cambridge, MA.

Stein, M. L. (1987). "Large sample properties of simulations using latin hypercube sampling," *Technometrics* **29,** 143–151.

Swayne, D. F., Cook, D., and Buja, A. (1997). "XGobi: Interactive dynamic graphics in the X window system," *J. Comput. Graphical Stat.* **7,** 113–130.

Tennenbaum, J. B., de Silva, V., and Langford, J. C. (2000). "A global geometric framework for nonlinear dimensionality reduction," *Science* **290,** 2319–2323.

Tibshirani, R. (1988). "Estimating optimal transformations for regression via additivity and variance stabilization," *J. Am. Stat. Assoc.* **83,** 394–405.

Van Ness, J. W. (1999). "Recent results in clustering admissibility." *In* "Applied Stochastic Models and Data Analysis," (H. Bacelar-Nicolau, F. Costa Nicolau, and J. Janssen, eds.), pp. 19–29, Instituto Nacional de Estatistica, Lisbon, Portugal.

Vapnik, V. N. (2000). "The Nature of Statistical Learning," 2nd ed., Springer-Verlag, New York.

Zhang, H., and Singer, B. (1999). "Recursive Partitioning in the Health Sciences," Springer-Verlag, New York.

Zhao, Y., and Atkeson, C. G. (1992). "Some approximation properties of projection pursuit networks." *In* "Advances in Neural Information Processing Systems 4" (J. Moody, S. J. Hanson, and R. P. Lippmann, eds.), pp. 936–943, Morgan Kaufmann, San Mateo, CA.



Designs and Error-Correcting Codes

K. T. Phelps
C. A. Rodger
Auburn University

I. Introduction
II. Perfect Codes
III. Constant Weight Codes
IV. Maximum Distance Separable Codes
V. Convolutional Codes

GLOSSARY

Binary sum, $u+v$ Componentwise addition of u and v with $1+1=0$ (exclusive or).
Binary word A string (or vector or sequence) of 0's and 1's
Code A set of (usually binary) words, often all of the same length.
Combinatorial design A collection of subsets of a set satisfying additional regularity properties.
Decoding Finding the most likely codeword (or message) transmitted.
Distance, $d(u, v)$ The number of coordinates in which u and v differ.
Encoding The assignment of codewords to messages.
Information rate The fraction of information per transmitted bit.
Weight, $wt(u)$ The number of nonzero bits in the word u.

I. INTRODUCTION

Both error-correcting codes and combinatorial designs are areas of discrete (not continuous) mathematics that began in response to applied problems, the first in making the electronic transmission of information reliable and the second in the design of experiments with results being statistically analyzed. It turns out that there is substantial overlap between these two areas, mainly because both are looking for uniformly distributed subsets within certain finite sets. In this article we provide a brief introduction to both areas and give some indication of their interaction.

A. Error-Correcting Codes

Error-correcting codes is a branch of discrete mathematics, electrical engineering, and computer science that has developed over the past 50 years, largely in response to

the dramatic growth of electronic transfer and storage of information. Coding Theory began in the late 1940s and early 1950s with the seminal work of Shannon (1948), Hamming (1950), and Golay (1949). Error-correcting codes' first significant application was in NASA's deep space satellite communications. Other important applications since then have been in storage devices (e.g., compact discs), wireless telephone channels, and geopositioning systems. They are now routinely used in all satellite communications and mobile wireless communications systems.

Since no communication system is ideal, information can be altered, corrupted, or even destroyed by *noise*. Any communication system needs to be able to recognize or *detect* such errors and have some scheme for recovering the information or *correcting* the error. In order to protect against the more likely errors and thus improve the reliability, redundancy must be incorporated into the message. As a crude example, one could simply transmit the message several times in the expectation that the majority will appear correctly at their destination, but this would greatly increase the cost in terms of time or the rate of transmission (or space in storage devices).

The most basic error control scheme involves simply detecting errors and requesting retransmission. For many communications systems, requests for retransmission are impractical or impose unacceptable costs on the communication system's performance. In deep space communications, a request for retransmission would take too long. In speech communications, noticeable delays are unacceptable. In broadcast systems, it is impractical given the multitude of receivers. The problem of correcting errors and recovering information becomes of paramount importance in such constrained situations.

Messages can be thought of as words over some alphabet, but for all practical purposes, all messages are simply strings of 0's and 1's, or *binary words*. Information can be partitioned or blocked up into a sequence of binary words or messages of fixed length k. A (*block*) *code*, C, is a set of binary words of fixed length n, each element of which is called a *codeword*. Mathematically, codewords can be considered to be vectors of length n with elements being chosen from a finite field, normally of order 2, but in some cases from the field $GF(2^r)$. [So, for example, the binary codeword 01101 could also be represented as the vector (0, 1, 1, 0, 1).] There are also convolutional codes where the codewords do not have fixed length (or have infinite length), but these will be discussed later. *Encoding* refers to the method of assigning messages to codewords which are then transmitted. Clearly, the number of codewords has to be at least as large as the number of k-bit messages. The *rate* of the code is k/n, since k bits of information result in n bits being transmitted.

The key to being able to detect or correct errors that occur during transmission is to have a code, C, such that no two codewords are close. The *distance* between any two words u and v, denoted by $d(u, v)$, is simply the number of coordinates in which the two words differ. The *weight* of u, $wt(u)$, is just the number of nonzero coordinates in u. Using the binary sum (so $1 + 1 = 0$), we have $d(u, v) = wt(u + v)$.

Under the assumptions that bits are more likely to be transmitted correctly than incorrectly (a natural assumption) and that messages are equally likely to be transmitted (this condition can be substantially relaxed), it is easy to show that for any received word w, the most likely codeword originally transmitted is the codeword $c \in C$ for which $d(c, w)$ is least; that is, the most likely codeword sent is the one closest to the received word. This leads to the definition of the *minimum distance* $d(C) = \min_{c_1, c_2 \in C}\{d(c_1, c_2)\}$ of a code C, the distance between the two closest codewords in C. Clearly, if a word w is received such that $d(c, w) \leq t = \lfloor (d(C) - 1)/2 \rfloor$ for some $c \in C$, then the unique closest codeword to w is c; therefore, c is the most likely codeword sent and we *decode* w to c. (Notice that if c was the codeword that was originally transmitted, then at most t bits were altered in c during transmission to result in w being received.) So decoding each received word to the closest codeword (known as *maximum likelihood decoding*, or MLD) will always result in correct decoding provided at most t errors occur during transmission. Furthermore, if c_1 and c_2 are the two closest codewords [so $d(c_1, c_2) = d(C)$], then it is clearly possible to change $t + 1$ bits in c_1 so that the resulting word w satisfies $d(c_1, w) > d(c_2, w)$. Therefore, if these $t + 1$ bits are altered during the transmission of c_1, then, using MLD, we would incorrectly decode the received word w to c_2. Since MLD results in correct decoding for C no matter which codeword is transmitted and no matter which set of up to t bits are altered during transmission, and since this is not true if we replace t with $t + 1$, C is known as a *t-error-correcting code*, or as an $(n, |C|, d)$ code where $d = d(C) \geq 2t + 1$.

The *construction problem* is to find an $(n, |C|, d)$ code, C, such that the minimum distance d (and thus t) is large—this improves the error-correction ability and thus the reliability of transmission; and where $|C|$ is large, so the rate of transmission ($\log_2 |C|/n = k/n$) is closer to 1. Since messages are usually blocked up into k-bit words, one usually has $\log_2 |C| = k$. Clearly, these aims compete against each other. The more codewords one packs together in C, the harder it is to keep them far apart. In practice one needs to make decisions about the reliability of the channel and the need to get the message transmitted correctly and then weigh that against the cost of decreasing the rate of transmission (or increasing the amount of data to be stored). It

is possible to obtain bounds on one of the three parameters n, d, and $|C|$ in terms of the other two, and in some cases families of codes have been constructed which meet these bounds. Two of these families are discussed in this article: the *perfect codes* are codes that meet the Hamming Bound and are described in Section II; the *maximum distance separable* (or MDS) codes are codes that meet the Singleton Bound and are described in Section IV.

The second problem associated with error-correcting codes is the *encoding problem*. Each message m is to be assigned a unique codeword c, but this must be done efficiently. One class of codes that have an efficient encoding algorithm are *linear* codes; that is, the *binary sum* of any pair of codewords in the code C is also a codeword in C (here, *binary sum* means componentwise binary addition of the binary digits or the *exclusive or* of the two codewords). This means that C is a vector space and thus has a basis which we can use to form the rows of a *generating* matrix G. Then each message m is encoded to the codeword $c = mG$. One might also require that C have the additional property of being *cyclic*; that is, the cyclic shift $c' = x_n x_1 x_2 \ldots x_{n-1}$ of any codeword $c = x_1 x_2 \ldots x_n$ (where $x_i \in \{0, 1\}$) is also a codeword for every codeword c in C. If C is cyclic and linear, then encoding can easily be completed using a shift register design. The representation of a code is critical in decoding. For example, Hamming codes (see Section II) possess a cyclic representation but also have equivalent representations that are not cyclic.

The final main problem associated with error-correcting codes is the *decoding problem*. It is all well and good to know from the design of the code that all sets of up to t errors occurring during transmission result in a received word, w, that is closer to the codeword c that was sent than it is to any other codeword; but given w, how do you efficiently find c and recover m? Obviously, one could test w against each possible codeword and perhaps eventually decode which is closest, but some codes are very big. Not only that, but it can also be imperative that decoding be done extremely quickly, as the following example demonstrates.

The introduction of the compact disc (CD) by Phillips in 1979 revolutionized the recording industry. This may not have been possible without the heavy use of error-correcting codes in each CD. (Errors can occur, for example, from incorrect cutting of the CD.) Each codeword on each CD represents less than 0.0001 sec of music, is represented by a binary word of length 588 bits, and is initially selected from a Reed-Solomon code (see Section III) that contains 2^{192} codewords. Clearly, this is an application where all decoding must take place with no delay, as nobody will buy a CD that stops the music while the closest codeword is being found! It turns out that not only are the Reed-Solomon codes excellent in that they meet the Singleton Bound (see Section III), but they also have an extremely efficient decoding algorithm which finds the closest codeword without having to compare the received word to all 2^{192} codewords.

Again, the class of linear codes also has a relatively efficient decoding algorithm. Associated with each linear code C is the dual code $C^{\perp}$ consisting of all vectors (codewords) such that the dot product with any codeword in C is 0 (again using *xor* or *binary* arithmetic). This is useful because of the fact that if H is a generating matrix for $C^{\perp}$, then $Hw^T = 0$ if and only if w is a codeword. H is also known as the *parity check* matrix for C. The word $s = Hw^T$ is known as the *syndrome* of w. Syndromes are even more useful because it turns out that for each possible syndrome s, there exists a word e_s with the property that a closest codeword to *any* received word w with syndrome s is $w + e_s$. This observation is taken even further to obtain a very efficient decoding algorithm for the Reed-Solomon codes that can deal with the 2^{196} codewords in real time; it incorporates the fact that these codes are not only linear but also cyclic.

Another family of codes that NASA uses is the *convolutional codes*. Theoretically, these codes are infinite in length, so a completely different decoding algorithm is required in this case (see Section V).

In the following sections, we focus primarily on the construction of some of the best codes, putting aside discussion of the more technical problem of describing decoding algorithms for all except the convolutional codes in Section V. This allows the interaction between codes and designs to be highlighted.

B. Combinatorial Designs

Although (combinatorial) designs were studied earlier by such people as Euler, Steiner, Kirkman, it was Yates (1936) who gave the subject a shot in the arm in 1935 by pointing out their use in statistics in the design of experiments. In particular, he defined what has become known as an (n, k, λ) *balanced incomplete block design* (BIBD) to be a set V of n elements and a set B of subsets (called *blocks*) of V such that

1. Each block has size $k < n$.
2. Each pair of elements in V occur as a subset of exactly λ blocks in B.

Fisher and Yates (1938) went on to find a table of small designs, and Bose (1939) soon after began a systematic study of the existence of such designs. Bose made use of finite geometries and finite fields in many of his constructions.

A natural generalization of BIBD is to replace (2) with

2′. Each $(t + 1)$-element subset of V occurs as a subset of exactly λ blocks of B.

Such designs are known as $(t+1)$ *designs*, which can be briefly denoted by $S_\lambda(t+1, k, n)$; in the particular case when $\lambda = 1$, they are known as *Steiner* $(t+1)$ *designs* which are denoted by $S(t+1, k, n)$. By elementary counting techniques, one can show that if $s < t$, then an $S_\lambda(t+1, k, n)$ design is also an $S_\mu(s, k, u)$ design where $\mu = \lambda\binom{n-s}{t+1-s}/\binom{k-s}{t+1-s}$. Since μ must be an integer, this provides several necessary conditions for the existence of a $(t+1)$ design.

For many values of n, k, and t, an $S(t+1, k, n)$ design cannot exist. A *partial* $S(t+1, k, n)$ design then is a set of k subsets of an n set where any $(t+1)$ subset is contained in at most one block. This is equivalent to saying that any two k subsets intersect in at most t elements. Partial designs are also referred to as *packings*, and much research has focused on finding maximum packings for various parameters.

There are very few results proving the existence of s designs once $s \geq 3$. Hanani found exactly when there exists an $S_\lambda(3, 4, v)$ [also called Steiner Quadruple Systems; see Lindner and Rodger (1975) for a simple proof and Hartman and Phelps (1992) for a survey], and Teirlinck (1980) proved that there exists an $S_\lambda(s, s+1, v)$ whenever $\lambda = ((s+1)!)^{2s+1}$, $v \geq s+1$, and $v \equiv s$ (modulo $((s+1)!)^{2s+1}$. Otherwise, just a few s designs are known [see Colbourn and Dinitz (1996)]. Much is known about their existence when $s = 2$. In particular, over 1000 papers have been written [see Colbourn and Rosa (1999)] about $S_\lambda(2, 3, v)$ designs (known as *triple systems*, and as *Steiner triple systems* if $\lambda = 1$). We only need to consider designs with $\lambda = 1$ in this article.

Certainly, $(t+1)$ designs and maximum packings are of interest in their own right, but they also play a role in the construction of good codes. To see how, suppose we have a $(t+1)$ design (or packing) (V, B). For each block b in B, we form its *characteristic vector* c_b of length n, indexed by the elements in V, by placing a 1 in position i if $i \in B$ and placing a 0 in position i if $i \notin B$. Let C be the code $\{c_b \mid b \in B\}$. Then C is a code of length n in which all codewords have exactly k 1's (we say each codeword has *weight* k). The fact that (V, B) is a $t+1$ design (or packing) also says something about the minimum distance of C: since each pair of blocks intersect in at most t elements, each pair of codewords have at most t positions where both are 1, so each pair of codewords disagree in at least $2k - 2t$ positions, so $d(C) \geq 2k - 2t$. This connection is considered in some detail in Section III. We also show in Section II that the codewords of weight d in perfect codes together form the characteristic vectors of a $(t+1)$ design.

There is much literature on the topic of combinatorial designs [see Colbourn and Dinitz (1996) for an encyclopedia of designs and Lindner and Rodger (1997) for an introductory text], and this topic is also considered elsewhere in this encyclopedia, so here we restrict our attention to designs that arise in connection with codes.

II. PERFECT CODES

Let C be a code of length n with minimum distance d. Let $t = \lfloor (d-1)/2 \rfloor$. Then, as described in Section I, for each codeword c in C, and for each binary word w of length n with $d(w, c) \leq t$, the *unique* closest codeword to w is c. Since we can choose any i of the n positions to change in c in order to form a word of length n distance exactly i from c, the number of words distance i from c is $\binom{n}{i} = n!/(n-i)!i!$. So the total number of words of length n that are distance at most t from c is $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{t}$, one of which is c, thus the number of words of length n distance at most t from some codeword in C is $|C| \sum_{i=0}^{t} \binom{n}{i}$ (by the definition of t, no codeword is within distance t of two codewords). Of course, the total number of binary words of length n is 2^n. Therefore, it must be the case that

$$|C| \leq 2^n \Big/ \sum_{i=0}^{t} \binom{n}{i}.$$

This bound is known as the *Hamming Bound* or the *sphere packing bound*. Any code that satisfies equality in the Hamming Bound is known as a *perfect code*, in which case $d = 2t + 1$.

From the argument above, it is clear that for any perfect code, each word of length n must be within distance t of a unique codeword (if a code is not perfect, then there exist words for which the distance to any closest codeword is more than t). In particular, if C is a perfect code with $d = 2t + 1$, then the codewords of minimum weight d in C are the characteristic vectors of the blocks of an $S(t+1, 2t+1, n)$ design. To see this, note that each word w of weight $t+1$ is within distance t of a unique codeword c, where clearly c must have weight $d = 2t + 1$. Equivalently, each $(t+1)$ subset is contained in a unique d subset, the characteristic vector of which is a codeword. In fact, for any codeword $c \in C$, one can define the neighborhood packing,

$$NS(c) = \{x + c \mid x \in C \text{ and } d(x, c) = d\}.$$

Then, the code C will be perfect *if and only if* every neighborhood packing is, in fact, the characteristic vectors of an $S(t+1, 2t+1, n)$ design. To see the converse, suppose C is a code with every $NS(c)$ an $S(t+1, 2t+1, n)$ design. If w is any word, let $c \in C$ be the closest codeword and assume $d(c, w) \geq t+1$. Choose any $t+1$ coordinates where c and w disagree. Since $NS(c)$ is an $S(t+1, 2t+1, n)$ design, these coordinates uniquely determine a block of size

$2t+1$ in the design and, hence, a codeword c' such that $d(c,c')=2t+1$. So c' disagrees with c in the same $t+1$ coordinates as does w, and thus, c' agrees with w in these $t+1$ coordinates. Thus,

$$\begin{aligned} d(c',w) &\le d(c,c') - (t+1) + d(c,w) - (t+1) \\ &\le d(c,w) - 1, \end{aligned}$$

which contradicts the assumption that c was the closest codeword and thus $d(c,w) \le t$.

It turns out that perfect binary codes are quite rare. The Hamming codes are an infinite family of perfect codes of length 2^r-1, for any $r \ge 2$, in which the distance is 3 and the number of codewords is 2^{2^r-1-r}. A linear Hamming code can always be formed by defining its parity check matrix, H, to be the $r \times n$ matrix in which the columns form all nonzero binary words of length r. Notice also that, in view of the comments earlier in this section, the codewords of weight 3 in any Hamming code form a Steiner triple system, $S(2,3,n)$.

The other perfect binary code is the Golay code, which has length 23, distance 7, and 2^{12} codewords.

Tietäväinen (1973), based on work by van Lint, showed that these are the only perfect binary codes and, in fact, generalized this result to codes over finite fields [see also Zinov'ev and Leont'ev (1973)]. For a survey of results on perfect codes, see van Lint (1975).

III. CONSTANT WEIGHT CODES

A *constant weight code* (CW) with parameters n, d, w is a set C of binary words of length n all having weight w such that the distance between any two codewords is at least d. All nontrivial (n,d,w) CW codes have $d \le 2w$. Let $A(n,d,w)$ be the largest number of codewords in any CW code with these parameters. The classic problem then is to detemine this number or find the best upper and lower bounds on $A(n,d,w)$.

Binary CW codes have found application in synchronization problems, in areas such as optical codedivision multiple-acces (CDMA) communications systems, frequency-hopping spread-spectrum communications, modile radio, radar and sonar signal design, and the construction of protocol sequences for multiuser collision channel without feedback. Constant weight codes over other alphabets have received some attention, but so far there have been few applications. We will only discuss the binary CW codes.

Constant weight codes have been extensively studied, and a good reference is MacWilliams and Sloane (1977). Eric Raines and Neil Sloane maintain a table of the best known lower bounds on $A(n,d,w)$ on the web site: http://www.research.att.com/njas/codes/Andw/. We will present an overview of this topic with an emphasis on the connections with designs.

Since the sum of any two binary words of the same weight always has even weight, we have $A(n,2\delta-1,w) = A(n,2\delta,w)$. We will assume from now on that the distance d is even. We also have $A(n,d,w) = A(n,d,n-w)$, since whenever two words are distance d apart, so are their complements. This means one only needs to consider the case $w \le n/2$.

The connection between CW codes and designs is immediate. In terms of sets, a CW code is just a collection of w subsets of an n set where the intersection of any two w subsets contains at most $t = w - \frac{d}{2}$ elements. Equivalently, a CW code is a partial $S(w-\frac{d}{2}+1, w, n)$ Steiner system. We then have

$$A(n,d,w) \le \frac{n(n-1)\cdots(n-w+d/2)}{w(w-1)\cdots(d/2)}$$

with equality if and only if a Steiner system $S(w-\frac{d}{2}+1, w, n)$ exists.

The interest in CW codes also comes from the problem of finding linear (or nonlinear) codes (n, M, d) of maximum size M. Obviously, $A(n,d,w)$ is an upper bound on the number of words of a given weight in such a maximum code. Conversely, such codes (or their cosets) can give lower bounds for $A(n,d,w)$. In particular, the stronger version of the Hamming Bound (given in the section on perfect codes) was originally proved using $A(n, 2t+2, 2t+1)$.

$A(n,2t+2,2t+1)$ is just the number of blocks in a maximum partial $S(t+1, 2t+1, n)$ design or packing. If C is a t-error-correcting code, then for any $c \in C$, the number of blocks in a neighborhood packing $|NS(c)| \le A(n,2t+2,2t+1)$. The number of words that are distance $t+1$ from c but not distance t from any other codeword is

$$\begin{aligned} \binom{n}{t+1} - \binom{2t+1}{t+1}|NS(c)| &\ge \binom{n}{t+1} \\ &- \binom{2t+1}{t+1} A(n,2t+2,2t+1). \end{aligned}$$

Each such word is distance $t+1$ from at most $\lfloor n/t+1 \rfloor$ other codewords. Thus, summing over all $c \in C$, each such word is counted at most this many times. This gives a stronger version of the Hamming bound:

$$|C|\left(\left(\sum_{i=0}^{t}\binom{n}{i}\right) + \frac{\binom{n}{t+1} - \binom{2t+1}{t+1}A(n,2t+2,2t+1)}{\lfloor n/(t+1) \rfloor}\right) \le 2^n.$$

Constant weight codes cannot be linear, since this would mean the zero vector was in the code, but one can have

a code with all nonzero words having the same weight. These codes are sometime referred to as linear *equidistant codes*. The dual of the Hamming code (also called the *simplex code*) is an example of such a code. In fact, it has been proved that the only such codes are formed by taking several copies of a simplex code. The proofs that all such codes are generalized simplex codes come explicitly from coding theory (Bonisoli, 1983) and also implicitly from results on designs and set systems (Teirlinck, 1980). There is a close connection between linear equidistant codes and finite geometries. The words of a simplex code correspond to the hyperplanes of projective space [over $GF(2)$] just as the words of weight 3 in the Hamming code correspond to lines in this projective space. [For connections between codes and finite geometries, see Black and Mullin (1976).]

Another variation on CW codes are optical orthogonal codes (OOC) which were motivated by an application to optical CDMA communication systems. Briefly, an (n, w, t_a, t_b) OOC is a CW code, C, of length n and weight w such that for any $c = (c_0, c_1, \ldots, c_{n-1}) \in C$, and each $y \in C$, $c \neq y$ and each $i \not\equiv 0 \pmod n$,

$$\sum_{j=0}^{n-1} c_j c_{j+i} \leq t_a, \tag{1}$$

and

$$\sum_{j=0}^{n-1} c_j y_{j+i} \leq t_c. \tag{2}$$

Equation (1) is the autocorrelation property, and Eq. (2) is the cross-correlation property. Most research has focused on the case where $t_a = t_c = t$, in which case we refer to an (n, w, t) OOC. Again, one can refomulate these properties in terms of (partial) designs or packings. In this case, an OOC is a collection of w subsets of the integers $(\bmod n)$, such that for subsets $c, b \in C$,

$$|(c + i) \cap (c + j)| \leq t_a \ (i \neq j), \tag{3}$$

and

$$|(c + i) \cap (b + j)| \leq t_c. \tag{4}$$

Here, $c + i = \{x + i (\bmod n) \mid x \in c\}$.

An OOC code is equivalent to a cyclic design or packing. A code or packing is said to be cyclic if every cyclic shift of a codeword (or block) is another codeword. The set of all cyclic shifts of a codeword is said to be an *orbit*. A representative from that orbit is often called a base block. An (n, w, t) OOC is a set of base blocks for a cyclic (partial) $S(t + 1, w, n)$ design or packing (assuming $t < w$). Conversely, given such a cyclic partial $S(t + 1, w, n)$ design or packing, one can form an (n, w, t) OOC by taking one representative block or codeword from each orbit.

IV. MAXIMUM DISTANCE SEPARABLE CODES

For any linear code C, recall that the minimum distance equals the minimum weight of any nonzero codeword. Also, if C has dimension k, then $C^\perp$ has dimension $n - k$ and any parity check matrix H of C has rank $n - k$. If $c \in C$ is a codeword of minimum weight, $wt(c) = d$, then $Hc^\top = 0$ implies that d columns of H are dependent, but no $d - 1$ columns are dependent. Since H has rank $n - k$, every $n - k + 1$ columns of H are dependent. Thus,

$$d \leq n - k + 1.$$

This is known as the *Singleton Bound*, and any code meeting equality in this bound is known as a *maximum distance separable* code (or MDS code).

There are no interesting binary MDS codes, but there are such codes over other alphabets, for example, the Reed-Solomon codes used in CD encoding of order 256 (Reed-Solomon codes are described below). Even though such codes are treated mathematically as codes with 256 different "digits," each still has an implementation as a binary code, since each of the digits in the finite field $GF(2^8)$ can be represented by a binary word of length 8; that is, by one byte. So the first step in encoding the binary string representing all the music onto a CD is to divide it into bytes and to regard each such byte, as a field element in $GF(2^8)$.

We now consider a code $C \subseteq F^n$ as a set of codewords over the alphabet F, where F is typically the elements of a finite field. The are several equivalent definitions for a linear code C of length n and dimension k to be an MDS code:

1. C has minimum distance $d = n - k + 1$,
2. Every k column of G, the generating matrix for C, is linearly independent.
3. Every $n - k$ column of H, the parity check matrix for C, is linearly independent.

Note, that from item (3) above C is MDS if and only if $C^\perp$ is MDS.

If one arranges the codewords of C in an $|C| \times n$ array, then from item (2) this array will have the property that for any choice of k columns (or coordinates) and any word of length k, $w \in F^k$, there will be exactly one row of this array that has w in these coordinates. An *orthogonal array* is defined to be a $q^k \times n$ array with entries from a set F, $|F| = q$, with precisely this property: restricting the array to any k columns, every word $w \in F^k$ occurs exactly once in this $q^k \times k$ subarray. Two rows of an orthogonal array can agree in at most $k - 1$ coordinates, which means that they must disagree in at least $n - (k - 1)$ coordinates. Thus, the distance between any two rows of an orthogonal

array is $d = n - k + 1$. Obviously, the row vectors of an orthogonal array are also codewords of an MDS code, except that orthogonal arrays (and MDS codes) do not need to be linear and exist over arbitrary alphabets.

Orthogonal arrays were also introduced in the design of experiments in statistical studies and are closely related to designs and finite geometries. In fact, the construction for Reed-Solomon codes was first published by Bush as a construction of orthogonal arrays (see Bush, 1952; Colbourn and Dinitz, 1996).

There are various representations of the Reed-Solomon codes. We present the most perspicuous one. Again, let F denote a finite field and let $F_k[x]$ denote the space of all polynomials of degree less than k with coefficients from F. Choose $n > k$ different (nonzero) field elements $\alpha_1, \alpha_2, \ldots, \alpha_n \in F$. For each polynomial $f(x) \in F_k[x]$, form the *valuation vector* $c_f = (f(\alpha_1), f(\alpha_2), \ldots, f(\alpha_n))$. Define

$$C = \{c_f \mid f(x) \in F_k[x]\}.$$

First, we note that C is a linear code, since $c_f + c_g = c_{f+g}$. Second, for any two different polynomials $f(x), g(x) \in F_k[x]$, we have $c_f \neq c_g$; if c_f and c_g were equal, then the polynomial $f(x) - g(x)$ would have n roots but degree $< k (< n)$. This means that $|F_k[x]| = |F^k| = q^k = |C|$, and thus, C has dimension k and length n. Finally, since $f(\alpha_i) = 0$ if and only if α_i is a root of $f(x)$, and moreover any polynomial of degree $\leq k - 1$ has at most $k - 1$ roots, then c_f has at most $k - 1$ zeros and at least $n - k + 1$ nonzero entries. Therefore, the minimum distance for C is $n - k + 1$ and C is an MDS code.

Reed-Solomon codes also have a representation as a cyclic code and a relatively efficient decoding algorithm (see Hoffman *et al.*, 1991, for example).

V. CONVOLUTIONAL CODES

Convolutional codes are practical codes, having been adopted for use by both NASA and the European Space Agency. In fact, they encode messages twice: first using a Reed-Solomon code, then the resulting codeword is encoded using a convolutional code.

Convolutional codes are infinite length codes that are both linear and cyclic. The messages to be considered are strung together into a stream of bits which form a single message m that is encoded by feeding m into a shift register (see Fig. 1). Initially, μ codewords are formed: for $1 \leq i \leq \mu$, and for each *tick* $t \geq 0$, the contents of certain registers are added together to form the tth bit in the output of the codeword $c_i = c_{i,0}, c_{i,1}, c_{i,2}, \ldots$. The single codeword c to which m is encoded is then formed by $c = c_{1,0}c_{2,0}, \ldots, c_{\mu,0}c_{1,1}c_{2,1}, \ldots, c_{\mu,1}, \ldots$.

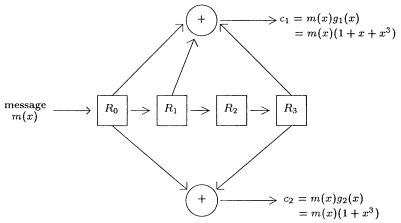


FIGURE 1 Convolutional code encoding.

Algebraically, one can represent the registers whose contents are added to form c_i by the polynomial $g_i = g_{i,0}x^0 + g_{i,1}x + \ldots g_{i,\ell}x^\ell$, where $g_{i,j} = 1$ if the contents in register R_j are used when forming c_i, and is 0 otherwise. Then by writing the message $m = m_0 m_1 m_2, \ldots$ in polynomial form $m(x) = m_0 + m_1 x + m_2 x^2 + \cdots$, it turns out that $c_i(x) = m(x)g_i(x)$. This representation of convolutional code encoding makes it clear why the code is cyclic and linear.

Convolutional codes can also be represented graphically. Assuming one message bit is moved into the shift register at each tick, the vertices of the directed graph are the 2^ℓ binary words $(r_0, r_1, \ldots, r_{\ell-1})$ that can make up the contents of the first ℓ registers $R_0, R_1, \ldots, R_{\ell-1}$, and there is a directed edge from $r = (r_0, r_1, \ldots, r_{\ell-1})$ to $s = (s_0, s_1, \ldots, s_{\ell-1})$ if and only if $s_i = r_{i-1}$ for $1 \leq i \leq \ell - 1$. So there is a directed edge from r to s if in one tick the contents of the first ℓ registers can change from r into s. Furthermore, the directed edge from $(m_{t+\ell}, m_{t+\ell-1}, \ldots, m_t)$ to $(m_{t+\ell+1}, m_{t+\ell}, \ldots, m_{t+1})$ is labeled with the tth bits of $c_1, c_2, \ldots, c_\mu$; so this label is the contribution to c made at the tth tick. This directed graph is essentially the transition diagram of the finite state machine formed by the shift register.

Representing a convolutional code with this directed graph D is helpful in understanding both encoding and decoding. Codewords simply correspond to walks in D with the codeword being the concatenation of the labels on the edges in the walk. If message bits are moved into the shift register one at a time, then there are exactly two arcs directed out of each vertex (corresponding to a message bit of a 0 or a 1 being moved into R_0), so the walk can head off in one of two possible directions. It is this observation that makes it clear what is involved in decoding. At each tick, decoding one message bit requires deciding which of the two directions to take from the current vertex (state). This decision cannot be based on knowing the entire received word, since it has arbitrarily long length. Instead, one gathers the next τ ticks worth of the received word, which together form, say, w', then finds which walk W emanating from the current state most closely matches w', then finally makes a decoding decision to take one step along W. This process can

be efficiently implemented by using the Viterbi Decoder. Deciding on how large to make τ can affect the codeword to which w is decoded (see Hoffman *et al.*, 1991, for example).

In deciding which convolutional code to use, choices have to made about $g_1(x), \ldots, g_\mu(x)$ and the number k of message symbols to move into the shift register at each tick, usually chosen to be 1. The rate of the code is then k/μ.

SEE ALSO THE FOLLOWING ARTICLES

COMMUNICATION SATELLITE SYSTEMS • DATABASES • DISCRETE MATHEMATICS AND COMBINATORICS • WIRELESS COMMUNICATIONS

BIBLIOGRAPHY

Beth, Th., Jungnickel, D., and Lenz, H. (1999, 2000). "Design Theory," Vols. 1 and 2, Cambridge Univ. Press, Cambridge, UK.

Blake, I. F., and Mullin, R. C. (1976). "The Mathematical Theory of Coding Theory," Academic Press, New York.

Bonisoli, A. (1983) "Every equidistant linear code is a sequence of dual Hamming codes," *Ars Combinatoria* **18,** 181–186.

Bose, R. C. (1939) "On the construction of balanced incomplete block designs," *Ann. Eugen.* **9,** 353–399.

Bush, K. A. (1952) "Orthogonal arrays of index unity," *Ann. Math. Stat.* **23,** 426–434.

Colbourn C. J., and Dinitz, J. H., eds. (1996). "The CRC Handbook of Combinatorial Designs," CRC Press, Boca Raton, FL.

Colbourn, C. J., and Rosa, A. (1999) "Triple Systems," Oxford Univ. Pess, Oxford.

Fisher, R. A., and Yates, F. (1938). "Statistical Tables for Biological, Agricultural and Medical Research," Oliver & Boyd, Edinburgh.

Golay, M. J. E. (1949). "Notes on digital coding," *Proc. IEEE* **37,** 657.

Hamming, R. S. (1950) "Error-detecting and error-correcting codes," *Bell Syst. Tech. J.* **29,** 147–160.

Hanani, H. (1960) "On quadruple systems," *Canad. J. Math.*, **12,** 145–157.

Hartman, A., and Phelps, K. T. (1992) "Steiner Quadruple Systems, Contemporary Design Theory" (J. H. Dinitz and D. R. Stinson, eds.), Wiley, New York.

Hoffman, D. G., Leonard, D. A., Lindner, C. C., Phelps, K. T., Rodger, C. A., and Wall, J. R. (1991). "Coding Theory: The Essentials," Dekker, New York.

Lindner, C. C., and Rodger, C. A. (1997) "Design Theory," CRC Press, Boca Raton, FL.

van Lint, J. H. (1975) "A survey of perfect codes," *Rocky Mount. J. Math.* **5,** 199–224.

MacWilliams, F. J., and Sloane, N. J. A. (1977). "The Theory of Error-Correcting Codes," North-Holland, Amsterdam.

Shannon, C. E. (1948) "A mathematical theory of communication," *Bell Syst. Tech. J.* **27,** 379–423 and 623–656.

Teirlinck, L. (1980). "On projective and affine hyperplanes," *J. Combinatorial Theory, Ser. A* **28,** 290–306.

Tietäväinen, A. (1973) "On the nonexistence of perfect codes over finite fields," *SIAM J. Appl. Math.* **24,** 88–96.

Yates, F. (1939) "Complex experiments," *J. R. Stat. Soc.* **2,** 181–247.

Yates, F. (1936) "Incomplete randomized blocks," *Ann. Eugen.* **7,** 121–140.

Zinov'ev, V. A., and Leont'ev, V. K. (1973). "The nonexistence of perfect codes over Galois fields," *Probl. Control Inf. Theory* **2**(2), 123–132.



Differential Equations, Ordinary

Anthony N. Michel
University of Notre Dame

I. Introduction
II. Initial-Value Problems
III. Fundamental Theory
IV. Linear Systems
V. Stability

GLOSSARY

Equilibrium or rest position For the system of equations $x' = f(t, x)$, any point x_e such that $f(t, x_e) = 0$ for all t is an equilibrium point or a rest position.

Fundamental matrix If $\{\phi_1, \ldots, \phi_n\}$ denotes a set of linearly independent solutions for the equation $x' = A(t)x$, then the matrix $\Phi = [\phi_1, \ldots, \phi_n]$ is a fundamental matrix of $x' = A(t)x$.

Initial-value problem The system of ordinary differential equations $x' = f(t, x)$ along with the initial data $x(\tau) = \xi$ is an initial value problem, where τ denotes initial time and ξ denotes the initial condition or the initial state.

***n*th-Order ordinary differential equations** Equation of the form $y^{(n)} = h(t, y^{(1)}, \ldots, y^{(n-1)})$ is an nth-order ordinary differential equation, where $y^{(i)}$ denotes the ith derivative of y with respect to t.

Qualitative theory of ordinary differential equations Study of families of solutions of ordinary differential equations, such as, for example, the (stability) properties of solutions near an equilibrium point.

Solution of an initial-value problem n-Vector-valued function ϕ is a solution of an initial-value problem if ϕ satisfies the equation $x' = f(t, x)$ andif $\phi(\tau) = \xi$.

State transition matrix For the initial value problem $x' = A(t)x$, $x(\tau) = \xi$, the matrix given by $\Phi(t, \tau) \triangleq \Phi(t)\Phi^{-1}(\tau)$ is the state transition matrix, where Φ denotes a fundamental matrix for $x' = A(t)x$ and Φ^{-1} denotes the inverse of Φ; the unique solution of the initial value problem is then $\phi(t, \tau, \xi) = \Phi(t, \tau)\xi$.

System of linear homogeneous differential equations System of equations given by $x' = A(t)x$ where x is an n vector and $A(t)$ denotes an $n \times n$ (time-varying) matrix is a linear homogeneous system of ordinary differential equations.

System of ordinary differential equations The system of equations given by $x' = f(t, x)$ where x is an n vector, t is real (time), f is a function, and x' denotes differentiation of x with respect to t is a system of n ordinary differential equations of the first order.

EQUATIONS containing the derivatives or differentials of one or more dependent variables, with respect to one or more independent variables, are called *differential equations*. If such equations contain only ordinary derivatives of one or more dependent variables, with respect to a single independent variable, then one speaks of *ordinary differential equations*. Equations involving the partial

derivatives of one or more dependent variables of two or more independent variables are called *partial differential equations*.

I. INTRODUCTION

In what follows, we will concern ourselves only with ordinary differential equations. The study of such equations may be divided into qualitative theory and quantitative theory (e.g., the numerical solution of differential equations). We will concern ourselves almost exclusively only with a qualitative theory for such equations.

Ordinary differential equations, which were first considdered in the seventeenth century by Leibnitz and Newton, arise in nearly all disciplines of science (physics, chemistry, biology, and the like) and engineering (aerospace, chemical, civil, electrical, mechanical, nuclear, and so forth) as well as in economics and societal systems. It is not an overstatement to say that a very great deal of applied mathematics involves in some way the study of differential equations.

The study of ordinary differential equations can be from fairly elementary and down-to-earth to rather advanced and abstract levels. In the current treatment we will follow a path between these two extremes. The study of ordinary differential equations at a very basic level requires as a prerequisite some knowledge of elementary calculus. At an intermediate level, the study of such equations demands some background in real variables and linear algebra, while at the advanced level, the study of differential equations may involve facts from measure and integration theory as well as functional analysis.

II. INITIAL-VALUE PROBLEMS

A. First-Order Ordinary Differential Equations

We let D denote a domain (i.e., an open, non-empty, and connected set) in the R^{n+1} space. We call R^{n+1} the (t, x) space, and we denote elements of R^{n+1} by $(t, x_1, \ldots, x_n) = (t, x)$ and elements of R^n by $(x_1, \ldots, x_n) = x$. Next we consider n functions $f_i, i = 1, \ldots, n$, which map D into the real numbers R. To express this, we write $f_i : D \to R$. We assume that each f_i is continuous at all points in D and we express this by writing $f_i \in C(D)$. Finally, we let $x_i^{(n)}$ denote the nth derivative of x_i with respect to t (provided that it exists) (i.e., $d^n x_i/dt^n = x_i^{(n)}$). In particular, when $n = 1$, we frequently write

$$x_i^{(1)} \triangleq x_i' = dx_i/dt$$

We call the system of equations given by:

$$x_i' = f_i(t, x_1, \ldots, x_n), \qquad i = 1, \ldots, n \qquad (\mathrm{E}_i)$$

a **system of n ordinary differential equations of the first order.** By a **solution** of the system of ordinary differential equations (E_i) we shall mean n continuously differentiable functions $\phi_1, \ldots, \phi_n$ defined on an interval $J = (a, b)$ (recall that (a, b) is the set of all t in R with the property $a < t < b$) such that $(t, \phi_1(t), \ldots, \phi_n(t)) \in D$ for all $t \in J$ and such that

$$\phi_i'(t) = f_i(t, \phi_1(t), \ldots, \phi_n(t))$$
$$i = 1, \ldots, n$$

for all $t \in J$.

Next, we let $(\tau, \xi_1, \ldots, \xi_n) \in D$. Then the **initial-value problem** associated with (E_i) is given by:

$$\begin{aligned} x_i' &= f_i(t, x_1, \ldots, x_n), && i = 1, \ldots, n \\ x_i(\tau) &= \xi_i, && i = 1, \ldots, n \end{aligned} \qquad (\mathrm{I}_i)$$

A set of functions $(\phi_1, \ldots, \phi_n)$ is a **solution** of (I_i) if $(\phi_1, \ldots, \phi_n)$ is a solution of (E_i) on some interval J containing τ and if $(\phi_1(\tau), \ldots, \phi_n(\tau)) = (\xi_1 \ldots, \xi_n)$.

In Fig. 1, the solution of a hypothetical initial-value problem is given when $n = 1$. Note that at (τ, ξ), $\phi'(\tau) = f(\tau, \phi(\tau)) = m$ is the slope of line L in the figure.

In dealing with systems of equations, we find it convenient to use vector notation, such as:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \xi = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix}, \qquad \phi = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_n \end{bmatrix}$$

$$f(t, x) = \begin{bmatrix} f_1(t, x_1, \ldots, x_n) \\ \vdots \\ f_n(t, x_1, \ldots, x_n) \end{bmatrix}$$

$$= \begin{bmatrix} f_1(t, x) \\ \vdots \\ f_n(t, x) \end{bmatrix}$$

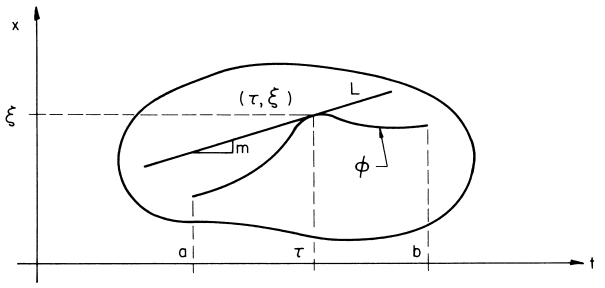


FIGURE 1 Solution of an initial-value problem; t interval $J = (a, b)$, m(slope of line L) $= f(\tau, \phi(\tau))$.

$$x' = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}, \qquad \begin{bmatrix} \int_\tau^t f_1(s, \phi(s))\, ds \\ \vdots \\ \int_\tau^t f_n(s, \phi(s))\, ds \end{bmatrix}$$

$$= \int_\tau^t f(s, \phi(s))\, ds$$

With this notation we can express the initial-value problem (I_i) by:

$$x' = f(t, x), \qquad x(\tau) = \xi \tag{I}$$

It is an easy matter to verify that the initial-value problem (I) can equivalently be expressed by an integral equation of the form:

$$\phi(t) = \xi + \int_\tau^t f(s, \phi(s))\, ds \tag{V}$$

where ϕ denotes the solution of (I).

B. Classification of Systems of First-Order Differential Equations

We are now ready to classify systems of first-order differential equations in a variety of ways:

1. If in (I), $f(t, x) \equiv f(x)$ for all (t, x) in D, then

$$x' = f(x) \tag{A}$$

and we call (A) an **autonomous system** of first-order ordinary differential equations.

2. If $(t + T, x) \in D$ when $(t, x) \in D$ and if $f(t, x) = f(t + T, x)$ for all $(t, x) \in D$ then (I) assumes the form

$$x' = f(t, x) = f(t + T, x) \tag{P}$$

Such a system is called a **periodic system** of first-order differential equations of period T. The smallest $T > 0$ for which (P) is true is the **least period** of this system of equations.

3. When in (I), $f(t, x) = A(t)x$, where $A(t) = [a_{ij}(t)]$ is a real $n \times n$ matrix with elements $a_{ij}(t)$ which are defined and at least piecewise continuous on a t interval J, then we have

$$x' = A(t)x \tag{LH}$$

and we speak of a **linear homogeneous system** of ordinary differential equations.

4. If for (LH) $A(t)$ is defined for all real t and if there is a $T > 0$ such that $A(t) = A(t + T)$ for all t, then we have

$$x' = A(t)x = A(t + T)x \tag{LP}$$

This system is called a **linear periodic system** of ordinary differential equations.

5. If in (I), $f(t, x) = A(t)x + g(t)$, where $g(t)^{\mathrm{T}} = [g_1(t), \ldots, g_n(t)]$ and where $g_i: J \to R$, then we have

$$x' = A(t)x + g(t) \tag{LN}$$

In this case we speak of a **linear nonhomogeneous system of ordinary differential equations.**

6. If in (I), $f(t, x) = Ax$, where $A = [a_{ij}]$ is a real $n \times n$ matrix with constant coefficients, then we have

$$x' = Ax \tag{L}$$

This type of system is called a **linear, autonomous, homogeneous** system of ordinary differential equations.

C. *n*th-Order Ordinary Differential Equations

Thus far we have concerned ourselves with systems of first-order ordinary differential equations. It is also possible to characterize initial value problems by means of nth-order ordinary differential equations. To this end, we let h be a real function which is defined and continuous on a domain D of the real $(t, y_1, \ldots, y_n)$ space. Then

$$y^{(n)} = h\big(t, y^{(1)}, \ldots, y^{(n-1)}\big) \tag{E_n}$$

is an nth-**order ordinary differential equation.** A **solution** of (E_n) is a real function ϕ which is defined on a t interval $J = (a, b)$ which has n continuous derivatives on J and satisfies $(t, \phi(t), \ldots, \phi^{(n-1)}(t)) \in D$ for all $t \in J$ and

$$\phi^{(n)}(t) = h\big(t, \phi(t), \ldots, \phi^{(n-1)}(t)\big)$$

for all $t \in J$.

Now for a given $(\tau, \xi_1, \ldots, \xi_n) \in D$, the initial value problem for (E_n) *is*

$$\begin{aligned} y^{(n)} &= h\big(t, y, y^{(1)}, \ldots, y^{(n-1)}\big) \\ y^{(\tau)} &= \xi_1, \ldots, y^{(n-1)}(\tau) = \xi_n \end{aligned} \tag{I_n}$$

A function ϕ is a solution of (I_n) if ϕ is a solution of Eq. (E_n) on some interval containing τ and if $\phi(\tau) = \xi_1, \ldots, \phi^{(n-1)}(\tau) = \xi_n$.

As in the case of systems of first-order equations, we single out several special cases.

1. Consider equations of the form

$$y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y^{(1)} + a_0(t)y = g(t) \tag{1}$$

where $a_{n-1}(t), \ldots, a_0(t)$ are real continuous functions defined on the interval J. We refer to Eq. (1) as a **linear homogeneous ordinary differential equation of order** n.

2. If in (1) we let $g(t) \equiv 0$, then

$$y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y^{(1)} + a_0(t)y = 0 \tag{2}$$

We call Eq. (2) a **linear homogeneous ordinary differential equation of order** n.

3. If in (2) we have $a_i(t) \equiv a_i$, $i = 0, 1, \ldots, n-1$, then

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y^{(1)} + a_0 y = 0 \qquad (3)$$

and we call Eq. (3) a **linear, autonomous, homogeneous ordinary differential equation of order** n.

We can, of course, also define **periodic** and **linear periodic ordinary differential equations of order** n in the obvious way.

It turns out that the theory of nth-order ordinary differential equations reduces to the theory of a system of n first-order ordinary differential equations. To this end, we let $y = x_1$, $y^{(1)} = x_2, \ldots, y^{(n-1)} = x_n$ in Eq. (I_n). Then we have the system of first-order ordinary differential equations:

$$\begin{aligned} x_1' &= x_2 \\ x_2' &= x_3 \\ &\vdots \\ x_n' &= h(t, x_1, \ldots, x_n) \end{aligned} \qquad (4)$$

which is defined for all $(t, x_1, \ldots, x_n) \in D$. Assume that the vector $\phi = (\phi_1, \ldots, \phi_n)^{\mathrm{T}}$ is a solution of (4) on an interval J. Since $\phi_2 = \phi_1'$, $\phi_3 = \phi_2', \ldots, \phi_n = \phi_1^{(n-1)}$, and since

$$\begin{aligned} &h(t, \phi_1(t), \ldots, \phi_n(t)) \\ &\quad = h\big(t, \phi_1(t), \ldots, \phi_1^{(n-1)}(t)\big) = \phi_1^{(n)}(t) \end{aligned}$$

it follows that the first component ϕ_1 of the vector ϕ is a solution of Eq. (E_n) on the interval J. Conversely, if ϕ_1 is a solution of (E_n) on J, then the vector $(\phi, \phi^{(1)}, \ldots, \phi^{(n-1)})^{\mathrm{T}}$ is clearly a solution of Eq. (4). Moreover, if $\phi_1(\tau) = \xi_1, \ldots, \phi_1^{(n-1)}(\tau) = \xi_n$, then the vector ϕ satisfies $\phi(\tau) = \xi = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$. The converse is also true.

D. Examples of Initial-Value Problems

We conclude the present section with several representative examples.

1. Consider the second-order ordinary differential equation given by:

$$m\,d^2x/dt^2 + g(x) = 0 \qquad (5)$$

where $m > 0$ is a constant and $g: R \to R$ is continuous. This equation, along with $x(0) = \xi_1$ and $x'(0) = \xi_2$ specified, constitutes an initial-value problem. If we let $x_1 = x$ and $x_2 = x'$, then Eq. (5) can equivalently be represented by a system of two first-order ordinary differential equations given by:

$$x_1' = x_2, \qquad x_2' = -(1/m)g(x_1) \qquad (6)$$

with $x_1(0) = \xi_1$ and $x_2(0) = \xi_2$.

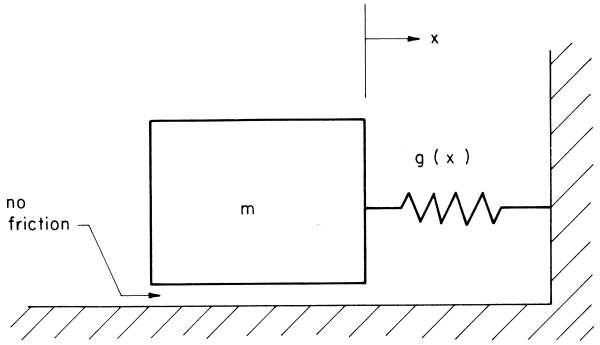


FIGURE 2 Mass–spring system.

Equation (5) can be used to describe a variety of physical phenomena. Consider, for example, the mass–spring system depicted in Fig. 2. When the system is in the rest position, then $x = 0$; otherwise, x is positive in the direction of the arrow or negative otherwise. The function $g(x)$ denotes the restoring force of the spring while the mass is expressed by m (in a consistent set of units).

There are several well-known special cases for Eq. (5):

(a) If $g(x) = kx$, where $k > 0$ is known as Hooke's constant, then Eq. (5) is a linear ordinary differential equation called the *harmonic oscillator*.

(b) If $g(x) = k(1 + a^2x^2)x$, where $k > 0$ and $a^2 > 0$ are parameters, then Eq. (5) is a nonlinear ordinary differential equation and one refers to the resulting system as a "mass and a hard spring."

(c) If $g(x) = k(1 - a^2x^2)x$, where $k > 0$ and $a^2 > 0$ are parameters, then Eq. (5) is a nonlinear ordinary differential equation and one refers to the resulting system as a "mass and a soft spring."

Alternatively, Eq. (5) can be used to describe the behavior of the pendulum shown in Fig. 3 with $\theta = x$. In this case, the restoring force $g(x)$ is specified by:

$$g(x) = m(g/l)\sin x$$

2. Using Kirchhoff's voltage and current laws, the circuit of Fig. 4 can be modeled by the linear system of equations:

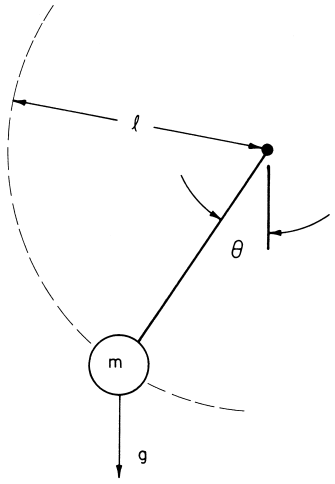


FIGURE 3 Simple pendulum.

$$
\begin{aligned}
v_1' &= -\frac{1}{C_1}\left(\frac{1}{R_1}+\frac{1}{R_2}\right)v_1 + \frac{1}{R_2C_2}v_2 + \frac{v}{R_1C_1} \\
v_2' &= -\frac{1}{C_1}\left(\frac{1}{R_1}+\frac{1}{R_2}\right)v_1 - \left(\frac{R_2}{L}-\frac{1}{R_2C_1}\right)v_2 \\
&\quad + \frac{R_2}{L}v_3 + \frac{v}{R_1C_1} \\
v_3' &= \frac{1}{R_2C_2}v_1 - \frac{1}{R_2C_2}v_2
\end{aligned}
\tag{7}
$$

In order to complete the description of this circuit, we need to specify the initial conditions $v_1(0)$, $v_2(0)$, and $v_3(0)$.

III. FUNDAMENTAL THEORY

In this section, we address the following questions:

1. Under what conditions has the initial-value problem (I) *at least one solution* for a given set of initial data (τ, ξ)?
2. Under what conditions has (I) *exactly one solution* for given (τ, ξ)?

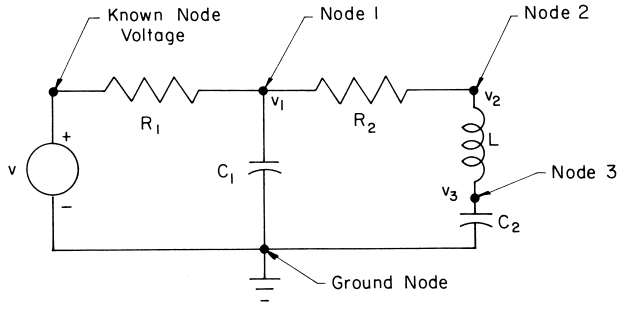


FIGURE 4 An example of an electric circuit.

3. What is the extent of the time interval over which one or more solutions exist for (I)?
4. How do solutions for (I) behave when the initial data (τ, ξ) (or some other parameters for the differential equation) are varied?

The significance of the preceding questions is brought further to light when the following examples are considered.

1. For the initial-value problem,

$$x' = -\operatorname{sgn} x, \qquad x(0) = 0, \quad t \geq 0 \tag{8}$$

where

$$\operatorname{sgn} x = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

no continuously differentiable function ϕ exists which satisfies Eq. (8). Hence, *no solution* (as defined in Section II) exists for the present initial-value problem.

2. The initial-value problem,

$$x' = 1/(2x), \qquad x(0) = 0, \quad t \geq 0 \tag{9}$$

has *two solutions* given by $\phi(t) = \pm t^{1/2}$ which exist for all $t \geq 0$.

3. The initial-value problem,

$$x' = 1 + x^2, \qquad x(0) = 0, \quad t \geq 0 \tag{10}$$

has the *unique solution* given by $\phi(t) = \tan t$. This solution exists only when $0 \leq t < \pi/2$, since ϕ is not continuously differentiable at $t = \pi/2$. In this case, we say that this solution has *finite escape time*.

4. The initial-value problem given by:

$$x' = ax, \qquad x(\tau) = \xi \tag{11}$$

where a is a fixed parameter, has a unique solution given by $\phi(t) = \phi(t, \tau, \xi) = \xi e^{a(t-\tau)}$ which exists for all real t. Note that the solution ϕ is continuous with respect to the parameters, a, τ, and ξ.

A. Existence of Solutions

In order to simplify our presentation, we will consider in the next few results one-dimensional initial value problems (i.e., we will assume that for (I), $n = 1$). Later in this section we will show how these results are modified for higher dimensional systems. Thus, we have a domain $D \in R^2$, we are given $(\tau, \xi) \in D$ and $f \in C(D)$, and *we seek a solution for the initial-value problem*,

$$x' = f(t, k), \qquad x(\tau) = \xi \tag{I'}$$

In doing so, it suffices to find a solution of the integral equation:

$$\phi(t) = \xi + \int_{\tau}^{t} f(s, \phi(s))\, ds \qquad (V')$$

One way of solving the above problem is by considering approximations to a solution first; an ε-**approximate solution** for (I′) on an interval J containing τ is a real-valued function ϕ which is piecewise continuously differentiable on J and satisfies $\phi(\tau) = \xi$, $(t, \phi(t)) \in D$ for all $t \in J$, and which satisfies:

$$|\phi'(t) - f(t, \phi(t))| < \varepsilon$$

at all points t of J where $\phi'(t)$ exists.

Let us now consider a subset S of D defined by:

$$S = \{(t, x) \in D: |t - \tau| \le a, |x - \xi| \le b\} \subset D$$

Since $f \in C(D)$, there is an $M \ge 0$ such that $|f(t, x)| \le M$ for all $(t, x) \in S$. Now define $c \triangleq \min\{a, b/M\}$ and, depending on the size of M relative to a, define one of the triangular regions shown either in Fig. 5a or 5b.

It is now not too difficult to prove the following results:

If $f \in C(D)$ and if c is as defined above, then for any $\varepsilon > 0$ there is an ε-approximate solution of (I′) *on the interval $|t - \tau| \le c$.*

Indeed, for a given $\varepsilon > 0$, such a solution will be of the form:

$$\begin{aligned} &\phi(t) = \phi(t_j) + f(t_j, \phi(t_j))(t - t_j) \\ &t_j \le t \le t_{j+1}, \qquad j = 0, 1, 2, \ldots, m-1 \end{aligned} \qquad (12)$$

when $t \ge \tau = t_0$ (the case in which $t \le \tau$ is modified in the obvious way). In Eq. (12), the choice of m and of $\max|t_{j+1} - t_j|$ will depend on ε but not on t, but in any

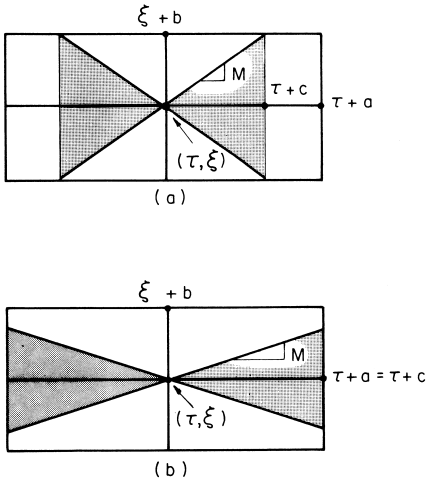


FIGURE 5 (a) Case $c = b/M$. (b) Case $c = a$.

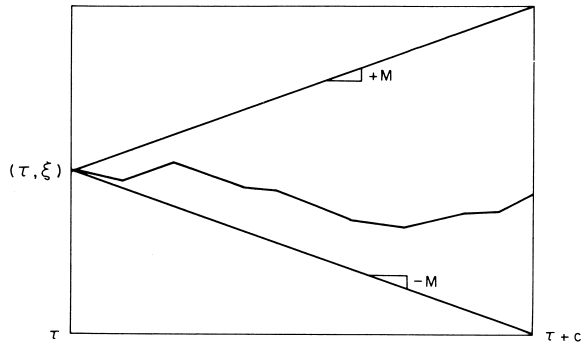


FIGURE 6 Typical ε-approximation solution.

case, we have $\sum_{j=0}^{m-1} |t_{j+1} - t_j| = c$. In Fig. 6, a typical ε-approximate solution is shown.

Next, let us consider a monotone decreasing sequence of real numbers with limit zero, and let us denote this sequence by $\{\varepsilon_m\}$. An example of such a sequence would be the case when $\varepsilon_m = 1/m$, where m denotes all positive integers greater or equal to one. Corresponding to each ε_m, let us consider now an ε_m-approximate solution which we denote by ϕ_m. Next, let us consider the family of ε_m-approximate solutions $\{\phi_m\}$, $m = 1, 2, \ldots$. This family $\{\phi_m\}$ is an example of an *equicontinuous* family of functions. Now, according to *Ascoli's lemma*, an equicontinuous family $\{\phi_m\}$, as constructed above, will contain a subsequence of functions, which we denote by $\{\phi_{m_k}\}$, which converges uniformly on the interval $[\tau - c, \tau + c]$ to a continuous function ϕ; that is,

$$\lim_{k \to \infty} \phi_{m_k}(t) = \phi(t), \qquad \text{uniformly in } t \qquad (13)$$

Now it turns out that ϕ is continuously differentiable on the interval $(\tau - c, \tau + c)$ and that it satisfies the integral equation (V′) and, hence, the initial-value problem (I′). In other words, ϕ is a solution of (I′).

The preceding discussion gives rise to the **Cauchy–Peano existence theorem:**

If $f \in C(D)$ and $(\tau, \xi) \in D$, then the initial-value problem (I′) *has a solution defined on $|t - \tau| \le c$ where c is as defined in Fig. 5.*

We mention in passing that a special case of Eq. (12),

$$\phi(t_{j+1}) = \phi(t_j) + f(t_j, \phi(t_j))(t_{j+1} - t_j) \qquad (14)$$

is known as **Euler's method** of solving ordinary differential equations.

It should be noted that the above result yields only a *sufficient condition.* In other words, when f is not continuous on the domain D, then the initial-value problem (I′) may or may not possess a solution in the sense defined in Section II.

The above result asserts the existence of a solution (I′) "locally," that is, only on a sufficiently short time interval (determined by $|t-\tau| \le c$). In general, this assertion cannot be changed to existence of a solution for all $t \ge \tau$ (or for all $t \le \tau$) as the following example shows:

The initial-value problem,

$$x' = x^2, \qquad x(\tau) = \xi$$

has a solution $\phi(t) = \xi[1 - \xi(t-\tau)]^{-1}$ which exists forward in time for $\xi > 0$ only until $t = \tau + \xi^{-1}$.

B. Continuation of Solutions

Our next task is to determine if it is possible to extend a solution ϕ to a larger interval than was indicated above ($|t-\tau| \le c$). The answer to this is affirmative. To see this, suppose that $f \in C(D)$ and suppose also that f is bounded on D. Suppose also that by some procedure, as above, it was possible to show that ϕ is a solution of the scalar differential equation,

$$x' = f(t, x) \qquad (\mathrm{E}')$$

on an interval $J = (a, b)$. Using expression (V′) for ϕ it is an easy matter to show that the limit of $\phi(t)$ as t approaches a from the right exists and that the limit of $\phi(t)$ as t approaches b from the left exists; that is,

$$\lim_{t \to a^+} \phi(t) = \phi(a^+)$$

and

$$\lim_{t \to b^-} \phi(t) = \phi(b^-)$$

Now clearly, if the point $(a, \phi(a^+)) \in D$ (resp., if $(b, \phi(b^-)) \in D$), then by repeating the procedure given in the above results (ε-approximate solution result and Peano–Cauchy theorem), the solution ϕ can be continued to the left past the point $t = a$ (resp., to the right past the point $t = b$). Indeed, it should be clear that repeated applications of these procedures will make it possible to continue the solution ϕ to the boundary of D. This is depicted in Fig. 7. It is worthwhile to note that the solution ϕ in this figure exists over the interval J' and not over the interval $\tilde{J}$.

We summarize the preceding discussion in the following **continuation result**:

If $f \in C(D)$ and if f is bounded on D, then all solutions of (E′) *can be extended to the boundary of D. These solutions are then* **noncontinuable**.

C. Uniqueness of Solutions

Next, we address the question of uniqueness of solutions. To accomplish this, we require the following concept:

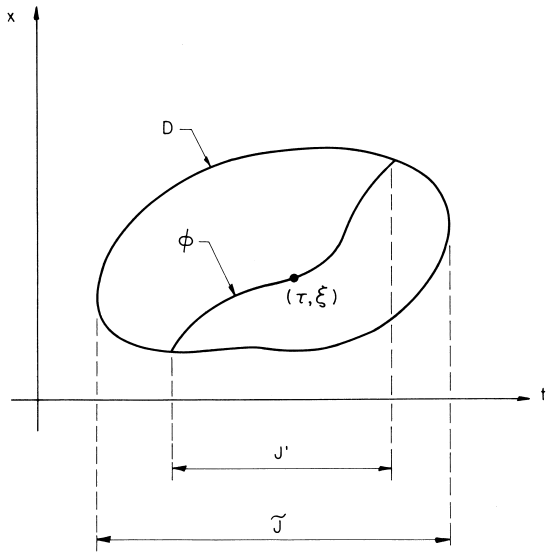


FIGURE 7 Continuation of a solution ϕ to ∂D.

$f \in C(D)$ is said to satisfy a **Lipschitz condition** on D (with respect to x) with **Lipschitz constant** L if

$$|f(t, \bar{\bar{x}}) - f(t, \bar{x})| \le L|\bar{\bar{x}} - \bar{x}| \qquad (15)$$

for all $(t, \bar{x})(t, \bar{\bar{x}})$ in D. The function f is said to be **Lipschitz continuous** in x on D in this case.

For example, it can be shown that if $\partial f(t, x)/\partial x$ exists and is continuous on D, then f will be Lipschitz continuous on any compact and convex subsect D_0 of D.

In order to establish a uniqueness result for solutions of the initial value problem (I′), we will also require a result known as the **Gronwall inequality**: Let r and k be continuous nonnegative real functions defined on an interval $[a, b]$ and let $\delta \ge 0$ be a constant. If

$$r(t) \le \delta + \int_a^t k(s) r(s)\, ds \qquad (16)$$

then

$$r(t) \le \delta \exp\left(\int_a^t k(s)\, ds \right) \qquad (17)$$

Now suppose that for (I′) the Cauchy–Peano theorem holds and suppose that for *one* given $(\tau, \xi) \in D$, *two* solutions ϕ_1 and ϕ_2 exist over some interval $|t-\tau| \le d$, $d > 0$. On the interval $\tau \le t \le \tau + d$ we now have, using (V′) to express ϕ_1 and ϕ_2,

$$\phi_1(t) - \phi_2(t) = \int_\tau^t [f(s, \phi_1(s)) - f(s, \phi_2(s))]\, ds \qquad (18)$$

Now if, in addition, f is Lipschitz continuous in x, then Eq. (18) yields:

$$|\phi_1(t) - \phi_2(t)| \le \int_\tau^t L|\phi_1(s) - \phi_2(s)|\, ds$$

Letting $r(t) = |\phi_1(t) - \phi_2(t)|$, $\delta = 0$, and $k(t) \equiv L$, and applying the Gronwall inequality, we now obtain:

$$|\phi_1(t) - \phi_2(t)| \le 0 \quad \text{for all} \quad \tau \le t \le \tau + d$$

Hence, it must be true that $\phi_1(t) = \phi_2(t)$ on $\tau \le t \le \tau + d$. A similar argument will also work for the interval $\tau - d \le t \le \tau$.

Summarizing, we have the following **uniqueness result**:

If $f \in C(D)$ and if f satisfies a Lipschitz condition on D with Lipschitz constant L, then the initial-value problem (I′) *has at most one solution on any interval $|t - \tau| \le d$, $d > 0$.*

If the solution ϕ of (I′) is unique, then the ε-approximate solutions constructed before will tend to ϕ as $\varepsilon \to 0^+$ and this is the basis for justifying Euler's method—a numerical method of constructing approximations to ϕ. Now, if we assume that f satisfies a Lipschitz condition, an alternative classical method of approximation is the **method of successive approximations**. Specifically, let $f \in C(D)$ and let S be the rectangle in D centered at (τ, ξ) shown in Fig. 5 and let c be defined as in Fig. 5. Successive approximations for (I′), or equivalently for (V′), are defined as:

$$\phi_0(t) = \xi$$

$$\phi_{m+1}(t) = \xi + \int_\tau^t f(s, \phi_m(s))\, ds, \qquad (19)$$

$$m = 0, 1, 2, \ldots$$

for $|t - \tau| \le c$.

The following result is the basis for justifying the method of successive approximations:

If $f \in C(D)$ and if f is Lipschitz continuous on S with constant L, then the successive approximations ϕ_m, $m = 0, 1, 2, \ldots$, given in Eq. (19) *exist on $|t - \tau| \le c$, are continuous there, and converge uniformly, as $m \to \infty$, to the unique solution of* (I′).

D. Continuity of Solutions with Respect to Parameters

Our next objective is to study the dependence of solutions ϕ of (I′) on initial data (τ, ξ). In this connection, we find it advantageous to highlight this dependence by writing $\phi(t) = \phi(t, \tau, \xi)$.

Now suppose that $f \in C(D)$ and suppose that f satisfies a Lipschitz condition on D with Lipschitz constant L. Furthermore, suppose that ϕ and ψ solve:

$$x' = f(t, x) \qquad (E')$$

on an interval $|t - \tau| \le d$ with $\psi(\tau) = \xi_0$ and $\phi(\tau) = \xi$. Then, by using (V′), we obtain:

$$|\phi(t, \tau, \xi) - \psi(t, \tau, \xi_0)| \le |\xi - \xi_0| + \int_\tau^t L|\phi(s, \tau, \xi) - \psi(s, \tau, \xi_0)|\, ds$$

and by using the Gronwall inequality with $\delta = |\xi - \xi_0|$, $L \equiv k(s)$, and $r(t) = |\phi(t) - \psi(t)|$, we obtain the estimate:

$$|\phi(t, \tau, \xi) - \psi(t, \tau, \xi_0)| \le |\xi - \xi_0| \exp(L|t - \tau|)$$

$$t \in |t - \tau| \le d \qquad (20)$$

If, in particular, we consider a sequence of initial conditions $\{\xi_m\}$ having the property that $\xi_m \to \xi_0$ as $m \to \infty$, then it follows from Eq. (20) that $\phi(t, \tau, \xi_m) \to \phi(t, \tau, \xi_0)$, uniformly in t on $|t - \tau| \le d$.

Summarizing, we have the following **continuous dependence result**:

Let $f \in C(D)$ and assume that f satisfies a Lipschitz condition on D. Then, the unique solution $\phi(t, \tau, \xi)$ of (I′), *existing on some bounded interval containing τ, depends continuously on ξ, uniformly in t.*

This means that if $\xi_m \to \xi_0$ then $\phi(t, \tau, \xi_m) \to \phi(t, \tau, \xi_0)$, uniformly in t on $|t - \tau| \le d$ for some $d > 0$.

In a similar manner we can show that $\phi(t, \tau, \xi)$ will depend continuously on the initial time τ. Furthermore, if the differential equation (E′) depends on a parameter, say μ, then the solutions of the corresponding initial value problem may also depend in a continuous manner on μ, provided that certain safeguards are present. We consider a specific case in the following.

Consider the initial-value problem,

$$x' = f(t, x, \mu), \qquad x(\tau) = \xi_\mu = \mu + 1 \qquad (I'_\mu)$$

where μ is a scalar parameter. Let f satisfy Lipschitz conditions with respect to x and μ for $(t, x) \in D$ and for $|\mu - \mu_0| < c$. Using an argument similar to the one employed in connection with Eq. (20), we can show that the solution $\phi(t, \tau, \xi_\mu, \mu)$ of (I'_μ), where ξ_μ depends continuously on μ, is a continuous function of μ (i.e., as $\mu \to \mu_0$, $\xi_\mu \to \xi_{\mu 0}$, and $\phi(t, \tau, \xi_\mu, \mu) \to \phi(t, \tau, \xi_{\mu 0}, \mu_0)$).

As an example, consider the initial-value problem

$$x' = x + \mu t, \qquad x(\tau) = \xi_\mu = \mu + 1 \qquad (21)$$

The right-hand side of Eq. (21) has a Lipschitz constant with respect to x equal to one and with respect to μ equal to $|a - b|$, where $J = (a, b)$ is assumed to be a bounded t-interval. The solution of (I'_μ) is

$$\phi(t, \tau, \xi_\mu, \mu) = [\mu(\tau + 2) + 1]e^{(t-\tau)} - \mu(t + 1) \qquad (22)$$

At $t = \tau$, we have $\phi(\tau, \tau, \xi_\mu, \mu) = \mu + 1 = x(\tau)$. Now what happens when $\mu \to 0$? In this case, Eq. (21) becomes:

$$x' = x, \qquad x(\tau) = 1 \qquad (23)$$

while the solution of Eq. (23) is

$$\phi(t) = e^{t-\tau} \tag{24}$$

If we let $\mu \to 0$ in Eq. (22), then we also obtain:

$$\phi(t, \tau, \xi_0, 0) = e^{t-\tau} \tag{25}$$

as expected.

E. Systems of Equations

In the interests of simplicity, we have considered thus far in this section the one-dimensional initial-value problem (I′). It turns out that the preceding results can be restated and proved for initial-value problems (I) involving systems of equations (E). In doing so, one must replace absolute values of scalars by norms of vectors or matrices, convergence of scalars by convergence of vectors and matrices, and so forth.

Rather than go through the task of restating the preceding results for systems of equations, we present as an illustration an additional result. Specifically, consider the linear system of equations

$$x' = A(t)x + g(t) = f(t, x) \tag{LN}$$

where $A(t) = [a_{ij}(t)]$ is an $n \times n$ matrix and $g(t)$ is an n-vector. We assume that $a_{ij}(t)$, $i, j = 1, \ldots, n$, and $g_i(t)$, $i = 1, \ldots, n$, are real and continuous functions defined on an interval J.

By making use of the taxicab norm given by

$$|x| = \sum_{i=1}^{n} |x_i|$$

it is an easy matter to show that for t on any compact subinterval J_0 of J there exists $L_0 \geq 0$ such that

$$|f(t, \bar{x}) - f(t, \bar{\bar{x}})| \leq \left(\sum_{i=1}^{n} \max_{l \leq j \leq n} |a_{ij}(t)| \right) |\bar{x} - \bar{\bar{x}}|$$
$$\leq L_0 |\bar{x} - \bar{\bar{x}}| \tag{26}$$

If we now invoke the results of the present section (rephrased for systems of equations), we obtain the following:

Suppose that $A(t)$ and $g(t)$ in (LN) *are defined and continuous on an interval J. Then for any τ in J and any ξ in R^n, Eq.* (LN) *has a unique solution satisfying $x(\tau) = \xi$. This solution exists on the entire interval J and is continuous in (t, τ, ξ). If A and g depend continuously on parameters $\lambda \in R^l$, then the solution will also vary continuously with λ.*

It is emphasized that the above result is a *global* result, since the solution ϕ exists over the *entire* interval J. On the other hand, as noted before, our earlier results in this section will in general be of a *local* nature.

F. Differentiability of Solutions with Respect to Parameters

In some of the preceding results we investigated the continuity of solutions with respect to parameters. Next, we address the question of differentiability of solutions with respect to parameters for the initial-value problem,

$$x' = f(t, x), \qquad x(\tau) = \xi \tag{I}$$

Again we assume that $f \in C(D)$, $(\tau, \xi) \in D \subset R^{n+1}$, where D is a domain. In addition, we assume that $f(t, x)$ is differentiable with respect to $x_1, \ldots, x_n$ and we form the Jacobian matrix $f_x(t, x)$ given by:

$$f_x(t, x) = \frac{\partial f}{\partial x}(t, x)$$
$$= \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1}(t, x) & \dfrac{\partial f_1}{\partial x_2}(t, x) & \cdots & \dfrac{\partial f_1}{\partial x_n}(t, x) \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial f_n}{\partial x_1}(t, x) & \dfrac{\partial f_n}{\partial x_2}(t, x) & \cdots & \dfrac{\partial f_n}{\partial x_n}(t, x) \end{bmatrix} \tag{27}$$

Under the above conditions, it can be shown that when f_x exists and is continuous, then the solution ϕ of (I) depends smoothly on the parameters of the problem. More specifically,

Let $f \in C(D)$, let f_x exist, and let $f_x \in C(D)$. If $\phi(t, \tau, \xi)$ is the solution of (E) *such that $\phi(\tau, \tau, \xi) = \xi$, then ϕ is continuously differentiable in (t, τ, ξ). Each vector-valued function $\partial\phi/\partial\xi_i$ or $\partial\phi/\partial\tau$ will solve*:

$$y' = f_x(t, \phi(t, \tau, \xi))y$$

as a function of t while

$$\frac{\partial \phi}{\partial \tau}(\tau, \tau, \xi) = -f(\tau, \xi)$$

and

$$\frac{\partial \phi}{\partial \xi}(\tau, \tau, \xi) = E_n$$

where E_n denotes the $n \times n$ identity matrix.

G. Comparison Theory

We conclude this section by touching on the **comparison theory** for differential equations. This theory is useful in continuation of solutions, in establishing estimates on bounds of solutions, and, as we will see later, in stability theory. Before we can present some of the main results of this theory we need to introduce a few concepts.

Once again we consider the scalar initial-value problem,

$$x' = f(t, x), \qquad x(\tau) = \xi \tag{I'}$$

where $f \in C(D)$ and $(\tau, \xi) \in D$. For (I′) we define the **maximal solution** ϕ_M as that noncontinuable solution of (I′) having the property that if ϕ is any other solution of (I′), then $\phi_M(t) \geq \phi(t)$ as long as both solutions are defined. The **minimal solution** ϕ_m of (I′) is defined in a similar manner. It is not too difficult to prove that ϕ_M and ϕ_m actually exist.

In what follows, we also need the concept of **upper right Dini derivative** D^+x. Given $x: (\alpha, \beta) \to R$ and $x \in C(\alpha, \beta)$ (i.e., x is a continuous real-valued function defined on the interval (α, β)), we define:

$$D^+x(t) = \lim_{h \to 0^+} \sup[x(t+h) - x(t)]/h$$

$$\triangleq \overline{\lim}[x(t+h) - x(t)]/h$$

where lim sup denotes the limit supremum. The **lower right Dini derivative** D^-x is defined similarly, replacing the lim sup by the lim inf.

We will also require the concept of differential inequalities, such as, for example, the inequality,

$$D^+x(t) \leq f(t, x(t)) \qquad \text{on} \quad D \tag{28}$$

Any function ϕ satisfying Eq. (28) is called a *solution* for (28). Differential inequalities involving D^-x are defined similarly.

Our first comparison result is now as follows:

Suppose that the maximal solution ϕ_M of (I′) stays in D for all $t \in [\tau, T]$. If a continuous function $\psi(t)$ with $\psi(\tau) = \xi$ satisfies

$$\psi'(t) \triangleq D^+\psi(t) \leq f(t, \psi(t)) \qquad \text{on} \quad D$$

then it is true that

$$\psi(t) \leq \phi_M(t) \qquad \text{for all} \quad t \in [\tau, T]$$

A similar result involving minimal solutions can also be established.

The above result can now be applied to systems of equations to obtain estimates for the norms of solutions. We have the following:

Let $f \in C(D)$, $D \subset R^{n+1}$, and let ϕ be a solution of

$$x' = f(t, x), \qquad x(\tau) = \xi \tag{I}$$

Let $F(t, v)$ be a scalar-valued continuous function such that

$$|f(t, x)| \leq F(t, |x|) \qquad \text{for all} \quad (t, x) \in D$$

where $|f(t, x)|$ denotes any one of the equivalent norms of $f(t, x)$ on R^n. If $\eta \leq |\phi(\tau)|$ and if v_M denotes the maximal solution of the scalar comparison equation given by:

$$v' = F(t, v), \qquad u(\tau) = \eta \tag{29}$$

then

$$|\phi(t)| \leq v_M(t)$$

for as long as both functions exist. (Here, $|\phi(t)|$ denotes the norm of $\phi(t)$.)

As an application of this result, suppose that $f(t, x)$ in (E) is such that

$$|f(t, x)| \leq A|x| + B$$

for all $t \in J$ and for all $x \in R^n$, where $J = [t_0, T]$, and where $A > 0$, $B > 0$ are parameters. Then Eq. (29) assumes the form:

$$v' = Au + B, \qquad v(\tau) = \eta$$

According to the above result, we now have

$$v_M(t) = e^{A(t-\tau)}(\eta - B/A) + B/A$$

Since $v_M(t)$ exists for all $t \in J$, then so do the solutions of (E). Also, if ϕ is any solution of (E) with $|\phi(\tau)| \leq \eta$, then the estimate,

$$|\phi(t)| \leq e^{A(t-\tau)}(\eta - (\beta/A)) + (\beta/A)$$

is true.

IV. LINEAR SYSTEMS

Both in the theory of differential equations and in their applications, linear systems of ordinary differential equations are extremely important. In this section we first present the general properties of linear systems. We then turn our attention to the special cases of linear systems of ordinary differential equations with constant coefficients and linear systems of ordinary differential equations with periodic coefficients. We also address some of the properties of nth-order linear ordinary differential equations.

A. Linear Homogeneous and Nonhomogeneous Systems

We first consider linear homogeneous systems,

$$x' = A(t)x \tag{LH}$$

As noted in Section III, this system possesses unique solutions for every $(\tau, \xi) \in D$ where $x(\tau) = \xi$,

$$D = \{(t, x): t \in J = (a, b),\ x \in R^n (\text{or } x \in C^n)\}$$

when each element $a_{ij}(t)$ of matrix $A(t)$ is continuous over J. These solutions exist over the entire interval $J = (a, b)$

and they depend continuously on the initial conditions. In applications it is typical that $j = (-\infty, \infty)$. We note that $\phi(t) \equiv 0$, for all $t \in J$, is a solution of (LH), with $\phi(\tau) = 0$. This is called the **trivial solution** of (LH).

In this section we consider matrices and vectors which will be either real or complex valued. In the former case, the field of scalars for the x space is the field of real numbers ($F = R$) and in the latter case, the field for the x space is the field of complex numbers ($F = C$).

Now let V denote the set of all solutions of (LH) on J; let α_1, α_2 be scalars (i.e., $\alpha_1, \alpha_2 \in F$); and let ϕ_1, ϕ_2 be solutions of (LH) (i.e., $\phi_1, \phi_2 \in V$). Then it is easily verified that $\alpha_1\phi_1 + \alpha_2\phi_2$ will also be a solution of (LH) (i.e., $\alpha_1\phi_1 + \alpha_2\phi_2 \in V$). We have thus shown that V is a vector space. Now if we choose n linearly independent vectors $\xi_1, \ldots, \xi_n$ in the n-dimensional x-space, then there exist n solutions $\phi_1, \ldots, \phi_n$ of (LH) such that $\phi_1(\tau) = \xi_1, \ldots, \phi_n(\tau) = \xi_n$. It is an easy matter to verify that this set of solutions $\{\phi_1, \ldots, \phi_n\}$ is linearly independent and that it spans V. Thus, $\{\phi_1, \ldots, \phi_n\}$ is a basis of V and any solution ϕ can be expressed as a linear combination of the vectors $\phi_1, \ldots, \phi_n$.

Summarizing, we have

The set of solutions of (LH) *on the interval J forms an n-dimensional vector space.*

In view of the above result it now makes sense to define a **fundamental set of solutions** for (LH) as a set of n linearly independent solutions of (LH) on J. If $\{\phi_1, \ldots, \phi_n\}$ is such a set, then we can form the matrix:

$$\Phi = [\phi_1, \ldots, \phi_n] \tag{30}$$

which is called a **fundamental matrix** of (LH).

In the following, we enumerate some of the important properties of fundamental matrices. All of these properties are direct consequences of definition (30) and of the properties of solutions of (LH). We have

1. *A fundamental matrix Φ of* (LH) *satisfies the matrix equation*:

$$X' = A(t)X \tag{31}$$

where $X = [x_{ij}]$ denotes an $n \times n$ matrix. (*Observe that Eq.* (31) *consists of a system of n^2 first-order ordinary differential equations.*)

2. *If Φ is a solution of the matrix equation* (31) *on an interval J and if τ is any point of J, then*

$$\det \Phi(t) = \det \Phi(\tau) \exp\left(\int_\tau^t \operatorname{tr} A(s)\, ds \right) \quad \text{for every } t \in J$$

(*Here*, det Φ *denotes the determinant of* Φ *and* tr A *denotes the trace of the matrix A.*) *This result is known as* **Abel's formula**.

3. *A solution Φ of matrix equation* (31) *is a fundamental matrix of* (LH) *if and only if its determinant is nonzero for all $t \in J$.* (*This result is a direct consequence of Abel's formula.*)

4. *If Φ is a fundamental matrix of* (LH) *and if C is any nonsingular constant $n \times n$ matrix, then ΦC is also a fundamental matrix of* (LH). *Moreover, if Ψ is any other fundamental matrix of* (LH), *then there exists a constant $n \times n$ nonsingular matrix P such that $\Psi = \Phi P$.*

In the following, we let $\{e_1, e_2, \ldots, e_n\}$ denote the set of vectors $e_1^{\mathrm{T}} = (1, 0, \ldots, 0)$, $e_2^{\mathrm{T}} = (0, 1, 0, \ldots, 0)$, $\ldots$, $e_n^{\mathrm{T}} = (0, \ldots, 0, 1)$. We call a fundamental matrix Φ of (LH) whose columns are determined by the linearly independent solutions $\phi_1, \ldots, \phi_n$ with

$$\phi_1(\tau) = e_1, \ldots, \phi_n(\tau) e_n, \qquad \tau \in J$$

the **state transition matrix** Φ for (LH). Equivalently, if Ψ is *any* fundamental matrix of (LH), then the matrix Φ determined by:

$$\Phi(t, \tau) \triangleq \Psi(t)\Psi^{-1}(\tau) \qquad \text{for all} \quad t, \tau \in J$$

is said to be the **state transition matrix** of (LH).

We now enumerate several properties of state transition matrices. All of these are direct consequences of the definition of state transition matrix and of the properties of fundamental matrices. In the following we let $\tau \in J$, we let $\phi(\tau) = \xi$, and we let $\Phi(t, \tau)$ denote the state transition matrix for (LH) for all $t \in J$. Then,

1. $\Phi(t, \tau)$ *is the unique solution of the matrix equation*:

$$\frac{\partial}{\partial t}\Phi(t, \tau) \triangleq \Phi'(t, \tau) = A(t)\Phi(t, \tau)$$

with $\Phi(\tau, \tau) = E$, the $n \times n$ identity matrix.

2. $\phi(t, \tau)$ *is nonsingular for all $t \in J$.*

3. *For any $t, \sigma, \tau \in J$, we have*

$$\Phi(t, \tau) = \Phi(t, \sigma)\Phi(\sigma, \tau)$$

4. $[\Phi(t, \tau)]^{-1} \triangleq \Phi^{-1}(t, \tau) = \Phi(\tau, t)$ *for all* $t, \tau \in J$.

5. *The unique solution $\phi(t, \tau, \xi)$ of* (LH) *with $\phi(\tau, \tau, \xi) = \xi$ specified, is given by*:

$$\phi(t, \tau, \xi) = \Phi(t, \tau)\xi \qquad \text{for all} \quad t \in J \tag{32}$$

In engineering and physics applications, $\phi(t)$ is interpreted as representing the "state" of a (dynamical) system represented by (LH) at time t and $\phi(\tau) = \xi$ is interpreted as representing the "state" at time τ. In Eq. (32), $\Phi(t, \tau)$

relates the "states" of (LH) at t and τ. This motivated the name "state transition matrix."

Let us now consider a couple of specific examples.

1. For the system of equations

$$\begin{aligned} x_1' &= 5x_1 - 2x_2 \\ x_2' &= 4x_1 - x_2 \end{aligned} \tag{33}$$

we have

$$A(t) \equiv A = \begin{bmatrix} 5 & -2 \\ 4 & -1 \end{bmatrix} \qquad \text{for all} \quad t \in (-\infty, \infty)$$

Two linearly independent solutions for Eq. (33) are

$$\phi_1(t) = \begin{bmatrix} e^{3t} \\ e^{3t} \end{bmatrix}, \qquad \phi_2(t) = \begin{bmatrix} e^{t} \\ 2e^{t} \end{bmatrix}$$

The matrix

$$\Phi(t) = \begin{bmatrix} e^{3t} & e^{t} \\ e^{3t} & 2e^{t} \end{bmatrix}$$

satisfies the equation $\Phi' = A\Phi$ and

$$\det \Phi(t) = e^{4t} \neq 0 \qquad \text{for all} \quad t \in (-\infty, \infty)$$

Thus, Φ is a fundamental matrix for Eq. (33). Also, in view of Abel's formula, we obtain:

$$\begin{aligned} \det \Phi(t) &= \det \Phi(\tau) \exp\left[\int_\tau^t \operatorname{tr} A(s)\, ds \right] \\ &= e^{4\tau} \exp\left[\int_\tau^t 4\, ds \right] = e^{4t} \end{aligned}$$

for all $t \in (-\infty, \infty)$

as expected. Finally, since

$$\Phi^{-1}(t) = \begin{bmatrix} 2e^{-3t} & -e^{-3t} \\ -e^{-t} & -e^{-t} \end{bmatrix}$$

we obtain for the transition matrix of Eq. (33),

$$\Psi(t)\Psi^{-1}(\tau) = \begin{bmatrix} 2e^{3(t-\tau)} - e^{t-\tau} & -e^{3(t-\tau)} + e^{t-\tau} \\ 2e^{3(t-\tau)} - 2e^{t-\tau} & -e^{3(t-\tau)} + 2e^{t-\tau} \end{bmatrix}$$

2. For the system,

$$x_1' = x_2, \qquad x_2' = tx_2 \tag{34}$$

we have

$$A(t) \begin{bmatrix} 0 & 1 \\ 0 & t \end{bmatrix} \qquad \text{for all} \quad t \in (-\infty, \infty)$$

Two linearly independent solutions of Eq. (34) are

$$\phi_1(t) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \phi_2(t) = \begin{bmatrix} \displaystyle\int_\tau^t e^{\eta^2/2}\, d\eta \\ e^{t^2/2} \end{bmatrix}$$

The matrix

$$\Phi(t) = \begin{bmatrix} 1 & \displaystyle\int_\tau^t e^{\eta^2/2} \\ 0 & e^{t^2/2} \end{bmatrix}$$

satisfies the matrix equation $\Phi' = A(t)\Phi$ and

$$\det \Phi(t) = e^{t^2/2} \qquad \text{for all} \quad t \in (-\infty, \infty)$$

Therefore, Φ is a fundamental matrix for Eq. (34). Also, in view of Abel's formula, we have

$$\begin{aligned} \det \Phi(t) &= \det \Phi(\tau) \exp\left[\int_\tau^t \operatorname{tr} A(s)\, ds \right] \\ &= e^{\tau^2/2} \exp\left[\int_\tau^t \eta\, d\eta \right] = e^{-t^2/2} \end{aligned}$$

for all $t \in (-\infty, \infty)$

as expected. Also, since

$$\Phi^{-1}(t) = \begin{bmatrix} 1 & -e^{-t^2/2} \displaystyle\int_\tau^t e^{\eta^2/2}\, d\eta \\ 0 & -e^{-t^2/2} \end{bmatrix}$$

we obtain for the state transition of Eq. (34),

$$\Phi(t)\Phi^{-1}(\tau) = \begin{bmatrix} 1 & e^{-\tau^2/2} \displaystyle\int_\tau^t e^{\eta^2/2}\, d\eta \\ 0 & e^{(t^2-\tau^2)/2} \end{bmatrix}$$

Finally, suppose that $\phi(\tau) = \xi = [1, 1]^{\mathrm{T}}$. Then

$$\phi(t, \tau, \xi) = \Phi(t, \tau)\xi = \begin{bmatrix} 1 + e^{-\tau^2/2} \displaystyle\int_\tau^t e^{\eta^2/2}\, d\eta \\ e^{(t^2-\tau^2)/2} \end{bmatrix}$$

Next, we consider linear nonhomogeneous systems,

$$x' = A(t)x + g(t) \tag{LN}$$

We assume that $A(t)$ and $g(t)$ are defined and continuous over $R = (-\infty, \infty)$ (i.e., each component $a_{ij}(t)$ of $A(t)$ and each component $g_k(t)$ of $g(t)$ is defined and continuous on R). As noted in Section III, system (LN) has for any $\tau \in R$ and any $\xi \in R^n$ a unique solution satisfying $x(\tau) = \xi$. This solution exists on the *entire* real line R and is continuous in (t, τ, ξ). Furthermore, if A and g depend continuously on parameters $\lambda \in R^l$, then the solution will also vary continuously with λ. Indeed, if we differentiate the function

$$\phi(t, \tau, \xi) = \Phi(t, \tau)\xi + \int_\tau^t \Phi(t, \eta) g(\eta)\, d\eta \tag{35}$$

with respect to t to obtain $\phi'(t, \tau, \xi)$, and if we substitute ϕ and ϕ' into (LN) (for x), then it is an easy matter to

verify that Eq. (35) is in fact the unique solution of (LN) with $\phi(t, \tau, \xi) = \xi$.

We note that when $\xi = 0$, then Eq. (35) reduces to

$$\phi_{\mathrm{p}}(t) = \int_{\tau}^{t} \Phi(t, \eta) g(\eta)\, d\eta \tag{36}$$

and when $\xi \neq 0$ but $g(t) \equiv 0$, then Eq. (35) reduces to

$$\phi_{\mathrm{h}}(t) = \Phi(t, \tau)\xi \tag{37}$$

Thus, the solution of (LN) may be viewed as consisting of a component due to the "forcing term" $g(t)$ and another component due to the initial data ξ. This type of separation is in general possible only in linear systems of differential equations. We call ϕ_p the **particular solution** and ϕ_h the **homogeneous solution** of (LN).

Before proceeding to linear systems with constant coefficients, we introduce adjoint equations. Let Φ be a fundamental matrix for the linear homogeneous system (LH). Then,

$$(\Phi^{-1})' = -\Phi^{-1}\Phi'\Phi^{-1} = -\Phi^{-1}A(t)$$

Taking the conjugate transpose of both sides, we obtain:

$$(\Phi^{*-1})' = -A^{*}(t)\Phi^{*-1}$$

This implies that Φ^{*-1} is a fundamental matrix for the system:

$$y' = -A^{*}(t)y, \qquad t \in J \tag{38}$$

We call Eq. (38) the **adjoint** to (LH), and we call the matrix equation,

$$Y' = -A^{*}(t)Y, \qquad t \in J$$

the **adjoint** to matrix equation (31).

One of the principal properties of adjoint systems is summarized in the following result:

If Φ is a fundamental matrix for (LH), *then Ψ is a fundamental matrix for its adjoint* (38) *if and only if*

$$\Psi^{*}\Phi = C$$

where C is some constant nonsingular matrix.

B. Linear Systems with Constant Coefficients

We now turn our attention to linear systems with constant coefficients. For purposes of motivation, we first consider the scalar initial-value problem:

$$x' = ax, \qquad x(\tau) = \xi \tag{39}$$

It is easily verified that Eq. (39) has the solution:

$$\phi(t) = e^{a(t-\tau)}\xi \tag{40}$$

It turns out that a similar result holds for the system of linear equations with constant coefficients,

$$x' = Ax \tag{L}$$

By making use of the Weierstrass M test, it is not difficult to verify the following result:

Let A be a constant $n \times n$ matrix which may be real or complex and let $S_N(t)$ denote the partial sum of matrices defined by the formula,

$$S_N(t) = E + \sum_{k=1}^{N} \frac{t^k}{k!} A^k \tag{41}$$

where E denotes the $n \times n$ identity matrix and $k!$ stands for k factorial. Then each element of the matrix $S_N(t)$ converges absolutely and uniformly on any finite t interval $(-a, a)$, $a > 0$, as $N \to \infty$.

This result enables us to define the matrix,

$$e^{At} = E + \sum_{k=1}^{\infty} \frac{t^k}{k!} A^k \tag{42}$$

for any $-\infty < t < \infty$.

It should be clear that when $A(t) \equiv A$, system (LH) reduces to system (L). Consequently, the results we established above for (LH) are also applicable to (L). Now by making use of these results, the definition of e^{At} in Eq. (42), and the convergence properties of $S_N(t)$ in Eq. (42), it is not difficult to establish several important properties of e^{At} and of (L). To this end we let $J = (-\infty, \infty)$ and $\tau \in j$, and we let A be a given constant $n \times n$ matrix for (L). Then the following is true:

1. $\Phi(t) \triangleq e^{At}$ *is a fundamental matrix for* (L) *for* $t \in J$.
2. *The state transition matrix for* (L) *is given by* $\Phi(t, \tau) = e^{A(t-\tau)} \triangleq \Phi(t - \tau)$, $t \in J$.
3. $e^{At_1}e^{At_2} = e^{A(t_1+t_2)}$ *for all* $t_1, t_2 \in J$.
4. $Ae^{At} = e^{At}A$ *for all* $t \in J$.
5. $(e^{At})^{-1} = e^{-At}$ *for all* $t \in J$.
6. *The unique solution ϕ of* (L) *with* $\phi(\tau) = \xi$ *is given by*:

$$\phi(t, \tau, \xi) = e^{A(t-\tau)}\xi \tag{43}$$

Notice that solution (43) of (L) such that $\phi(\tau) = \xi$ depends on t and τ only via the difference $t - \tau$. This is the typical situation for *general autonomous* systems that satisfy uniqueness conditions. Indeed, if $\phi(t)$ is a solution of

$$x' = F(x), \qquad x(0) = \xi$$

then clearly $\phi(t - \tau)$ will be a solution of

$$x' = F(x), \qquad x(\tau) = \xi$$

Next, we consider the "forced" system of equations,

$$x' = Ax + g(t) \tag{44}$$

where $g: J \to R^n$ is continuous. Clearly, Eq. (44) is a special case of (LN). In view of Eq. (35) we thus have

$$\phi(t) = e^{A(t-\tau)}\xi + e^{At}\int_\tau^t e^{-A\eta}g(\eta)\,d\eta \tag{45}$$

for the solution of Eq. (44).

Next, we address the problem of evaluating the state transition matrix. While there is no general procedure for evaluating such a matrix for a time-varying matrix $A(t)$, there are several such procedures for determining e^{At} when $A(t) \equiv A$. In the following, we consider two such methods.

We begin by recalling the Laplace transform. To this end, we consider a vector $f(t)=[f_1(t), \ldots, f_n(t)]^{\mathrm{T}}$, where $f_i: [0, \infty) \to R$, $i = 1, \ldots, n$. Letting s denote a complex variable, we define the *Laplace transform* of f_i as:

$$\hat{f}_i(s) = \mathscr{L}[f_i(t)] \triangleq \int_0^\infty f_i(t)e^{-st}\,dt \tag{46}$$

provided, of course, that the integral in Eq. (46) exists. (In this case f_i is said to be Laplace transformable.) Also, we define the Laplace transform of the vector $f(t)$ by:

$$\hat{f}(s) = [\hat{f}_1(s), \ldots, \hat{f}_n(s)]^{\mathrm{T}},$$

and we define the Laplace transform of a matrix $C(t) = [c_{ij}(t)]$ similarly. Thus, if $c_{ij}: [0, \infty) \to R$ and if each c_{ij} is Laplace transformable, then the Laplace transform of $C(t)$ is defined by:

$$\hat{C}(s) = \mathscr{L}[c_{ij}(t)] = [\mathscr{L}c_{ij}(t)] = [\hat{c}_{ij}(s)]$$

Now consider the initial value problem,

$$x' = Ax, \qquad x(0) = \xi \tag{47}$$

Taking the Laplace transform of both sides of Eq. (47), we obtain:

$$sx(s) - \xi = Ax(s)$$

or

$$(sE - A)x(s) = \xi$$

or

$$x(s) = (sE - A)^{-1}\xi \tag{48}$$

where E denotes the $n \times n$ identity matrix. It can be shown by analytic continuation that $(sE - A)^{-1}$ exists for all s, except at the eigenvalues of A (i.e., except at those values of s where the equation $\det(sE - A) = 0$ is satisfied). Taking the inverse Laplace transform of Eq. (48) (i.e., by reversing the procedure and obtaining, for example, in Eq. (46) $f_i(t)$ from $f_i(s)$ we obtain for the solution of Eq. (47),

$$\phi(t) = \mathscr{L}^{-1}[(sE - A)^{-1}]\xi = \Phi(t, 0)\xi = e^{At}\xi \tag{49}$$

where $\mathscr{L}^{-1}[\hat{f}(s)] = f(t)$ denotes the inverse Laplace transform of $\hat{f}(s)$. It follows from Eqs. (49) and (48) that

$$\hat{\Phi}(s) = (sE - A)^{-1}$$

and

$$\Phi(t, 0) \triangleq \Phi(t) = \mathscr{L}^{-1}[(sE - A)^{-1}] = e^{At} \tag{50}$$

Finally, note that when the initial time $\tau \neq 0$, we can immediately compute $\Phi(t, \tau) = \Phi(t - \tau) = e^{A(t-\tau)}$.

Next, let us consider a "forced" system of the form:

$$x' = Ax + g(t), \qquad x(0) = \xi \tag{51}$$

and let us assume that the Laplace transform of g exists. Taking the Laplace transform of both sides of Eq. (51) yields:

$$s\hat{x}(s) - \xi = A\hat{x}(s) + \hat{g}(s)$$

or

$$(sE - A)\hat{x}(s) = \xi + \hat{g}(s)$$

or

$$\begin{aligned}\hat{x}(s) &= (sE - A)^{-1}\xi + (sE - A)^{-1}\hat{g}(s)\\ &= \hat{\Phi}(s)\xi + \hat{\Phi}(s)\hat{g}(s) \triangleq \hat{\phi}_{\mathrm{h}}(s) + \hat{\phi}_{\mathrm{p}}(s)\end{aligned} \tag{52}$$

Taking the inverse Laplace transform of both sides of Eq. (52) and using Eq. (45), we obtain:

$$\begin{aligned}\phi(t) &= \phi_{\mathrm{h}}(t) + \phi_{\mathrm{p}}(t)\\ &= \mathscr{L}^{-1}[(sE - A)^{-1}]\xi + \mathscr{L}^{-1}[(sE - A)^{-1}\hat{g}(s)]\\ &= \Phi(t, 0)\xi + \int_0^t \Phi(t - \eta)g(\eta)\,d\eta\end{aligned} \tag{53}$$

Therefore,

$$\phi_{\mathrm{p}}(t) = \int_0^t \Phi(t - \eta)g(\eta)\,d\eta \tag{54}$$

as expected. We call the expression in Eq. (54) the **convolution** of Φ and g. Clearly, convolution of Φ and g in the time domain corresponds to multiplication of Φ and g in the s domain.

Let us now consider the specific initial-value problem,

$$\begin{aligned}x_1' &= -x_1 + x_2, &\qquad x_1(0) &= -1\\ x_2' &= -2x_2 + u(t), &\qquad x_2(0) &= 0\end{aligned} \tag{55}$$

where

$$u(t) = \begin{cases}1 & \text{for} \quad t > 0\\ 0 & \text{elsewhere}\end{cases}$$

We have, in this case,

$$(sE - A) = \begin{bmatrix} s+1 & -1 \\ 0 & s+2 \end{bmatrix}$$

$$\hat{\Phi}(s) = (sE - A)^{-1}$$

$$= \begin{bmatrix} \dfrac{1}{s+1} & \left(\dfrac{1}{s+1} - \dfrac{1}{s+2}\right) \\ 0 & \dfrac{1}{s+2} \end{bmatrix}$$

and

$$\Phi(t) = e^{At} = \mathscr{L}^{-1}[\hat{\Phi}(s)]$$

$$= \begin{bmatrix} e^{-t} & (e^{-t} - e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix}$$

It now follows that

$$\phi_{\mathrm{h}}(t) = \begin{bmatrix} e^{-t} & (e^{-t} - e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix} \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} -e^{-t} \\ 0 \end{bmatrix}$$

Also, since $\hat{u}(s) = 1/s$, we have

$$\hat{\phi}_{\mathrm{p}}(s) = \begin{bmatrix} \dfrac{1}{s+1} & \left(\dfrac{1}{s+1} - \dfrac{1}{s+2}\right) \\ 0 & \dfrac{1}{s+2} \end{bmatrix} \begin{bmatrix} 0 \\ \dfrac{1}{s} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{1}{2} & \left(\dfrac{1}{s}\right) + \dfrac{1}{2} & \left(\dfrac{1}{s+2}\right) - \dfrac{1}{s+1} \\ \dfrac{1}{2} & \left(\dfrac{1}{s}\right) - \dfrac{1}{2} & \left(\dfrac{1}{s+2}\right) \end{bmatrix}$$

and

$$\phi_{\mathrm{p}}(t) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^{-2t} - e^{-t} \\ \frac{1}{2} - \frac{1}{2}e^{-2t} \end{bmatrix}$$

Therefore, the solution of the initial value problem (55) is

$$\phi(t) = \phi_{\mathrm{p}}(t) + \phi_{\mathrm{h}}(t) = \begin{bmatrix} \frac{1}{2} - 2e^{-t} + \frac{1}{2}e^{-2t} \\ \frac{1}{2} - \frac{1}{2}e^{-2t} \end{bmatrix}$$

A second method of evaluating e^{At} and of solving initial value problems for (L) and Eq. (44) involves the transformation of A into a Jordan canonical form. Specifically, it is shown in linear algebra that for every complex $n \times n$ matrix A there exists a nonsingular $n \times n$ matrix P (i.e., $\det P \neq 0$) such that the matrix,

$$J = P^{-1}AP \tag{56}$$

is in the canonical form,

$$J = \begin{bmatrix} J_0 & & & \\ & J_1 & 0 & \\ & 0 & \ddots & \\ & & & J_s \end{bmatrix} \tag{57}$$

where J_0 is a diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_k$ (not necessarily distinct); that is,

$$J_0 = \begin{bmatrix} \lambda_1 & & & \\ & & 0 & \\ & \ddots & & \\ & 0 & & \\ & & & \lambda_k \end{bmatrix} \tag{58}$$

and each J_p is an $n_p \times n_p$ matrix of the form ($p = 1, \ldots, s$):

$$J_q = \begin{bmatrix} \lambda_{k+p} & 1 & 0 & \cdots & 0 \\ 0 & \lambda_{k+p} & 1 & & \\ \vdots & \vdots & \vdots & & \vdots \\ & & & \cdots & 1 \\ 0 & 0 & 0 & \cdots & \lambda_{k+p} \end{bmatrix} \tag{59}$$

where λ_{k+p} need not be different from λ_{k+q} if $p \neq q$ and $k + n_1 + \cdots + n_s = n$. The numbers λ_i, $i = 1, \ldots, k+s$, are the eigenvalues of A (i.e., the roots of the equation $\det(\lambda E - A) = 0$). If λ_i is a simple eigenvalue of A (i.e., it is not a repeated root of $\det(\lambda E - A) = 0$), then it appears in the block J_0. The blocks $J_0, J_1, \ldots, J_s$ are called **Jordan blocks** and J is called the **Jordan canonical form** of A.

Returning to the subject at hand, we consider once more the initial value problem (47) and let P be a real $n \times n$ nonsingular matrix which transforms A into a Jordan canonical form J. Consider the transformation $x = Py$ or, equivalently, $y = P^{-1}x$. Differentiating both sides with respect to t, we obtain:

$$\begin{aligned} y' &= P^{-1}x' = P^{-1}APy = Jy \\ y(\tau) &= P^{-1}\xi \end{aligned} \tag{60}$$

The solution of Eq. (60) is given as

$$\psi(t) = e^{J(t-\tau)}P^{-1}\xi \tag{61}$$

Using Eq. (61) and $x = Py$, we obtain for the solution of Eq. (47),

$$\phi(t) = Pe^{J(t-\tau)}P^{-1}\xi \tag{62}$$

In the case in which A has n *distinct* eigenvalues $\lambda_1, \ldots, \lambda_n$, we can choose $P = [p_1, p_2, \ldots, p_n]$ in such a way that p_i is an eigenvector corresponding to the eigenvalue λ_i, $i = 1, \ldots, n$ (i.e., $p_i \neq 0$ satisfies the equation

$\lambda_i p_i = A p_i$). Then the Jordan matrix $J = P^{-1}AP$ assumes the form:

$$J = \begin{bmatrix} \lambda_1 & & & 0 \\ & \ddots & & \\ 0 & & & \\ & & & \lambda_n \end{bmatrix}$$

Using the power series representation, Eq. (42), we immediately obtain the expression:

$$e^{Jt} = \begin{bmatrix} e^{\lambda_1 t} & & & 0 \\ & \ddots & & \\ 0 & & & \\ & & & e^{\lambda_n t} \end{bmatrix}$$

In this case, we have the expression for the solution of Eq. (47):

$$\phi(t) = P \begin{bmatrix} e^{\lambda_1(t-\tau)} & & & 0 \\ & \ddots & & \\ 0 & & & \\ & & & e^{\lambda_n(t-\tau)} \end{bmatrix} P^{-1}\xi$$

In the general case when A has repeated eigenvalues, we can no longer diagonalize A and we have to be content with the Jordan form given by Eq. (57). In this case, $P = [v_1, \ldots, v_n]$, where the v_i denote generalized eigenvectors. Using the power series representation, Eq. (42) and the very special nature of the Jordan blocks (58) and (59), it is not difficult to show that in the case of repeated eigenvalues we have

$$e^{Jt} = \begin{bmatrix} e^{J_0 t} & & & 0 \\ & e^{J_1 t} & & \\ & & \ddots & \\ 0 & & & e^{J_s t} \end{bmatrix} \quad -\infty < t < \infty$$

where

$$e^{J_0 t} = \begin{bmatrix} e^{\lambda_1 t} & & & 0 \\ & \ddots & & \\ 0 & & & \\ & & & e^{\lambda_k t} \end{bmatrix}$$

and

$$e^{tJ_i} = e^{\lambda_{k+i} t} \begin{bmatrix} 1 & t & \cdots & t^{n_i-1}/(n_i-1)! \\ 0 & 1 & \cdots & t^{n_i-2}/(n_i-2)! \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 1 \end{bmatrix}$$

$$i = 1, \ldots, s$$

From Eq. (62), it now follows that the solution of Eq. (47) is given by:

$$\phi(t) = P \begin{bmatrix} e^{J_0(t-\tau)} & & & 0 \\ & e^{J_1(t-\tau)} & & \\ & & \ddots & \\ 0 & & & e^{J_s(t-\tau)} \end{bmatrix} P^{-1}\xi$$

As a specific example of the above procedure of determining the state transition matrix, consider the initial-value problem:

$$x_1' = -x_1 + x_2, \qquad x_1(0) = 1$$
$$x_2' = -2x_2, \qquad x_2(0) = 2$$

In this case, we have

$$A = \begin{bmatrix} -1 & 1 \\ 0 & -2 \end{bmatrix}$$

with eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -2$ and with corresponding eigenvectors,

$$P_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad P_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

We thus have

$$P = [p_1, p_2] = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \qquad P^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

and

$$J = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}$$

Furthermore,we obtain:

$$\begin{bmatrix} \phi_1(t) \\ \phi_2(t) \end{bmatrix} = P e^{Jt} P^{-1}\xi$$
$$= \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{-2t} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$
$$= \begin{bmatrix} 3e^{-t} - 2e^{-2t} \\ 2e^{-2t} \end{bmatrix}$$

C. Linear Systems With Periodic Coefficients

Next, we consider linear homogeneous periodic systems of the form:

$$x' = A(t)x, \qquad -\infty < t < \infty \qquad \text{(P)}$$

where the elements of A are continuous functions on R and where

$$A(t) = A(t+T) \qquad (63)$$

for some $T > 0$ which is a period of A.

The principal result for (P) which we shall present here (called **Floquet theory**) involves the logarithm of a matrix, the existence of which is not too difficult to establish. We have the following:

Let B be a nonsingular $n \times n$ matrix. Then there exists an $n \times n$ matrix F, called the logarithm of B, such that

$$e^F = B$$

Using properties of the fundamental matrix as well as the concept of the logarithm of a matrix, we are in a position to prove the following fundamental result for (P):

Let Eq. (63) *be true and let $A(t)$ be continuous on $(-\infty, \infty)$. If $\Phi(t)$ is a fundamental matrix for* (P), *then so is $\Phi(t+T)$, $-\infty < t < \infty$. Moreover, corresponding to every Φ, there exists a nonsingular matrix P which is also periodic with period T and a constant matrix R, such that*

$$\Phi(t) = P(t)e^{tR} \tag{64}$$

It is not difficult to show, using Eq. (64), that if the fundamental matrix Φ for (P) is known over *any* interval of length T, then it is automatically known for all $-\infty < t < \infty$. For example, if $\Phi(t)$ is known for all t over the interval $[t_0, t_0 + T]$, then we can show that

$$\Phi(t) = P(t)e^{tT^{-1}\log C} \tag{65}$$

where $C = \Phi(t_0)^{-1}\Phi(t_0 + T)$. Thus, since $P(t)$ is periodic, $\Phi(t)$ as given in Eq. (65) will be known for all t over $(-\infty, \infty)$.

It can also be shown that even though the fundamental matrix Φ does not determine R uniquely in Eq. (64), the set of all fundamental matrices of (P), and hence of $A(t)$, determines uniquely all quantities associated with e^{TR} which are invariant under a similarity transformation. Specifically, the set of all fundamental matrices of $A(t)$ determine a unique set of eigenvalues of the matrix e^{TR}, $\lambda_1, \ldots, \lambda_n$, which are called the **Floquet multipliers** associated with $A(t)$. None of these vanishes since $\Pi\ \lambda_i = \det e^{TR} \neq 0$. Also, the eigenvalues of R, $\rho_1, \ldots, \rho_n$, are called the **characteristic exponents** of $A(t)$.

Now let us suppose that all ρ_i are such that $\text{Re}\ \rho_i < 0$ (i.e., the real part of each ρ_i is negative) and let $\alpha = \min_i |\text{Re}\ \rho_i|$. Now, if we arrange things so that R in Eq. (64) is in Jordan canonical form, then it is a simple matter to show that there exists a $k > 0$ such that for each component $\phi_i(t)$ of the solution $\phi(t)$,

$$|\phi_i(t)| \leq ke^{\alpha t} \qquad \text{for all} \quad t \geq 0 \tag{66}$$

and $|\phi_i(t)| \to 0$ as $t \to \infty$. In other words, if the eigenvalues ρ_i, $i = 1, \ldots, n$, of R have negative real parts, then the norm of any solution of (P) tends to zero as $t \to \infty$ at an exponential rate.

Finally, by using Eq. (64) we can write:

$$P(t) = \Phi(t)e^{-tR} \tag{67}$$

which in turn can be used to see that $AP - P' = PR$. Thus, for the transformation,

$$x = P(t)y \tag{68}$$

we compute

$$\begin{aligned} x' &= A(t)x = A(t)P(t)y \\ &= P'(t)y + P(t)y' = (P(t)y)' \end{aligned}$$

or

$$\begin{aligned} y' &= P^{-1}(t)[A(t)P(t) - P'(t)]y \\ &= P^{-1}(t)(P(t)R)y = Ry \end{aligned}$$

This shows that *transformation* (68) *reduces the linear, homogeneous, periodic system* (P) to

$$y' = Ry \tag{69}$$

a linear homogeneous system with constant coefficients. Since $P(t)$ is nonsingular, we are thus able to deduce the properties of the solutions of (P) from those of Eq. (69), provided of course that we can determine the matrix $P(t)$.

D. Linear *n* th-Order Ordinary Differential Equations

We conclude this section by considering some of the more important aspects of linear nth-order ordinary differential equations. We shall consider equations of the form,

$$y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y^{(1)} + a_0(t)y = b(t) \tag{70}$$

$$y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y^{(1)} + a_0(t)y = 0 \tag{71}$$

and

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1y^{(1)} + a_0y = 0 \tag{72}$$

In Eqs. (70) and (71) the functions $a_k(t)$ and $b(t)$, $k = 1, \ldots, n-1$, are continuous on some appropriate time interval J. If we define the differential operator L_n by:

$$L_n = \frac{d^n}{dt^n} + a_{n-1}(t)\frac{d^{n-1}}{dt^{n-1}} + \cdots + a_1(t)\frac{d}{dt} + a_0(t) \tag{73}$$

then we can rewrite Eqs. (70) and (71) more compactly as

$$L_n y = b(t) \tag{74}$$

and

$$L_n y = 0 \tag{75}$$

respectively. We can rewrite Eq. (72) similarly by defining a differential operator L in the obvious way.

Following the procedure in Section II, we can reduce the study of Eq. (71) to the study of the system of n first-order ordinary differential equations,

$$x' = A(t)x \tag{LH}$$

where $A(t)$ is the **companion matrix** given by:

$$A(t) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0(t) & -a_1(t) & -a_2(t) & \cdots & -a_{n-1}(t) \end{bmatrix} \tag{76}$$

Since $A(t)$ is continuous on J, we know from Section III that there exists a unique solution $\phi(t)$, for all $t \in J$, to the initial-value problem,

$$x' = A(t)x, \qquad x(\tau) = \xi, \qquad \tau \in J$$

where $\xi = (\xi_1, \ldots, \xi_n)^{\mathrm{T}} \in R^n$. The first component of this solution is a solution of $L_n y = 0$ satisfying $y(\tau) = \xi_1$, $y'(\tau) = \xi_2, \ldots, y^{(n-1)}(\tau) = \xi_n$.

Now let $\phi_1, \ldots, \phi_n$ be n solutions of Eq. (75). Then we can easily show that the matrix,

$$\Phi(t) = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_n \\ \phi_1' & \phi_2' & \cdots & \phi_n' \\ \vdots & \vdots & & \vdots \\ \phi_1^{(n-1)} & \phi_2^{(n-1)} & \cdots & \phi_n^{(n-1)} \end{bmatrix}$$

is a solution of the matrix equation,

$$X' = A(t)X \tag{77}$$

where $A(t)$ is defined by Eq. (76). We call the determinant of Φ the **Wronskian** for Eq. (75) with respect to the solutions $\phi_1, \ldots, \phi_n$ and we denote it by:

$$W(\phi_1, \ldots, \phi_n) = \det \Phi(t)$$

Note that $W(\phi_1, \ldots, \phi_n)(t)$ depends on $t \in J$. Since Φ is a solution of matrix equation (77), then by Abel's formula it follows that for any $\tau \in J$ and for any $t \in J$,

$$\begin{aligned} W(\phi_1, \ldots, \phi_n)(t) &= \det \Phi(\tau) \exp\left[\int_\tau^t \operatorname{tr} A(s)\, ds \right] \\ &= W(\phi_1, \ldots, \phi_n)(\tau) \\ &\quad \times \exp\left\{ \int_\tau^t -a_{n-1}(s)\, ds \right\} \end{aligned} \tag{78}$$

As an example, consider the second-order differential equation

$$t^2 y'' + t y' - y = 0, \qquad 0 < t < \infty$$

which can be written equivalently as:

$$y'' + (1/t)y' - (1/t^2)y = 0, \qquad 0 < t < \infty \tag{79}$$

The functions $\phi_1(t = t$ and $\phi_2(t) = 1/t$ are clearly solutions of Eq. (79). We now form the matrix,

$$\Phi(t) = \begin{bmatrix} \phi_1 & \phi_2 \\ \phi_1' & \phi_2' \end{bmatrix} = \begin{bmatrix} t & 1/t \\ 1 & 1/t^2 \end{bmatrix}$$

which yields the Wronskian

$$W(\phi_1, \phi_2)(t) = \det \Phi(t) = -2/t, \qquad t > 0$$

In the notation of Eq. (76) we have $a_1(t) = 1/t$, $a_0(t) = -1/t^2$, and thus $a_1(s) = 1/s$. In view of Eq. (78), we have for any $\tau > 0$,

$$\begin{aligned} W(\phi_1, \phi_2)(t) &= \det \Phi(t) \\ &= W(\phi_1, \phi_2)(\tau) \exp\left\{ \int_\tau^t -a_1(s)\, ds \right\} \\ &= -(2/\tau) e^{\ln(\tau/t)} = -2/t, \qquad t > 0 \end{aligned}$$

as expected.

Similarly as in the case of systems of equations, we can prove the following result for nth-order differential equations:

A set of n solutions of Eq. (75), $\phi_1, \ldots, \phi_n$, *is linearly independent on J if and only if* $W(\phi_1, \ldots, \phi_n)(t) \neq 0$ *for all* $t \in J$. *Moreover, every solution of Eq.* (75) *is a linear combination of any set of n linearly independent solutions.*

The above result enables us to make the following definition:

A set of n linearly independent solutions of Eq. (75) *on* j, $\phi_1, \ldots, \phi_n$, *is called a* ***fundamental set*** *of solutions for Eq.* (75).

Next, we turn our attention to nonhomogeneous linear nth-order ordinary differential equations of the form (70). As shown in Section II, the study of Eq. (70) reduces to the study of the system of n first-order ordinary differential equations,

$$x' = A(t)x + g(t) \tag{80}$$

where $A(t)$ is given by Eq. (76) and $g(t) = [0, \ldots, 0, b(t)]^{\mathrm{T}}$. Recall that for given $\tau \in J$ and given $x(\tau) = \xi \in R^n$, Eq. (80) has a unique solution given by $\phi = \phi_{\mathrm{h}} + \phi_{\mathrm{p}}$, where $\phi_h(t) = \Phi(t, \tau)\xi$ is a solution of (LH),

$\phi(t, \tau)$ denotes the state transition matrix of $A(t)$, and ϕ_p is a particular solution of Eq. (80), given by:

$$\phi_p(t) = \int_\tau^t \Phi(t, s)g(s)\,ds = \Phi(t)\int_\tau^t \Phi^{-1}(s)g(s)\,ds$$

We now specialize this result from the n-dimensional system (80) to the corresponding nth-order equation (70) to obtain the following result:

If $\{\phi_1, \ldots, \phi_n\}$ is a fundamental set for the equation $L_n y = 0$, then the unique solution ψ of the equation $L_n y = b(t)$ satisfying $\psi(\tau) = \xi_1, \ldots, \psi^{(n-1)}(\tau) = \xi_n$ is given by:

$$\psi(t) = \psi_h(t) + \psi_P(t) = \psi_h(t) + \sum_{k=1}^{n} \phi_k(t) \times \int_\tau^t \frac{W_k(\phi_1, \ldots, \phi_n)(s)}{W(\phi_1, \ldots, \phi_n)(s)} b(s)\,ds \tag{81}$$

*Here, ψ_h is the solution of $L_n y = 0$ such that $\psi(\tau) = \xi_1$, $\psi'(\tau) = \xi_2, \ldots, \psi^{(n-1)}(\tau) = \xi_n$, and $W_k(\phi_1, \ldots, \phi_n)(t)$ is obtained from $W(\phi_1, \ldots, \phi_n)(t)$ by replacing the k*th *column in $W(\phi_1, \ldots, \phi_n)(t)$* by $(0, \ldots, 0, 1)^T$.

We apply the above example to the second-order differential equation,

$$y'' + (1/t)y' - (y/t^2) = b(t), \qquad 0 < t < \infty$$

where $b(t)$ is a real continuous function for all $t > 0$. From the example involving Eq. (79) we have $\phi_1(t) = t$, $\phi_2(t) = 1/t$, and $W(\phi_1, \phi_2)(t) = -2/t$, $t > 0$. Also,

$$W_1(\phi_1, \phi_2)(t) = \begin{vmatrix} 0 & 1/t \\ 1 & -1/t^2 \end{vmatrix} = -\frac{1}{t},$$

$$W_1(\phi_1, \phi_2)(t) = \begin{vmatrix} t & 0 \\ 1 & 1 \end{vmatrix} = t$$

From Eq. (81) we now have

$$\psi(t) = \psi_h(t) + \psi_p(t) = \psi_h(t) + \frac{t}{2}\int_\tau^t b(s)\,ds - \frac{1}{2t}\int_\tau^t s^2 b(s)\,ds$$

Next, we consider nth-order ordinary differential equations with constant coefficients given by Eq. (72) which can equivalently be written as $L_n y = 0$, where

$$L_n = \frac{d^n}{dt^n} + a_{n-1}\frac{d^{n-1}}{dt^{n-1}} + \cdots + a_1\frac{d}{dt} + a_0$$

We assume that $J = (-\infty, \infty)$, we call

$$p(\lambda) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 \tag{82}$$

the **characteristic polynomial** of the differential equation (72), and we call

$$p(\lambda) = 0 \tag{83}$$

the **characteristic equation** of Eq. (72). The roots of $p(\lambda)$ are called the **characteristic roots** of Eq. (72).

We see that the study of Eq. (72) reduces to the study of the system of first-order ordinary differential equations with constant coefficients given by $x' = Ax$, where

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} \end{bmatrix} \tag{84}$$

The following result, which is proved in a straightforward manner, connects Eq. (72) and $x' = Ax$ with A given by Eq. (84):

The characteristic polynomial of A in Eq. (83) *is precisely the characteristic polynomial $p(\lambda)$ given by Eq.* (82), *that is*,

$$p(\lambda) = \det(\lambda E_n - A)$$

The next result enumerates a fundamental set for Eq. (72):

Let $\lambda_1, \ldots, \lambda_s$ be the distinct roots of the characteristic equation (83) *and suppose that λ_i has multiplicity m_i, $i = 1, \ldots, s$, with $\sum_{i=1}^{s} m_i = n$. Then the following set of functions is a fundamental set for Eq.* (72):

$$t^k e^{\lambda_i t}, \qquad k = 0, 1, \ldots, m_i - 1, \quad i = 1, \ldots, s \tag{85}$$

As a specific example, consider:

$$p(\lambda) = (\lambda - 2)(\lambda + 3)^2(\lambda + i)(\lambda - i)(\lambda - 4)^4 \tag{86}$$

Then, $n = 9$, and $\{e^{2t}, e^{-3t}, te^{-3t}, e^{-it}, e^{+it}, e^{4t}, te^{4t}, t^2e^{4t}, t^3e^{4t}\}$ is a fundamental set for the differential equation corresponding to the characteristic equation (86).

We conclude this section by considering adjoint equations. Corresponding to the operator L_n given in Eq. (73), we define a second linear operator L_n^+ of order n, which we call the **adjoint** of L_n, as follows. The domain of L_n^+ is the set of all continuous functions defined on J such that $[\bar{a}_j(t)y(t)]$ has j continuous derivatives on J. (Here, $\bar{a}_j(t)$ denotes the complex conjugate of $a_j(t)$.) For each function y, define:

$$L_n^+ y = (-1)^n y^{(n)} + (-1)^{n-1}(\bar{a}_{n-1}y)^{n-1} + \cdots + (-1)(\bar{a}_1 y)' + \bar{a}_0 y$$

The equation,

$$L_n^+ y = 0, \qquad t \in J$$

is called the **adjoint equation** to $L_n y = 0$.

When Eq. (75) is written in companion form (LH) with $A(t)$ given by Eq. (76), then the adjoint system is $z' = -A^*(t)z$, where

$$A^*(t) = \begin{bmatrix} 0 & 0 & \cdots & 0 & -\bar{a}_0(t) \\ 1 & 0 & \cdots & 0 & -\bar{a}_1(t) \\ 0 & 1 & \cdots & 0 & -\bar{a}_2(t) \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -\bar{a}_{n-1}(t) \end{bmatrix}$$

This adjoint system can be written in component form as:

$$\begin{aligned} z_1' &= \bar{a}_0(t) z_n \\ z_j' &= -z_{j-1} + \bar{a}_{j-1}(t) z_n, \qquad 2 \le j \le n \end{aligned} \tag{87}$$

If $\psi = [\psi_1, \psi_2, \ldots, \psi_n]^{\mathrm{T}}$ is a solution of Eq. (87) and if $a_j \psi_n$ has j derivatives, then

$$\psi_n' - (\bar{a}_{n-1}\psi_n) = -\psi_{n-1}$$

and

$$\psi_n'' - (\bar{a}_{n-1}\psi_n)' = -\psi_{n-1}' = \psi_{n-1} - (\bar{a}_{n-2}\psi_n)$$

or

$$\psi_n'' - (\bar{a}_{n-1}\psi_n)' + (\bar{a}_{n-2}\psi_n) = \psi_{n-2}$$

Continuing in this manner, we see that ψ_n solves $L_n^+ \psi = 0$.

V. STABILITY

Since there are no general rules for determining explicit formulas for the solutions of systems ofordinary differential equations (E), the analysis of initial-value problems (I) is accomplished along two lines: (1) a quantitative approach is used which usually involves the numerical solution of such problems by means of simulations on a digital computer, and (2) a qualitative approach is used which is usually concerned with the behavior of families of solutions of a given differential equation and which usually does not seek specific explicit solutions. As mentioned in Section I, we will concern ourselves primarily with qualitative aspects of ordinary differential equations.

The principal results of the qualitative approach include stability properties of an equilibrium point (rest position) and the boundedness of solutions of ordinary differential equations. We shall consider these topics in the present section.

A. The Concept of an Equilibrium Point

We shall concern ourselves with systems of equations,

$$x' = f(t, x) \tag{E}$$

where $x \in R^n$. When discussing *global results*, such as global asymptotic stability, we shall always assume that $f: R^+ \times R^n \to R^n$. On the other hand, when considering *local results*, we shall usually assume that $f: R^+ \times B(h) \to R^n$ for some $h > 0$, where $R^+ = [0, \infty)$, $B(h) = \{x \in R^n : |x| < h\}$ and $|\cdot|$ is any one of the equivalent norms on R^n. On some occasions we assume that $t \in R = (-\infty, \infty)$ rather than $t \in R^+$. Unless otherwise stated, we assume that for every (t_0, ξ), $t_0 \in R^+$, the initial-value problem,

$$x' = f(t, x), \qquad x(t_0) = \xi \tag{I}$$

possesses a unique solution $\phi(t, t_0, \xi)$ which is defined for all $t \ge t_0$ and which depends continuously on the initial data (t_0, ξ). Since it is natural in this section to think of t as representing time, we shall use the symbol t_0 in (I) to represent the initial time (rather than τ as was done earlier). Furthermore, we shall frequently use the symbol x_0 in place of ξ to represent the initial state.

A point $x_e \in R^n$ is called an **equilibrium point** of (E) (at time $t^* \in R^+$) if $f(t, x_e) = 0$ for all $t \ge t^*$. Other terms for equilibrium point include **stationary point, singular point, critical point**, and **rest position**.

We note that if x_e is an equilibrium point of (E) at t^*, then it is an equilibrium point at all $\tau \ge t^*$. Note also that in the case of autonomous systems,

$$x' = f(x) \tag{A}$$

and in case of T-periodic systems,

$$x' = f(t, x), \qquad f(t, x) = f(t + T, x) \tag{P}$$

a point $x_e \in R^n$ is an equilibrium at some t^* if and only if it is an equilibrium point at all times. Also note that if x_e is an equilibrium (at t^*) of (E), then the transformation $s = t - t^*$ reduces (E) to $dx/ds = f(s + t^*, x)$ and x_e is an equilibrium (at $s = 0$) of this system. For this reason, we shall henceforth assume that $t^* = 0$ in the above definition and we shall not mention t^* further. Note also that if x_e is an equilibrium point of (E), then for any $t_0 \ge 0$, $\phi(t, t_0, x_e) = x_e$ for all $t \ge t_0$ (i.e., x_e is a unique solution of (E) with initial data given by $\phi(t_0, t_0, x_e) = x_e$).

As a specific example, consider the simple pendulum introduced in Section II, which is described by equations of the form,

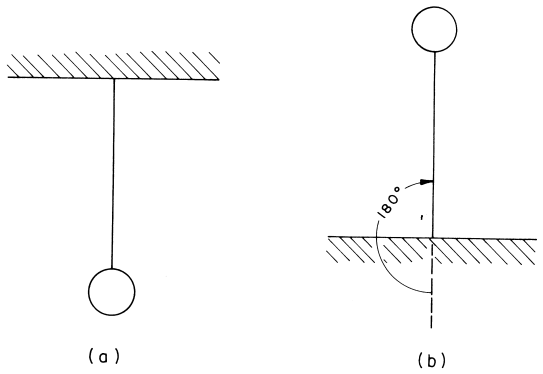


FIGURE 8 (a) Stable and (b) unstable equilibria of the simple pendulum.

$$\begin{aligned} x_1' &= x_2 \\ x_2' &= k \sin x_1, \qquad k > 0 \end{aligned} \tag{88}$$

Physically, the pendulum has two equilibrium points. One of these is located as shown in Fig. 8a and the second point is located as shown in Fig. 8b. However, the *model* of this pendulum, described by Eq. (88), has countably infinitely many equilibrium points located in R^2 at the points $(\pi n, 0)$, $n = 0, \pm 1, \pm 2, \ldots$.

An equilibrium point x_e of (E) is called an **isolated equilibrium point** if there is an $r > 0$ such that $B(x_e, r) \subset R^n$ contains no equilibrium points of (E) other than x_e itself. (Here, $B(x_e, r) = \{x \in R^n : |x - x_e| < r\}$.)

All equilibrium points of Eq. (88) are isolated equilibria in R^2. On the other hand, for the system,

$$\begin{aligned} x_1' &= -ax_1 + bx_1x_2 \\ x_2' &= -bx_1x_2 \end{aligned} \tag{89}$$

where $a > 0$, $b > 0$ are constants, every point on the positive x_2 axis is an equilibrium point for Eq. (89).

It should be noted that there are systems with *no equilibrium points* at all, as is the case, for example, in the system,

$$\begin{aligned} x_1' &= 2 + \sin(x_1 + x_2) + x_1 \\ x_2' &= 2 + \sin(x_1 + x_2) - x_1 \end{aligned} \tag{90}$$

Many important classes of systems possess only *one equilibrium point*. For example, the linear homogeneous system,

$$x' = A(t)x \tag{LH}$$

has a unique equilibrium at the origin if $A(t_0)$ is nonsingular for all $t_0 \geq 0$. Also, the system,

$$x' = f(x) \tag{A}$$

where f is assumed to be continuously differentiable with respect to all of its arguments and where

$$J(x_e) = \frac{\partial f}{\partial x}(x)\bigg|_{x=x_e}$$

denotes the $n \times n$ Jacobian matrix defined by $\partial f/\partial x = [\partial f_i/\partial x_j]$ has an isolated equilibrium at x_e if $f(x_e) = 0$ and $J(x_e)$ is nonsingular.

Unless otherwise stated, we shall assume throughout this section that a given equilibrium point is an isolated equilibrium. Also, we shall assume, unless otherwise stated, that in a given discussion, the equilibrium of interest is located at the origin of R^n. This assumption can be made without any loss of generality. To see this, assume that $x_e \neq 0$ is an equilibrium point of (E) (i.e., $f(t, x_e) = 0$ for all $t \geq 0$). Let $w = x - x_e$. Then $w = 0$ is an equilibrium of the transformed system,

$$w' = F(t, w) \tag{91}$$

where

$$F(t, w) = f(t, w + x_e) \tag{92}$$

Since Eq. (92) establishes a one-to-one correspondence between the solutions of (E) and Eq. (91), we may assume henceforth that (E) possesses the equilibrium of interest located at the origin. The equilibrium $x = 0$ will sometimes be referred to as the **trivial solution** of (E).

B. Definitions of Stability and Boundedness

We now state several definitions of stability of an equilibrium point, in the sense of Lyapunov.

The equilibrium $x = 0$ of (E) is **stable** if for every $\varepsilon > 0$ and any $t_0 \in R^+$ there exists a $\delta(\varepsilon, t_0) > 0$ such that:

$$|\phi(t, t_0, \xi)| < \varepsilon \qquad \text{for all} \quad t \geq t_0 \tag{93}$$

whenever

$$|\xi| < \delta(\varepsilon, t_0) \tag{94}$$

It is an easy matter to show that if the equilibrium $x = 0$ satisfies Eq. (93) for a single t_0 when Eq. (94) is true, then it also satisfies this condition at every initial time $t_0' > t_0$. Hence, in the preceding definition it suffices to take the single value $t = t_0$ in Eqs. (93) and (94).

Suppose that the initial-value problem (I) has a unique solution ϕ defined for t on an interval J containing t_0. By the **motion** through $(t_0, \xi = x(t_0))$ we mean the set $\{t, \phi(t) : t \in J\}$. This is, of course, the graph of the function ϕ. By the **trajectory** or **orbit** through $(t_0, \xi = x(t_0))$ we mean the set $C(x(t_0)) = \{\phi(t) : t \in J\}$. The **positive semitrajectory** (or **positive semiorbit**) is defined as $C^+(x(t_0)) = \{\phi(t) : t \in J \text{ and } t \geq t_0\}$. Also, the **negative trajectory** (or **negative semiorbit**) is defined as $C^-(x(t_0)) = \{\phi(t) : t \in J \text{ and } t \leq \tau\}$.

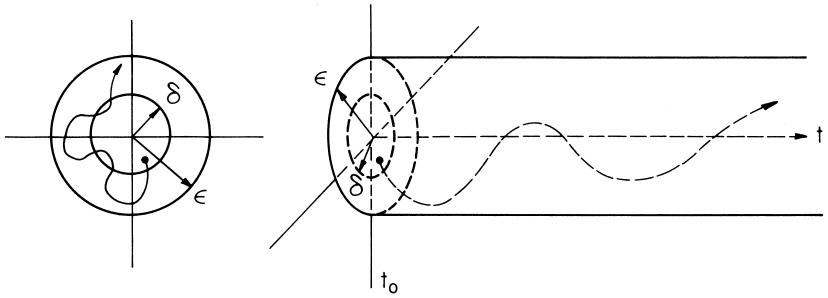


FIGURE 9 Stability of an equilibrium point.

Now, in Fig. 9 we depict the behavior of the trajectories in the vicinity of a stable equilibrium for the case $x \in R^2$.

When $x_e = 0$ is stable, by choosing the initial points in a sufficiently small spherical neighborhood, we can force the graph of the solution for $t \geq t_0$ to lie entirely inside a given cylinder.

In the above definition of stability, δ depends on ε and t_0 (i.e., $\delta = \delta(\varepsilon, t_0)$). If δ is independent of t_0 (i.e., $\delta = \delta(\varepsilon)$). then the equilibrium $x = 0$ of (E) is said to be **uniformly stable**.

The equilibrium $x = 0$ of (E) is said to be **asymptotically stable** if (1) it is stable, and (2) for every $t_0 \geq 0$ there exists an $\eta(t_0) > 0$ such that $\lim_{t \to \infty} \phi(t, t_0, \xi) = 0$ whenever $|\xi| < \eta$. Furthermore, the set of all $\xi \in R^n$ such that $\phi(t, t_0, \xi) \to 0$ as $t \to \infty$ for some $t_0 \geq 0$ is called the **domain of attraction** of the equilibrium $x = 0$ of (E). Also, if for (E) condition (2) is true, then the equilibrium $x = 0$ is said to be **attractive**.

The equilibrium $x = 0$ of (E) is said to be **uniformly asymptotically stable** if (1) it is uniformly stable, and (2) there is a $\delta_0 > 0$ such that for every $\varepsilon > 0$ and for any $t_0 \in R^+$ there exists a $T(\varepsilon) > 0$, independent of t_0, such that $|\phi(t, t_0, \xi)| < \varepsilon$ for all $t \geq t_0 + T(\varepsilon)$ whenever $|\xi| < \delta_0$.

In Fig. 10 we depict property (2), for uniform asymptotic stability, pictorially. *By choosing the initial points in a sufficiently small spherical neighborhood at $t = t_0$, we can force the graph of the solution to lie inside a given cylinder for all $t > t_0 + T(\varepsilon)$.* Condition (2) can be rephrased by saying that there exists a $\delta_0 > 0$ such that $\lim_{t \to \infty} \phi(t + t_0, t_0, \xi) = 0$ uniformly in (t_0, ξ) for $t_0 \geq 0$ and for $|\xi| \leq \delta_0$.

In applications, we are frequently interested in the following special case of uniform asymptotic stability: the equilibrium $x = 0$ of (E) is **exponentially stable** if there exists an $\alpha > 0$, and for every $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$, such that $|\phi(t, t_0, \xi)| \leq \varepsilon e^{\alpha(t - t_0)}$ for all $t \geq t_0$ whenever $|\xi| < \delta(\varepsilon)$ and $t \geq 0$.

In Fig. 11, the behavior of a solution in the vicinity of an exponentially stable equilibrium $x = 0$ is shown.

The equilibrium $x = 0$ of (E) is said to be **unstable** if it is not stable. In this case, there exists a $t_0 \geq 0$, $\varepsilon > 0$, a sequence $\xi_m \to 0$ of initial points, and a sequence $\{t_m\}$ such that $|\phi(t_0 + t_m, t_0, \xi_m)| \geq \varepsilon$ for all $m, t_m \geq 0$.

If $x = 0$ is an unstable equilibrium of (E), it still can happen that all the solutions tend to zero with increasing t. Thus, instability and attractivity are compatible concepts. Note that the equilibrium $x = 0$ is necessarily unstable if every neighborhood of the origin contains initial points corresponding to unbounded solutions (i.e., solutions whose norm $|\phi(t, t_0, \xi)|$ grows to infinity on a sequence $t_m \to \infty$). However, it can happen that a system (E) with unstable equilibrium $x = 0$ may have only bounded solutions.

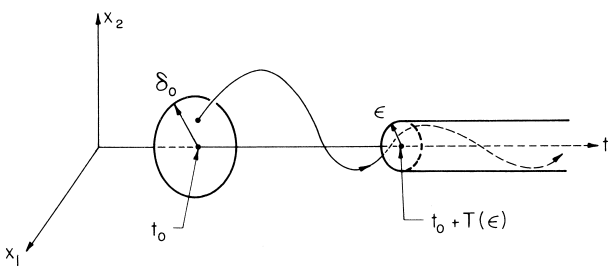


FIGURE 10 Attractivity of an equilibrium point.

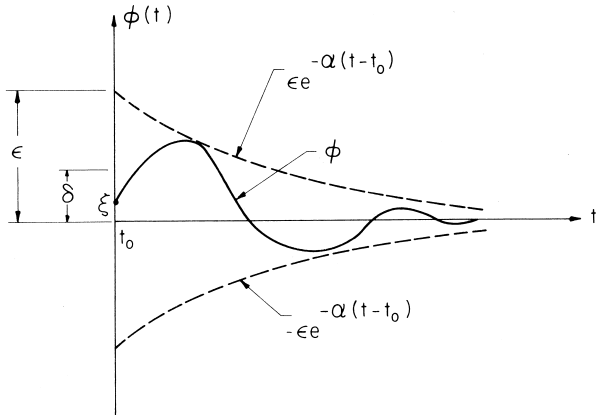


FIGURE 11 Exponential stability of an equilibrium point.

The above concepts pertain to *local properties of an equilibrium.* In the following definitions, we consider some *global characterizations of an equilibrium.*

A solution $\phi(t, t_0, \xi)$ of (E) is **bounded** if there exists a $\beta > 0$ such that $|\phi(t, t_0, \xi)| < \beta$ for all $t \geq t_0$, where β may depend on each solution. System (E) is said to possess **Lagrange stability** if for each $t_0 \geq 0$ and ξ the solution $\phi(t, t_0, \xi)$ is bounded.

The solutions of (E) are **uniformly bounded** if for any $\alpha > 0$ and $t_0 \in R^+$ there exists a $\beta = \beta(\alpha) > 0$ (independent of t_0) such that if $|\xi| < \alpha$, then $|\phi(t, t_0, \xi)| < \beta$ for all $t \geq t_0$.

The solutions of (E) are **uniformly ultimately bounded** (with bound B) if there exists a $B > 0$ and if corresponding to any $\alpha > 0$ and $t_0 \in R^+$ there exists a $T = T(\alpha)$ (independent of t_0) such that $|\xi| < \alpha$ implies that $|\phi(t, t_0, \xi)| < \beta$ for all $t \geq t_0 + T$.

In contrast to the boundedness properties given in the preceding three paragraphs, the concepts introduced earlier as well as those stated in the following are usually referred to as stability (respectively, instability) **in the sense of Lyapunov**.

The equilibrium $x = 0$ of (E) is **asymptotically stable in the large** if it is stable and if every solution of (E) tends to zero as $t \to \infty$. In this case, the domain of attraction of the equilibrium $x = 0$ of (E) is all of R^n. Note that in this case, $x = 0$ is the *only* equilibrium of (E).

The equilibrium $x = 0$ of (E) is **uniformly asymptotically stable in the large** if (1) it is uniformly stable, and (2) for any $\alpha > 0$ and any $\varepsilon > 0$, and $t_0 \in R^+$, there exists $T(\varepsilon, \alpha) > 0$, independent of t_0 such that if $|\xi| < \alpha$, then $|\phi(t, t_0, \xi)| < \varepsilon$ for all $t \geq t_0 + T(\varepsilon, \alpha)$.

Finally, the equilibrium $x = 0$ of (E) is **exponentially stable in the large** if there exists $\alpha > 0$ and for any $\beta > 0$, there exists $k(\beta) > 0$ such that $|\phi(t, t_0, \xi)| \leq k(\beta)|\xi|e^{-\alpha(t-t_0)}$ for all $t \geq t_0$ whenever $|\xi| < \beta$.

At this point it may be worthwhile to consider some specific examples:

1. The scalar equation,

$$x' = 0 \tag{95}$$

has for any initial condition $x(0) = c$ the solution $\phi(t, 0, c) = c$. All solutions are equilibria of Eq. (95). The trivial solution is *stable*; in fact, it is *uniformly stable*. However, it is not asymptotically stable.

2. The scalar equation,

$$x' = ax, \qquad a > 0 \tag{96}$$

has for every $x(0) = c$ the solution $\phi(t, 0, c) = ce^{at}$ and $x = 0$ is the only equilibrium of Eq. (96). This equilibrium is *unstable*.

3. The scalar equation,

$$x' = -ax, \qquad a > 0 \tag{97}$$

has for every $x(0) = c$ the solution $\phi(t, 0, c) = ce^{-at}$ and $x = 0$ is the only equilibrium of Eq. (97). This equilibrium is *exponentially stable in the large*.

4. The scalar equation,

$$x' = [-1/(t + 1)]x \tag{98}$$

has for every $x(t_0) = c$, $t_0 \geq 0$, a unique solution of the form $\phi(t, t_0, c) = (1 + t_0)c/(t + 1)$ and $x = 0$ is the only equilibrium of Eq. (98). This equilibrium is *uniformly stable* and *asymptotically stable in the large*, but it is not uniformly asymptotically stable.

C. Some Basic Properties of Autonomous and Periodic Systems

Making use of the properties of solutions and using the definitions of stability in the sense of Lyapunov it is not difficult to establish the following general stability results for systems described by:

$$x' = f(x) \tag{A}$$

and

$$x' = f(t, x), \qquad f(t, x) = f(t + T, x) \tag{P}$$

1. *If the equilibrium $x = 0$ of* (P) *(or of* (A)*) is stable, then it is in fact uniformly stable.*

2. *If the equilibrium $x = 0$ of* (P) *(or of* (A)*) is asymptotically stable, then it is uniformly asymptotically stable.*

D. Linear Systems

Next, by making use of the general properties of the solutions of linear autonomous homogeneous systems,

$$x' = Ax, \qquad t \geq 0 \tag{L}$$

and of linear homogeneous systems (with $A(t)$ continuous),

$$x' = A(t)x, \qquad t \geq t_0, \quad t_0 \geq 0 \tag{LH}$$

the following results are easily verified:

1. *The equilibrium $x = 0$ of* (LH) *is stable if and only if the solutions of* (LH) *are bounded. Equivalently, the equilibrium $x = 0$ of* (LH) *is stable if and only if*:

$$\sup_{t \geq t_0} |\Phi(t, t_0)| \stackrel{\Delta}{=} c(t_0) < \infty$$

where $|\Phi(t, t_0)|$ denotes the matrix norm induced by the vector norm used on R^n and sup denotes supremum.

2. *The equilibrium* $x = 0$ *of* (LH) *is uniformly stable if and only if*:

$$\sup_{t_0 \geq 0} c(t_0) \triangleq \sup_{t_0 \geq 0} \left(\sup_{t \geq t_0} |\Phi(t, t_0)| \right) \triangleq c_0 < \infty$$

3. *The following statements are equivalent*: (a) *The equilibrium* $x = 0$ *of* (LH) *is asymptotically stable.* (b) *The equilibrium* $x = 0$ *of* (LH) *is asymptotically stable in the large.* (c) $\lim_{t \to \infty} = 0$.

4. *The equilibrium* $x = 0$ *of* (LH) *is uniformly asymptotically stable if and only if it is exponentially stable*.

5. (a) *The equilibrium* $x = 0$ *of* (L) *is stable if all eigenvalues of A have nonpositive real parts and every eigenvalue of A that has a zero real part is a simple zero of the characteristic polynomial of A.* (b) *The equilibrium* $x = 0$ *of* (L) *is asymptotically stable if and only if all eigenvalues of A have negative real parts. In this case, there exist constants* $k > 0, \sigma > 0$ *such that* $|\Phi(t, t_0)| \leq k \exp[-\sigma(t - t_0)]$, $t_0 \leq t \leq \infty$, *where* $\Phi(t, t_0)$ *denotes the state transition matrix of* (L).

We shall find it convenient to use the following convention, which has become standard in the literature: A real $n \times n$ matrix A is called **stable** or a **Hurwitz matrix** if all of its eigenvalues have negative real parts. If at least one of the eigenvalues has a positive real part, then A is called **unstable**. A matrix A which is neither stable nor unstable is called **critical** and the eigenvalues of A with zero real parts are called **critical eigenvalues**.

Thus, the equilibrium $x = 0$ of (L) is asymptotically stable if and only if A is stable. If A is unstable, then $x = 0$ is unstable. If A is critical, then the equilibrium is stable if the eigenvalues with zero real parts correspond to a simple zero of the characteristic polynomial of A; otherwise, the equilibrium may be unstable.

Next, we consider the stability properties of linear periodic systems,

$$x' = A(t)x, \qquad A(t) = A(t + T) \qquad \text{(PL)}$$

where $A(t)$ is a continuous matrix for all $t \in R$. For such systems, the following results follow directly from Floquet theory:

1. *The equilibrium* $x = 0$ *of* (PL) *is uniformly stable if all eigenvalues of R in Eq.* (64) *have nonpositive real parts and any eigenvalue of R having zero real part is a simple zero of the characteristic polynomial of R.*

2. *The equilibrium* $x = 0$ *of* (PL) *is uniformly asymptotically stable if and only if all eigenvalues of R have negative real parts.*

E. Two-Dimensional Linear Systems

Before we present the principal results of the Lyapunov theory for general systems (E), we consider in detail the behavior of trajectories near an equilibrium point $x = 0$ of two-dimensional linear systems of the form

$$\begin{aligned} x_1' &= a_{11}x_1 + a_{12}x_2 \\ x_2' &= a_{21}x_1 + a_{22}x_2 \end{aligned} \qquad (99)$$

We can rewrite Eq. (99) equivalently as

$$x' = Ax \qquad (100)$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad (101)$$

When $\det A \neq 0$ system (99) will have one and only one equilibrium point, namely $x = 0$. We shall classify this equilibrium point (and, hence, system (99)) according to the following cases which the eigenvalues λ_1, λ_2 of A can assume:

1. λ_1, λ_2 are real and $\lambda_1 < 0, \lambda_2 < 0$: $x = 0$ is an asymptotically stable equilibrium point called a **stable node**.
2. λ_1, λ_2 are real and $\lambda > 0, \lambda_2 > 0$: $x = 0$ is an unstable equilibrium point called an **unstable node**.
3. λ_1, λ_2 are real and $\lambda_1 \lambda_2 < 0$: $x = 0$ is an unstable equilibrium point called a **saddle**.
4. λ_1, λ_2 are complex conjugates and $\text{Re } \lambda_1 = \text{Re } \lambda_2 < 0$: $x = 0$ is an asymptotically stable equilibrium point called a **stable focus**.
5. λ_1, λ_2 are complex conjugates and $\text{Re } \lambda_1 = \text{Re } \lambda_2 > 0$: $x = 0$ is an unstable equilibrium point called an **unstable focus**.
6. λ_1, λ_2 are complex conjugates and $\text{Re } \lambda_1 = \text{Re } \lambda_2 = 0$: $x = 0$ is a stable equilibrium called a **center**.

Using the results of Section IV, it is possible to solve Eq. (99) explicitly and verify that the qualitative behavior of the trajectories near the equilibrium $x = 0$ is as shown in Figs. 12–14 for the cases of a stable node, unstable node,

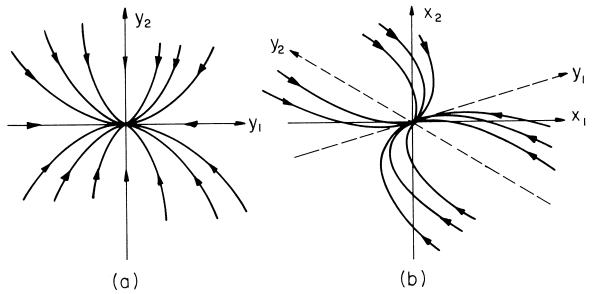


FIGURE 12 Trajectories near a stable node.

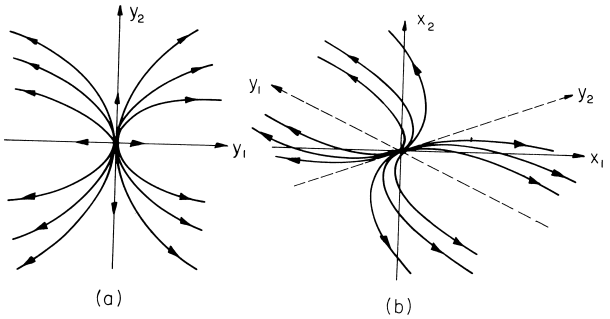


FIGURE 13 Trajectories near an unstable node.

and saddle, respectively. (The arrows on the trajectories point into the direction of increasing time.) In these cases, the figures labeled (a) correspond to systems in which A is in Jordan canonical form while the figures labeled (b) correspond to systems in which A is in some arbitrary form. In a similar manner, it is possible to verify that the qualitative behavior of the trajectories near the equilibrium $x = 0$ is as shown in Figs. 15–18 for the cases of a stable node (repeated eigenvalue case), an unstable focus, a stable focus, and a center, respectively. (For purposes of convention, we assumed in these figures that when λ_1, λ_2 are real and not equal, then $\lambda_1 > \lambda_2$.)

F. Lyapunov Functions

Next, we present general stability results for the equilibrium $x = 0$ of a system described by (E). Such results involve the existence of realvalued functions $v: D \to R$. In the case of local results (e.g., stability, instability, asymptotic stability, and exponential stability results), we shall usually only require that $D = B(h) \subset R^n$ for some $H > 0$, or $D = R^+ \times B(h)$. (Recall that $R^+ = (0, \infty)$ and $B(h) = \{x \in R^n : |x| < h\}$ where $|x|$ denotes any one of the equivalent norms of x on R^n.) On the other hand, in the case of global results (e.g., asymptotic stability in the large, exponential stability in the large, and uniform boundedness of solutions), we have to assume that $D = R^n$ or

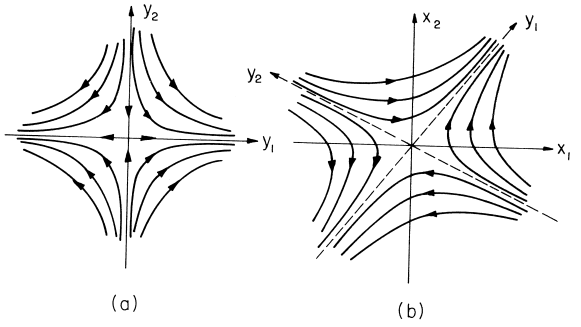


FIGURE 14 Trajectories near a saddle.

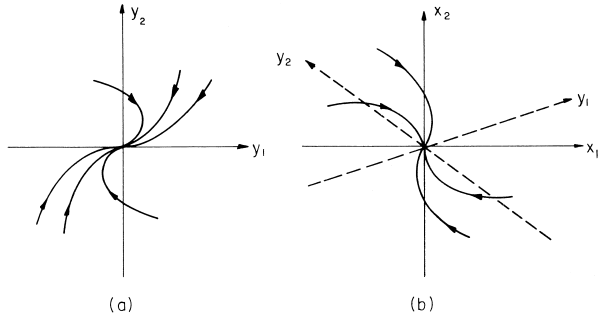


FIGURE 15 Trajectories near a stable node (repeated eigenvalue case).

$D = R^+ \times R^n$. Unless stated otherwise, we shall always assume that $v(t, 0) = 0$ for all $t \in R^+$ (resp., $v(0) = 0$).

Now let ϕ be an arbitrary solution of (E) and consider the function $t \mapsto v(t, \phi(t))$. If v is continuously differentiable with respect to all of its arguments, then we obtain (by the chain rule) the derivative of v with respect to t along the solutions of (E), $v'_{(E)}$, as:

$$v'_{(E)}(t, \phi(t)) = \frac{\partial v}{\partial t}(t, \phi(t)) + \nabla v(t, \phi(t))^T f(t, \phi(t))$$

Here, ∇v denotes the gradient vector of v with respect to x. For a solution $\phi(t, t_0, \xi)$ of (E), we have

$$v(t, \phi(t)) = v(t_0, \xi) + \int_{t_0}^{t} v'_{(E)}(\tau, \phi(\tau, t_0, \xi))\, d\tau$$

The above observations motivate the following: let v: $R^+ \times R^n \to R$ (resp., $v: R^+ \times B(h) \to R$) be continuously differentiable with respect to all of its arguments and let ∇v denote the gradient of v with respect to x. Then $v'_{(E)}: R^+ \times R^n \to R$ (resp., $v'_{(E)}: R^+ \times B(h) \to R$) is defined by:

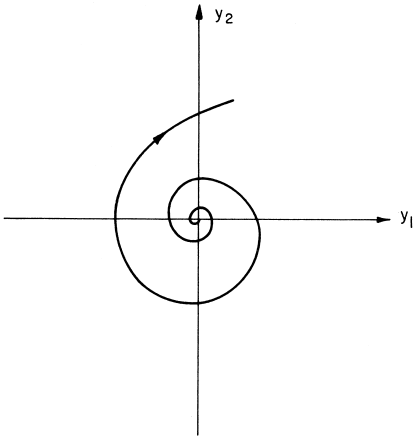


FIGURE 16 Trajectory near an unstable focus.

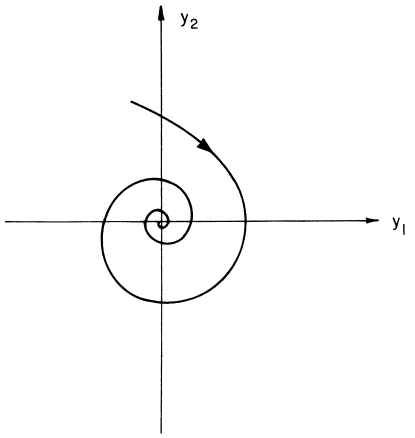


FIGURE 17 Trajectory near a stable focus.

$$v'_{(\mathrm{E})}(t,x) = \frac{\partial v}{\partial t}(t,x) + \sum_{i=1}^{n} \frac{\partial v}{\partial x_i}(t,x) f_i(t,x)$$

$$= \frac{\partial v}{\partial t}(t,x) + \nabla v(t,x)^{\mathrm{T}} f(t,x) \qquad (102)$$

We call $v'_{(\mathrm{E})}$ the **derivative of** v (**with respect to** t) **along the solutions of** (E).

It is important to note that in Eq. (102) the derivative of v with respect to t, along the solutions of (E), is evaluated *without having to solve* (E). The significance of this will become clear later. We also note that when $v\colon R^n \to R$ (resp., $v\colon B(h) \to R$), then Eq. (102) reduces to $v'_{(\mathrm{E})}(t,x) = \nabla v(x)^{\mathrm{T}} f(t,x)$. Also, in the case of autonomous systems (A), if $v\colon R^n \to R$ (resp., $v\colon B(h) \to R$), we have

$$v'_{(\mathrm{A})}(x) = \nabla v(x)^{\mathrm{T}} f(x) \qquad (103)$$

Occasionally, we shall require only that v be continuous on its domain of definition and that it satisfy locally a Lipschitz condition with respect to x. In such cases we define the **upper right-hand derivative of** v **with respect to** t **along the solutions of** (E) by:

$$v'_{(\mathrm{E})}(t,x) = \lim_{\theta \to 0^+} \sup(1/\theta)\{v(t+\theta, \phi(t+\theta,t,x)) - v(t,x)\}$$

$$= \lim_{\theta \to 0^+} \sup(1/\theta)\{v(t+\theta, x+\theta \cdot f(t,x)) - v(t,x)\} \qquad (104)$$

When v is continuously differentiable, then Eq. (104) reduces to Eq. (102).

We now give several important properties that v functions may possess. In doing so, we employ **Kamke comparison functions** defined as follows: a continuous function $\psi\colon [0, r_1] \to R^+$ (resp., $\psi\colon (0,\infty) \to R^+$) is said to belong to the **class** K (i.e., $\psi \in K$), if $\psi(0) = 0$ and if ψ is strictly increasing on $[0, r_1]$ (resp., on $[0,\infty)$). If $\psi\colon R^+ \to R^+$, if $\psi \in K$ and if $\lim_{r\to\infty} \psi(r) = \infty$, then ψ is said to belong to **class** KR.

We are now in a position to characterize v-functions in several ways. In the following, we assume that $v\colon R^+ \times R^n \to R$ (resp., $v\colon R^+ \times B(h) \to R$), that $v(0,t) = 0$ for all $t \in R^+$, and that v is continuous.

1. v is **positive definite** if, for some $r > 0$, there exists a $\psi \in K$ such that $v(t,x) \geq \psi(|x|)$ for all $t \geq 0$ and for all $x \in B(r)$.
2. v is **decrescent** if there exists a $\psi \in K$ such that $|v(t,x)| \leq \psi(|x|)$ for all $t \geq 0$ and for all $x \in B(r)$ for some $r > 0$.
3. $v\colon R^+ \times R^n \to R$ is **radially unbounded** if there exists a $\psi \in KR$ such that $v(t,x) \geq \psi(|x|)$ for all $t \geq 0$ and for all $x \in R^n$.
4. v is **negative definite** if $-v$ is positive definite.
5. v is **positive semidefinite** if $v(t,x) \geq 0$ for all $x \in B(r)$ for some $r > 0$ and for all $t \geq 0$.
6. v is **negative semidefinite** if $-v$ is positive semidefinite.

The definitions involving the above concepts, when $v\colon R^n \to R$ or $v\colon B(h) \to R$ (where $B(h) \subset R^n$ for some $h > 0$) involve obvious modifications. We now consider several specific cases:

1. The function $v\colon R^3 \to R$ given by $v(x) = x^{\mathrm{T}}x = x_1^2 + x_2^2 + x_3^2$ is positive definite and radially unbounded. (Here, x^{T} denotes the transpose of x.)
2. The function $v\colon R^3 \to R$ given by $v(x) = x_1^2 + (x_2 + x_3)^2$ is positive semidefinite (but not positive definite).
3. The function $v\colon R^2 \to R$ given by $v(x) = x_1^2 + x_2^2 - (x_1^2 + x_2^2)^3$ is positive definite but not radially unbounded.
4. The function $v\colon R^3 \to R$ given by $v(x) = x_1^2 + x_2^2$ is positive semidefinite (but not positive definite).

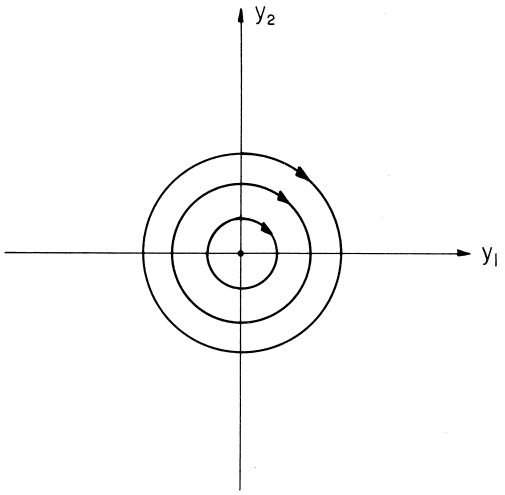


FIGURE 18 Trajectories near a center.

5. The function $v: R^2 \to R$ given by $v(x) = x_1^4/(1 + x_1^4) + x_2^4$ is positive definite but not radially unbounded.

6. The function $v: R^+ \times R^2 \to R$ given by $v(t, x) = (1 + \cos^2 t)x_1^2 + 2x_2^2$ is positive definite, decrescent, and radially unbounded.

7. The function $v: R^+ \times R^2 \to R$ given by $v(t, x) = (x_1^2 + x_2^2)\cos^2 t$ is positive semidefinite and decrescent.

8. The function $v: R^+ \times R^2 \to R$ given by $v(t, x) = (1 + t)(x_1^2 + x_2^2)$ is positive definite and radially unbounded but not decrescent.

9. The function $v: R^+ \times R^2 \to R$ given by $v(t, x) = x_1^2/(1 + t) + x_2^2$ is decrescent and positive semidefinite but not positive definite.

10. The function $v: R^+ \times R^2 \to R$ given by $v(t, x) = (x_2 - x_1)^2(1 + t)$ is positive semidefinite but not positive definite or decrescent.

Of special interest are functions $v: R^n \to R$ that are **quadratic forms** given by:

$$v(x) = x^{\mathrm{T}} B x = \sum_{i,k=1}^{n} b_{ik} x_i x_k \tag{105}$$

where $B = [b_{ij}]$ is a real symmetric $n \times n$ matrix (i.e., $B^T = B$). Since B is symmetric, it is diagonizable and all of its eigenvalues are real. For Eq. (105) one can prove the following:

1. *v is positive definite (and radially unbounded) if and only if all principal minors of B are positive; that is, if and only if*

$$\det \begin{bmatrix} b_{11} & \cdots & b_{1k} \\ \vdots & & \vdots \\ b_{k1} & \cdots & b_{kk} \end{bmatrix} > 0, \qquad k = 1, \ldots, n$$

2. *v is negative definite if and only if*

$$(-1)^k \det \begin{bmatrix} b_{11} & \cdots & b_{1k} \\ \vdots & & \vdots \\ b_{k1} & \cdots & b_{kk} \end{bmatrix} > 0, \qquad k = 1, \ldots, n$$

3. *v is* **definite** *(i.e., either positive definite or negative definite) if and only if all eigenvalues are nonzero and have the same sign.*

4. *v is* **semidefinite** *(i.e., either positive semidefinite or negative semidefinite) if and only if the nonzero eigenvalues of B have the same sign.*

5. *If λ_m and λ_M denote the smallest and largest eigenvalues of B and if $|x|$ denotes the Euclidean norm of x, then $\lambda_m |x|^2 \leqslant v(x) \leqslant \lambda_M |x|^2$ for all $x \in R^n$.* (The Euclidean norm of x is defined as $(x^{\mathrm{T}}x)^{1/2} = (\sum_{i=1}^{n} x_i^2)^{1/2}$.)

6. *v is* **indefinite** *(i.e., in every neighborhood of the origin $x = 0$, v assumes positive and negative values) if and only if B possesses both positive and negative eigenvalues.*

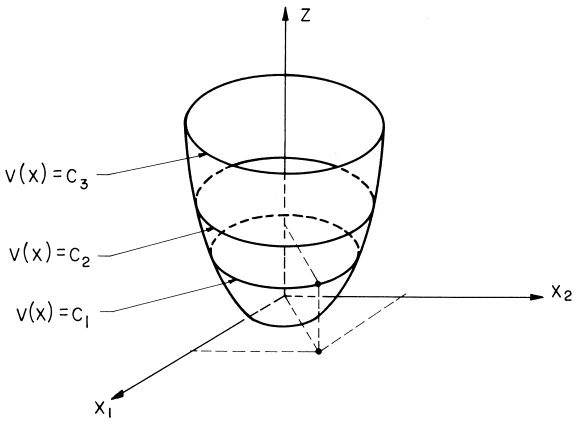


FIGURE 19 Surface described by a quadratic form.

Quadratic forms (105) have some interesting geometric properties. To see this, let $n = 2$, and assume that both eigenvalues of B are positive so that v is positive definite and radially unbounded. In R^3, let us now consider the surface determined by:

$$z = v(x) = x^{\mathrm{T}} B x \tag{106}$$

This equation describes a cup-shaped surface as depicted in Fig. 19. Note that corresponding to every point on this cup-shaped surface there exists one and only one point in the x_1x_2 plane. Note also that the loci defined by $C_i = \{x \in R^2: v(x) = c_i \geqslant 0\}$, $c_i =$ const, determine closed curves in the x_1x_2 plane as shown in Fig. 20. We call these curves **level curves**. Note that $C_0 = \{0\}$ corresponds to the case in which $z = c_0 = 0$. Note also that this function v can be used to *cover the entire* R^2 plane with closed curves by selecting for z all values in R^+.

In the case when $v = x^{\mathrm{T}} B x$ is a positive definite quadratic form with $x \in R^n$, the preceding comments are still true; however, in this case, the closed curves C_i must be replaced by closed hypersurfaces in R^n and a simple geometric visualization as in Figs. 19 and 20 is no longer possible.

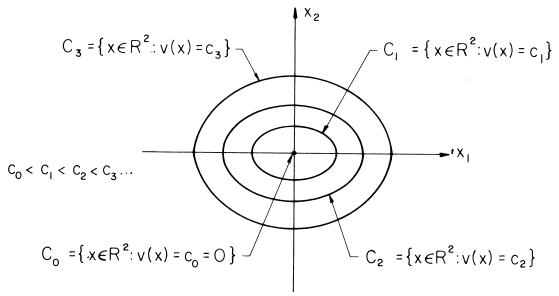


FIGURE 20 Level curves determined by a quadratic form.

G. Lyapunov Stability and Instability Results: Motivation

Before we summarize the principal Lyapunov-type of stability and instability results, we give a geometric interpretation of some of these results in R^2. To this end, we consider the system of equations,

$$x_1' = f_1(x_1, x_2), \qquad x_2' = f_2(x_1, x_2) \tag{107}$$

and we assume that f_1 and f_2 are such that for every (t_0, x_0), $t_0 \geqslant 0$, Eq. (107) has a unique solution $\phi(t, t_0, x_0)$ with $\phi(t_0, t_0, x_0) = x_0$. We also assume that $(x_1, x_2)^{\mathrm{T}} = (0, 0)^{\mathrm{T}}$ is the only equilibrium in $B(h)$ for some $h > 0$.

Next, let v be a positive definite, continuously differentiable function with nonvanishing gradient ∇v on $0 < |x| \leqslant h$. Then, $v(x) = c$, $c \geqslant 0$, defines for sufficiently small constants $c > 0$ a family of closed curves C_i which cover the neighborhood $B(h)$ as shown in Fig. 21. Note that the origin $x = 0$ is located in the interior of each such curve and in fact, $C_0 = \{0\}$.

Now suppose that all trajectories of Eq. (107) originating from points on the circular disk $|x| \leq r_1 < h$ cross the curves $v(x) = c$ from the exterior toward the interior when we proceed along these trajectories in the direction of increasing values of t. Then, we can conclude that these trajectories approach the origin as t increases (i.e., the equilibrium $x = 0$ is in this case asymptotically stable).

In terms of the given v function, we have the following interpretation. For a given solution $\phi(t, t_0, x_0)$ to cross the curve $v(x) = r$, $r = v(x_0)$, the angle between the outward normal vector $\nabla v(x_0)$ and the derivative of $\phi(t, t_0, x_0)$ at $t = t_0$ must be greater than $\pi/2$; that is,

$$v_{(107)}'(x_0) = \nabla v(x_0) f(x_0) < 0$$

For this to happen at all points, we must have $v_{(107)}'(x) < 0$ for $0 < |x| \leq r_1$. The same results can be arrived at from an analytic point of view. The function $V(t) = v(\phi(t, t_0, x_0))$ decreases monotonically as t increases. This implies that the derivative $v'(\phi(t, t_0, x_0))$ along the solution $(\phi(t, t_0, x_0))$ must be negative definite in $B(r)$ for $r > 0$ sufficiently small.

Next, let us assume that, Eq. (107) has only one equilibrium (at $x = 0$) and that v is positive definite and radially unbounded. It turns out that in this case, the relation $v(x) = c$, $c \in R^+$, can be used to cover *all* of R^2 by closed curves of the type shown in Fig. 21. If for arbitrary (t_0, x_0), the corresponding solution of Eq. (107), $\phi(t, t_0, x_0)$, behaves as already discussed, then it follows that the derivative of v along this solution, $v'(\phi(t, t_0, x_0))$, will be negative definite in R^2.

Since the foregoing discussion was given in terms of an arbitrary solution of Eq. (107), we may suspect that the following results are true:

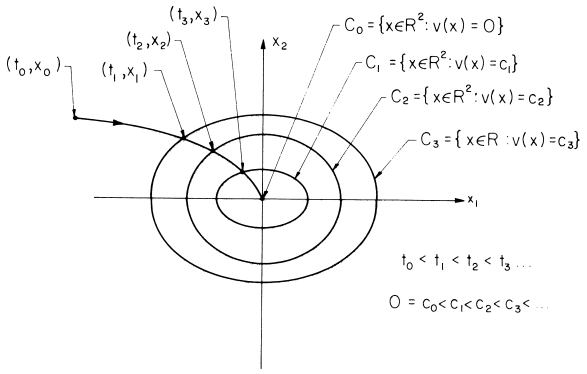


FIGURE 21 Trajectory near an asymptotically stable equilibrium point.

1. *If there exists a positive definite function v such that $v_{(107)}'$ is negative definite, then the equilibrium $x = 0$ of Eq.* (107) *is asymptotically stable.*
2. *If there exists a positive definite and radially unbounded function v such that $v_{(107)}'$ is negative definite for all $x \in R^2$, then the equilibrium $x = 0$ of Eq.* (107) *is asymptotically stable in the large.*

Continuing our discussion by making reference to Fig. 22, let us assume that we can find for Eq. (107) a continuously differentiable function $v: R^2 \to R$ which is indefinite and has the properties discussed below. Since v is indefinite, there exist in each neighborhood of the origin points for which $v > 0$, $v < 0$, and $v(0) = 0$. Confining our attention to $B(k)$, where $k > 0$ is sufficiently small, we let $D = \{x \in B(k): v(x) < 0\}$. ($D$ may consist of several subdomains.) The boundary of D, ∂D, as shown in Fig. 22, consists of points in $\partial B(k)$ and of points determined by $v(x) = 0$. Assume that in the interior of D, v is bounded. Suppose $v_{(107)}'(x)$ is negative definite in D

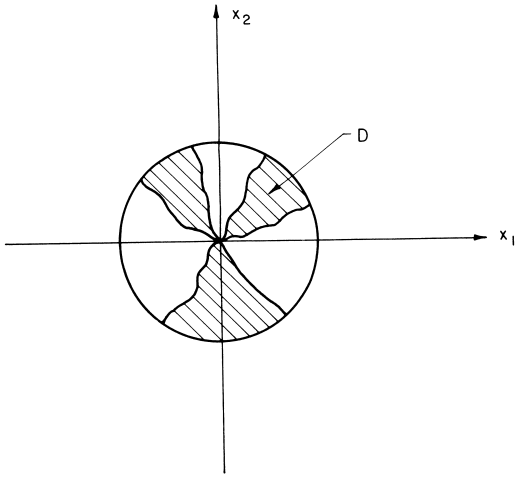


FIGURE 22 Instability of an equilibrium point.

and that $x(t)$ is a trajectory of Eq. (107) which originates somewhere on the boundary of D ($x(t_0) \in \partial D$) with $v(x(t_0)) = 0$. Then, this trajectory will penetrate the boundary of D at points where $v = 0$ as t increases and it can never again reach a point where $v = 0$. In fact, as t increases, this trajectory will penetrate the set of points determined by $|x| = k$ (since; by assumption, $v'_{(107)} < 0$ along this trajectory and $v < 0$ in D). But, this indicates that the equilibrium $x = 0$ of Eq. (107) is unstable.

We are once more led to a conjecture:

3. *Let a function* $v\colon R^2 \to R$ *be given which is continuously differentiable and which has the following properties*: (a) *There exist points x arbitrarily close to the origin such that* $v(x) < 0$; *they form the domain D bounded by the set of points determined by* $v = 0$ *and the disk* $|x| = k$. (b) *In the interior of D*, v *is bounded.* (c) *In the interior of D*, $v'_{(107)}$ *is negative. Then*, *the equilibrium* $x = 0$ *of Eq.* (107) *is unstable*.

H. Principal Lyapunov Stability and Instability Theorems

It turns out that results of the type given above for Eq. (107) are true for general systems (E). These results, which are proved by standard δ–ε arguments, comprise the **direct method of Lyapunov**, which is also sometimes called the **second method of Lyapunov**. The reason for this nomenclature is clear: Results of the kind presented here allow us to make qualitative statements about whole families of solutions of (E), without actually solving this equation.

In the following, we enumerate some of the more important results of the direct method. We shall assume that $v\colon R^+ \times B(h) \to R$ (resp., $v\colon R^+ \times R^n \to R$).

1. *If there exists a continuously differentiable positive definite function* v *with a negative semidefinite (or identically zero) derivative* $v'_{(E)}$, *then the equilibrium* $x = 0$ *of* (E) *is stable*.

As an example, consider the system given by:

$$x'_1 = x_2, \qquad x'_2 = -x_2 - e^{-t}x_1 \tag{108}$$

which has an equilibrium at $(x_1, x_2)^T = (0, 0)^T$. For Eq. (108), choose the positive definite function $v(t, x_1, x_2) = x_1^2 + e^t x_2^2$. We obtain $v'_{(108)}(t, x_1, x_2) = -e^t x_2^2$ which is negative semidefinite. We conclude that the equilibrium $x = 0$ of Eq. (108) is *stable*.

2. *If there exists a continuously differentiable, positive definite, decrescent function* v *with negative semidefinite derivative* $v'_{(E)}$, *then the equilibrium* $x = 0$ *of* (E) *is uniformly stable*.

As an example, consider the simple pendulum,

$$x'_1 = x_2, \qquad x'_2 = -k \sin x_1 \tag{109}$$

where $k > 0$ is a constant. As noted earlier, Eq. (109) has an isolated equilibrium at $x = 0$. Choose $v(x_1, x_2) = \frac{1}{2}x_2^2 + k\int_0^{x_1} \sin\eta\, d\eta$, which is continuously differentiable and positive definite. Also, since v does not depend on t, it will automatically be decrescent. Furthermore, $v'_{(109)}(x_1, x_2) = (k \sin x_1)x'_1 + x_2 x'_2 = (k \sin x_1)x_2 + x_2(-k \sin x_1) = 0$. Therefore, the equilibrium $x = 0$ of Eq. (109) is *uniformly stable*.

3. *If there exists a continuously differentiable, positive definite*, *decrescent function* v *with a negative definite derivative* $v'_{(E)}$, *then the equilibrium* $x = 0$ *of* (E) *is uniformly asymptotically stable*.

For an example, the system,

$$\begin{aligned} x'_1 &= (x_1 - c_2 x_2)\big(x_1^2 + x_2^2 - 1\big) \\ x'_2 &= (c_1 x_1 + x_2)\big(x_1^2 + x_2^2 - 1\big) \end{aligned} \tag{110}$$

has an isolated equilibrium at the origin $x = 0$. Choosing $v(x) = c_1 x_1^2 + c_2 x_2^2$, we obtain $v'_{(110)}(x) = 2(c_1 x_1^2 + c_2 x_2^2)(x_1^2 + x_2^2 - 1)$. If $c_1 > 0, c_2 > 0$, then v is positive definite (and decrescent) and $v'_{(110)}$ is negative definite in the domain $x_1^2 + x_2^2 < 1$. Therefore, the equilibrium $x = 0$ *of* Eq. (110) *is uniformly asymptotically stable*.

4. *If there exists a continuously differentiable*, *positive definite*, *decrescent*, *and radially unbounded function* v *such that* $v'_{(E)}$ *is negative definite for all* $(t, x) \in R^+ \times R^n$, *then the equilibrium* $x = 0$ *of* (E) *is uniformly asymptotically stable in the large*.

As an example, consider the system,

$$\begin{aligned} x'_1 &= x_2 + c x_1\big(x_1^2 + x_2^2\big) \\ x'_2 &= -x_1 + c x_2\big(x_1^2 + x_2^2\big) \end{aligned} \tag{111}$$

where c is a real constant. Note that $x = 0$ is the only equilibrium. Choosing the positive definite, decrescent, and radially unbounded function $v(x) = x_1^2 + x_2^2$, we obtain $v'_{(111)}(x) = 2c(x_1^2 + x_2^2)^2$. We conclude that if $c = 0$, then $x = 0$ of Eq. (111) is *uniformly stable* and if $c < 0$, then $x = 0$ of Eq. (111) is *uniformly asymptotically stable in the large*.

5. *If there exists a continuously differentiable function* v *and three positive constants* c_1, c_2, *and* c_3 *such that*

$$c_1|x|^2 \le v(t, x) \le c_2|x|^2$$
$$v'_{(E)}(t, x) \le -c_3|x|^2$$

for all $t \in R^+$ *and for all* $x \in B(r)$ *for some* $r > 0$, *then the equilibrium* $x = 0$ *of* (E) *is exponentially stable*.

6. *If there exist a continuously differentiable function* v *and three positive constants* c_1, c_2, *and* c_3 *such that*

$$c_1|x|^2 \le v(t,x) \le c_2|x|^2$$
$$v'_{(E)}(t,x) \le -c_3|x|^2$$

for all $t \in R^+$ *and for all* $x \in R^n$, *then the equilibrium* $x=0$ *of* (E) *is exponentially stable in the large*.

As an example, consider the system,

$$\begin{aligned} x_1' &= -a(t)x_1 - bx_2 \\ x_2' &= bx_1 - c(t)x_2 \end{aligned} \tag{112}$$

where b is a real constant and where a and c are real and continuous functions defined for $t \ge 0$ satisfying $a(t) \ge \delta > 0$ and $c(t) \ge \delta > 0$ for all $t \ge 0$. We assume that $x=0$ is the only equilibrium for Eq. (112). If we choose $v(x) = \frac{1}{2}(x_1^2 + x_2^2)$, then $v'_{(112)}(t,x) = -a(t)x_1^2 - c(t)x_2^2 \le -\delta(x_1^2 + x_2^2)$. Hence, the equilibrium $x=0$ of Eq. (112) is *exponentially stable in the large.*

7. *If there exists a continuously differentiable function* v *defined on* $|x| \ge R$ (*where* R *may be large*) *and* $0 \le t \le \infty$, *and if there exist* $\psi_1, \psi_2 \in KR$ *such that* $\psi_1(|x|) \le v(t,x) \le \psi_2(|x|)$, $v'_{(E)}(t,x) \le 0$ *for all* $|x| \ge R$ *and for all* $0 \le t < \infty$, *then the solutions of* (E) *are uniformly bounded.*

8. *If there exists a continuously differentiable function* v *defined on* $|x| \ge R$ (*where* R *may be large*) *and* $0 \le t < \infty$, *and if there exist* $\psi_1, \psi_2 \in KR$ *and* $\psi_3 \in K$ *such that* $\psi_1(|x|) \le v(t,x) \le \psi_2(|x|)$, $v'_{(E)}(t,x) \le -\psi_3(|x|)$ *for all* $|x| \ge R$ *and* $0 \le t < \infty$, *then the solutions of* (E) *are uniformly ultimately bounded.*

As an example, consider the system,

$$x' = -x - \sigma, \qquad \sigma' = -\sigma - f(\sigma) + x \tag{113}$$

where $f(\sigma) = \sigma(\sigma^2 - 6)$. There are isolated equilibrium points at $x = \sigma = 0$, $x = -\sigma = 2$, and $x = -\sigma = -2$. Choosing the radially unbounded and decrescent function $v(x,\sigma) = \frac{1}{2}(x^2 + \sigma^2)$, we obtain $v'_{(113)} = (x,\sigma) = -x^2 - \sigma^2(\sigma^2 - 5) \le -x^2 - (\sigma^2 - \frac{5}{2})^2 + \frac{25}{4}$. Also $v'_{(113)}$ is negative for all (x,σ) such that $x^2 + \sigma^2 > R^2$, where, for example, $R = 10$ will do. Therefore, all solutions of Eq. (113) are *uniformly bounded* and, in fact, *uniformly ultimately bounded.*

9. *The equilibrium* $x=0$ *of* (E) *is unstable* (*at* $t = t_0 \ge 0$) *if there exists a continuously differentiable, decrescent function* v *such that* $v'_{(E)}$ *is positive definite (negative definite) and if in every neighborhood of the origin there are points* x *such that* $v(t_0, x) > 0 (v(t_0, x) < 0)$.

Reconsider system (111), this time assuming that $c > 0$. If we choose $v(x) = x_1^2 + x_2^2$, then $v'_{(111)}(x) = 2c(x_1^2 + x_2^2)^2$ and we can conclude from the above result that the equilibrium $x=0$ of (E) is *unstable*.

10. *Let there exist a bounded and continuously differentiable function* $v: D \to R$, $D = \{(t,x) \ge t_0, x \in B(h)\}$, *with the following properties*: (a) $v'_{(E)}(t,x) = \lambda v(t,x) + w(t,x)$, *where* $\lambda > 0$ *is a constant and* $w(t,x)$ *is either identically zero or positive semidefinite*; (b) *in the set* $D_1 = \{(t,x): t = t_1, x \in B(h_1)\}$ *for fixed* $t_1 \ge t_0$ *and with arbitrarily small* h_1, *there exist values* x *such that* $v(t_1, x) > 0$. *Then the equilibrium* $x=0$ *of* (E) is *unstable*.

As a specific example, consider:

$$\begin{aligned} x_1' &= x_1 + x_2 + x_1 x_2^4 \\ x_2' &= x_1 + x_2 - x_1^2 x_2 \end{aligned} \tag{114}$$

which has an isolated equilibrium $x=0$. Choosing $v(x) = (x_1^2 - x_2^2)/2$, we obtain $v'_{(114)}(x) = \lambda v(x) + w(x)$, where $w(x) = x_1^2 x_2^4 + x_1^2 x_2^2$ and $\lambda = 2$. It follows from the above result that the equilibrium $x=0$ of Eq. (114) is *unstable*.

11. *Let there exist a continuously differentiable function* v *having the following properties*: (a) *For every* $\varepsilon > 0$ *and for every* $t \ge 0$, there exist points $\bar{x} \in B(\varepsilon)$ *such that* $v(t, \bar{x}) < 0$. *We call the set of all points* (t,x) *such that* $x \in B(h)$ *and such that* $v(t,x) < 0$ *the "domain* $v < 0$." *It is bounded by the hypersurfaces which are determined by* $|x| = h$ *and* $v(t,x) = 0$ *and it may consist of several component domains.* (b) *In at least one of the component domains* D *of the domain* $v < 0$, v *is bounded from below and* $0 \in \partial D$ *for all* $t \ge 0$. (c) *In the domain* D, $v'_{(E)} \le -\Psi(|v|)$, *where* $\psi \in K$. *Then, the equilibrium* $x=0$ *of* (E) *is unstable*.

As an example, consider the system,

$$\begin{aligned} x_1' &= x_1 + x_2 \\ x_2' &= x_1 - x_2 + x_1 x_2 \end{aligned} \tag{115}$$

which has an isolated equilibrium at the origin $x=0$. Choosing $v(x) = -x_1 x_2$, we obtain $v'_{(115)}(x) = -x_1^2 - x_2^2 - x_1^2 x_2$. Let $D = \{x \in R^2: x_1 > 0, x_2 > 0, \text{ and } x_1^2 + x_2^2 < 1\}$. Then, for all $x \in D$, $v < 0$ and $c'_{(115)} < 2v$. We see that the above result is applicable and conclude that the equilibrium $x=0$ of Eq. (115) is *unstable*.

The results given in items 1–11 are also true when v is continuous (rather than continuously differentiable). In this case, $v'_{(E)}$ must be interpreted in the sense of Eq. (104).

For the case of systems (A),

$$x' = f(x) \tag{A}$$

it is sometimes possible to relax the conditions on $v'_{(A)}$ when investigating the asymptotic stability of the equilibrium $x=0$, by insisting that $v'_{(A)}$ be only negative semidefinite. In doing so, we require the following concept: A set Γ of points in R^n is **invariant (with respect to** (A)) if

every solution of (A) starting in Γ remains in Γ for all time.

We are now in a position to state our next result which is part of **invariance theory** for ordinary differential equations.

12. *Assume that there exists a continuously differentiable, positive definite, and radially unbounded function* $v: R^n \to R$ *such that* (a) $v'_{(A)}(x) \leq 0$ *for all* $x \in R^n$, *and* (b) *the origin* $x = 0$ *is the only invariant subset of the set* $E = \{x \in R^n: v'_{(A)}(x) = 0\}$. *Then, the equilibrium* $x = 0$ *of* (A) *is asymptotically stable in the large.*

As a specific example, consider the *Liénard equation* given by:

$$x'_1 = x_2, \qquad x'_2 = -f(x_1)x_2 - g(x_1) \tag{116}$$

where it is assumed that f and g are continuously differentiable for all $x_1 \in R$, $g(x_1) = 0$ if and only if $x_1 = 0$, $x_1 g(x_1) > 0$ for all $x_1 \neq 0$ and $x_1 \in R$,

$$\lim_{|x_1| \to \infty} \int_0^{x_1} g(\eta)\, d\eta = \infty$$

and $f(x_1) > 0$ for all $x_1 \in R$. Then $(x_1, x_2) = (0, 0)$ is the only equilibrium of Eq. (116). Let us choose the v function,

$$v(x_1, x_2) = \frac{1}{2}x_2^2 + \int_0^{x_1} g(\eta)\, d\eta$$

(for Eq. (116)) which is positive definite and radially unbounded. Along the solutions of Eq. (116) we have $v'_{(116)}(x_1, x_2) = -x_2^2 f(x_1) \leq 0$ for all $(x_1, x_2) \in R^2$. It is easily verified that the set E in the above theorem is in our case the x_1 axis. Furthermore, a moment's reflection shows that the largest invariant subset (with respect to Eq. (116)) of the x_1 axis is the set $\{(0, 0)^T\}$. Thus, by the above result, the origin $x = 0$ of Eq. (116) is *asymptotically stable in the large*.

I. Linear Systems Revisited

One of the great drawbacks of the Lyapunov theory, as developed above, is that there exist no general rules of choosing v functions, which are called **Lyapunov functions**. However, for the case of linear systems,

$$x' = Ax \tag{L}$$

it is possible to construct Lyapunov functions, as shown by the following result:

Assume that the matrix A has no eigenvalues on the imaginary axis. Then, there exists a Lyapunov function v of the form:

$$v(x) = x^T B x, \qquad B = B^T \tag{117}$$

whose derivative $v'_{(L)}$, *given by*:

$$v'_{(L)}(x) = -x^T C x$$

where

$$-C = A^T B + BA, \qquad C = C^T \tag{118}$$

is definite (*i.e., negative definite or positive definite*).

The above result shows that if A is a stable matrix, then for (L), our earlier Lyapunov result for asymptotic stability in the large constitutes also necessary conditions for asymptotic stability. Also, if A is an unstable matrix with no eigenvalues on the imaginary axis, then according to the above result, our earlier instability result given in item 9 above yields also necessary conditions for instability.

In view of the above result, the v function in Eq. (117) is easily constructed by assuming a definite matrix C (either positive definite or negative definite) and by solving the **Lyapunov matrix equation** (118) for the $n(n+1)/2$ unknown elements of the symmetric matrix B.

J. Domain of Attraction

Next, we should address briefly the problem of estimating the domain of attraction of an equilibrium $x = 0$. Such questions are important when $x = 0$ is not the only equilibrium of a system or when $x = 0$ is not asymptotically stable in the large.

For purposes of discussion, we consider a system,

$$x' = f(x) \tag{A}$$

and we assume that for (A) there exists a Lyapunov function v which is *positive definite* and *radially unbounded*. Also, we assume that over some domain $D \subset R^n$ containing the origin, $v'_{(A)}(x)$ is *negative*, except at the origin, where $v'_{(A)} = 0$. Now let C_i denote $C_i = \{x \in R^n: v(x) \leq c_i\}$, $c_i > 0$. Using similar reasoning as was done in connection with Eq. (107), we can show that as long as $C_i \subset D$, C_i will be a subset of the domain of attraction of $x = 0$. Thus, if $c_i > 0$ is the largest number for which this is true, then it follows that C_i will be contained in the domain of attraction of $x = 0$. The set C_i will be the best estimate that we can obtain for our particular choice of v function.

K. Converse Theorems

Above we showed that for system (L) there exist actually converse Lyapunov (stability and instability) theorems. It turns out that for virtually every result which we gave above (in items 1–11) there exists a converse. Unfortunately, these **Lyapunov converse theorems** are not much help in constructing v functions in specific cases. For purposes of illustration, we cite here an example of such a converse theorem:

If f and $f_x = \partial f/\partial x$ are continuous on the set $(R^+ \times B(r))$ for some $r > 0$, and if the equilibrium $x = 0$ of (E) *is uniformly asymptotically stable, then there exists a Lyapunov function v which is continuously differentiable on $(R^+ \times B(r_1))$ for some $r_1 > 0$ such that v is positive definite and decrescent and such that $v'_{(E)}$ is negative definite.*

L. Comparison Theorems

Next, we consider once more some comparison results for (E), as was done in Section III. We shall assume that $f: R^+ \times B(r) \to R^n$ for some $r_1 > 0$ and that f is continuous there. We begin by considering a scalar ordinary differential equation of the form,

$$y' = G(t, y) \tag{$\tilde{C}$}$$

where $y \in R$, $t \in R^+$, and $F: R^+ \times [0, r) \to R$ for some $r > 0$. Assume that G is continuous on $R^+ \times [0, r)$ and that $G(t, 0) = 0$ for all t. Also, assume that $y = 0$ is an isolated equilibrium for ($\tilde{C}$).

The following results are the basis of the **comparison principle** in the stability analysis of the isolated equilibrium $x = 0$ of (E):

Let f and G be continuous on their respective domains of definition. Let $v: R^+ \times B(r) \to R$ be a continuously differentiable, positive-definite function such that

$$v'_{(E)}(t, x) \leq G(t, v(t, x)) \tag{119}$$

Then, the following statements are true: (1) *If the trivial solution of* ($\tilde{C}$) *is stable, then the trivial solution of system* (E) *is stable.* (2) *If v is decrescent and if the trivial solution of* ($\tilde{C}$) *is uniformly stable, then the trivial solution of* (E) *is uniformly stable.* (3) *If v is decrescent and if the trivial solution of* ($\tilde{C}$) *is uniformly asymptotically stable, then the trivial solution of* (E) *is uniformly asymptotically stable.* (4) *If there are constants $a > 0$ and $b > 0$ such that $a|x|^b \leq v(t, x)$, if v is decrescent, and if the trivial solution of* ($\tilde{C}$) *is exponentially stable, then the trivial solution of* (E) *is exponentially stable.* (5) *If $f: R^+ \times R^n \to R^n$, $G: R^+ \times R \to R$, $v: R^+ \times R^n \to R$ is decrescent and radially unbounded, if Eq.* (119) *holds for all $t \in R^+$, $x \in R^n$, and if the solutions of* ($\tilde{C}$) *are uniformly bounded* (*uniformly ultimately bounded*), *then the solutions of* (E) *are also uniformly bounded* (*uniformly ultimately bounded*).

The above results enable us to analyze the stability and boundedness properties of an n-dimensional system (E), which may be complex, in terms of the corresponding properties of a one-dimensional comparison system ($\tilde{C}$), which may be quite a bit simpler. The generality and effectiveness of the above results can be improved and extended by considering *vector-valued comparison equations and vector Lyapunov functions*.

M. Lyapunov's First Method

We close this section by answering the following question: Under what conditions does it make sense to linearize a nonlinear system about an equilibrium $x = 0$ and then deduce the properties of $x = 0$ from the corresponding linear system? This is known as **Lyapunov's first method** or **Lyapunov's indirect method.**

We consider systems of n real nonlinear first-order ordinary differential equations of the form:

$$x' = Ax + F(t, x) \tag{PE}$$

where $F: R^+ \times B(h) \to R^n$ for some $h > 0$ and A is a real $n \times n$ matrix. Here, we assume that Ax constitutes the **linear part** of the right-hand side of (PE) and $F(t, x)$ represents the remaining terms which are of order higher than one in the various components of x. Such systems may arise in the process of linearizing nonlinear equations of the form:

$$x' = g(t, x) \tag{G}$$

or they may arise in some other fashion during the modeling process of a physical system.

To be more specific, let $g: R \times D \to R^n$ where D is some domain in R^n. If g is continuously differentiable on $R \times D$ and if ϕ is a given solution of (E) defined for all $t \geq t_0 \geq 0$, then we can **linearize** (G) **about** ϕ in the following manner. Define $y = x - \phi(t)$ so that

$$\begin{aligned} y' &= g(t, x) - g(t, \phi(t)) \\ &= g(t, y + \phi(t)) - g(t, \phi(t)) \\ &= (\partial g/\partial t)(t, \phi(t))y + G(t, y) \end{aligned}$$

Here,

$$G(t, y) = [g(t, y+\phi(t)) - g(t, \phi(t))] - (\partial g/\partial x)(t, \phi(t))y$$

is $o(|y|)$ as $|y| \to 0$ uniformly in t on compact subsets of $[t_0, \infty)$.

Of special interest is the case when g is independent of t (i.e., when $g(t, x) \equiv g(x)$) and $\phi(t) = \xi_0$ is a constant (equilibrium point). Under these conditions we have

$$y' = Ay + G(y)$$

where $A = (\partial g/\partial x)(x)|_{x=\xi_0}$, where $(\partial g/\partial x)(x)$ denotes the Jacobian of $g(x)$.

By making use of the result for the Lyapunov function (117), we can readily prove the following results:

1. *Let A be a real, constant, and stable $n \times n$ matrix and let $F: R^+ \times B(h) \to R^n$ be continuous in (t, x) and satisfy $F(t, x) = o(|x|)$ as $|x| \to 0$, uniformly in $t \in R^+$.*

Then, the trivial solution of (PE) *is uniformly asymptotically stable.*

As a specific example, consider the *Liénard equation* given by:

$$x'' + f(x)x' + x = 0 \tag{120}$$

where $f: R \to R$ is a continuous function with $f(0) > 0$. We can rewrite Eq. (120) as:

$$x_1' = x_2, \qquad x_2' = -x_1 - f(0)x_2 + [f(0) - f(x_1)]x_2$$

and we can apply the above result with $x = (x_1, x_2)^{\mathrm{T}}$,

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -f(0) \end{bmatrix}$$

$$F(t, x) = \begin{bmatrix} 0 \\ [f(0) - f(x_1)]x_2 \end{bmatrix}$$

Noting that A is a stable matrix and that $F(t, x) = o(|x|)$ as $|x| \to 0$, uniformly in $t \in R^+$, we conclude that the trivial solution $(x, x') = (0, 0)$ of Eq. (120) is uniformly asymptotically stable.

2. *Assume that A is a real $n \times n$ matrix with no eigenvalues on the imaginary axis and that at least one eigenvalue of A has positive real part. If $F: R^+ \times B(h) \to R^n$ is continuous and satisfies $F(t, x) = o(|x|)$ as $|x| \to 0$, uniformly in $t \in R^+$, then the trivial solution of* (PE) *is unstable.*

As a specific example, consider the simple pendulum,

$$x'' + k \sin x = 0, \qquad k > 0 \tag{121}$$

Note that $x_e = \pi, x_e' = 0$ is an equilibrium of Eq. (121). Let $y = x - x_e$ so that

$$y'' + a \sin(y + \pi) = y'' - ay + a(\sin(y + \pi) + y) = 0$$

This equation can be put into the form (PE) with

$$A = \begin{bmatrix} 0 & 1 \\ a & 0 \end{bmatrix}$$

$$F(t, x) = \begin{bmatrix} 0 \\ a(\sin(y + \pi) + y) \end{bmatrix}$$

Applying the above result we conclude that the equilibrium point $(\pi, 0)$ is *unstable*.

SEE ALSO THE FOLLOWING ARTICLES

ARTIFICIAL NEURAL NETWORKS • CALCULUS • COMPLEX ANALYSIS • DIFFERENTIAL EQUATIONS, PARTIAL • FUNCTIONAL ANALYSIS • MEASURE AND INTEGRATION

BIBLIOGRAPHY

Antsaklis, P. J., and Michel, A. N. (1997). "Linear Systems," McGraw-Hill, New York.

Boyce, W. E., and DiPrima, R. C. (1997). "Elementary Differential Equations and Boundary Value Problems," John Wiley & Sons, New York.

Brauer, F., and Nohel, J. A. (1969). "Qualitative Theory of Ordinary Differential Equations," Benjamin, New York.

Carpenter, G. A., Cohen, M., and Grossberg, S. (1987). "Computing with neural networks," *Science* **235,** 1226–1227.

Coddington, E. A., and Levinson, N. (1955). "Theory of Ordinary Differential Equations," McGraw-Hill, New York.

Hale, J. K. (1969). "Ordinary Differential Equations," Wiley, New York.

Halmos, P. R. (1958). "Finite Dimensional Vector Spaces," Van Nostrand, Princeton, NJ.

Hille, E. (1969). "Lectures on Ordinary Differential Equations," Addison-Wesley, Reading, MA.

Hoffman, K., and Kunze, R. (1971). "Linear Algebra," Prentice-Hall, Englewood Cliffs, NJ.

Hopfield, J. J. (1984). "Neurons with graded response have collective computational properties like those of two-state neurons," *Proc. Nat. Acad. Sci. U.S.A.* **81,** 3088–3092.

Kantorovich, L. V., and Akilov, G. P. (1964). "Functional Analysis in Normed Spaces," Macmillan, New York.

Michel, A. N. (1983). "On the status of stability of interconnected systems," *IEEE Trans. Automat. Control* **28**(6), 639–653.

Michel, A. N., and Herget, C. J. (1993). "Applied Algebra and Functional Analysis," Dover, New York.

Michel, A. N., and Miller, R. K. (1977). "Qualitative Analysis of Large-Scale Dynamical Systems," Academic Press, New York.

Michel, A. N., and Wang, K. (1995). "Qualitative Theory of Dynamical Systems," Dekker, New York.

Michel, A. N., Farrell, J. A., and Porod, W. (1989). "Qualitative analysis of neural networks," *IEEE Trans. Circuits Syst.* **36**(2), 229–243.

Miller, R. K., and Michel, A. N. (1982). "Ordinary Differential Equations," Academic Press, New York.

Naylor, A. W., and Sell, G. R. (1971). "Linear Operator Theory in Engineering and Science," Holt, Rinehart & Winston, New York.

Royden, H. L. (1963). "Real Analysis," Macmillan, New York.

Rudin, W. (1953). "Principles of Mathematical Analysis," McGraw-Hill, New York.

Simmons, G. F. (1972). "Differential Equations," McGraw-Hill, New York.



Differential Equations, Partial

Martin Schechter
University of California, Irvine

I. Importance
II. How They Arise
III. Some Well-Known Equations
IV. Types of Equations
V. Problems Associated with Partial Differential Equations
VI. Methods of Solution

GLOSSARY

Boundary Set of points in the closure of a region not contained in its interiors.
Bounded region Region that is contained in a sphere of finite radius.
Eigenvalue Scalar λ for which the equation $Au = \lambda u$ has a nonzero solution u.
Euclidean *n* dimensional space $\mathbb{R}^n$ Set of vectors $x = (x_1, \ldots, x_n)$ where each component x_j is a real number.
Partial derivative Derivative of a function of more than one variable with respect to one of the variables keeping the other variables fixed.

A PARTIAL DIFFERENTIAL EQUATION is an equation in which a partial derivative of an unknown function appears. The order of the equation is the highest order of the partial derivatives (of an unknown function) appearing in the equation. If there is only one unknown function $u(x_1, \ldots, x_n)$, then a partial differential equation for u is of the form:

$$F\left(x_1, \ldots, x_n, u, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n}, \frac{\partial^2 u}{\partial x_1^2}, \ldots, \frac{\partial^k u}{\partial x_1^k}, \ldots\right) = 0$$

One can have more than one unknown function and more than one equation involving some or all of the unknown functions. One then has a system of j partial differential equations in k unknown functions. The number of equations may be more or less than the number of unknown functions. Usually it is the same.

I. IMPORTANCE

One finds partial differential equations in practically every branch of physics, chemistry, and engineering. They are also found in other branches of the physical sciences and in the social sciences, economics, business, etc. Many parts of theoretical physics are formulated in terms of partial differential equations. In some cases, the axioms require that the states of physical systems be given by solutions of partial differential equations. In other cases, partial differential equations arise when one applies the axioms to specific situations.

II. HOW THEY ARISE

Partial differential equations arise in several branches of mathematics. For instance, the Cauchy–Riemann equations,

$$\frac{\partial u(x, y)}{\partial x} = \frac{\partial v(x, y)}{\partial y}, \qquad \frac{\partial u(x, y)}{\partial y} = -\frac{\partial v(x, y)}{\partial x}$$

must be satisfied if

$$f(z) = u(x, y) + iv(x, y)$$

is to be an analytic function of the complex variable $z = x + iy$. Thus, the rich and beautiful branch of mathematics known as analytic function theory is merely the study of solutions of a particular system of partial differential equations.

As a simple example of a partial differential equation arising in the physical sciences, we consider the case of a vibrating string. We assume that the string is a long, very slender body of elastic material that is flexible because of its extreme thinness and is tightly stretched between the points $x = 0$ and $x = L$ on the x axis of the x,y plane. Let x be any point on the string, and let $y(x, t)$ be the displacement of that point from the x axis at time t. We assume that the displacements of the string occur in the x,y plane. Consider the part of the string between two close points x_1 and x_2. The tension T in the string acts in the direction of the tangent to the curve formed by the string. The net force on the segment $[x_1, x_2]$ in the y direction is

$$T \sin\varphi_2 - T \sin\varphi_1$$

where φ_i is the angle between the tangent to the curve and the x axis at x_i. According to Newton's second law, this force must equal mass times acceleration. This is

$$\int_{x_1}^{x_2} \rho\, \partial^2 y/\partial t^2\, dx$$

where ρ is the density (mass per unit length) of the string. Thus, in the limit

$$T \frac{\partial}{\partial x} \sin\varphi = \rho \frac{\partial^2 y}{\partial t^2}$$

We note that $\tan\varphi = \partial y/\partial x$. If we make the simplifying assumption (justified or otherwise) that

$$\cos\varphi \approx 1, \qquad \frac{\partial}{\partial x} \cos\varphi \approx 0$$

we finally obtain:

$$T\partial^2 y/\partial x^2 = \rho\, \partial^2 y/\partial t^2$$

which is the well-known equation of the vibrating string.

The derivation of partial differential equations from physical laws usually brings about simplifying assumptions that are difficult to justify completely. Most of the time they are merely plausibility arguments. For this reason, some branches of science have accepted partial differential equations as axioms. The success of these axioms is judged by how well their conclusions describe past observations and predict new ones.

III. SOME WELL-KNOWN EQUATIONS

Now we list several equations that arise in various branches of science. Interestingly, the same equation can arise in diverse and unrelated areas.

A. Laplace's Equation

In n dimensions this equation is given by:

$$\Delta u = 0$$

where

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}$$

It arises in the study of electromagnetic phenomena (e.g., electrostatics, dielectrics, steady currents, magnetostatics), hydrodynamics (e.g., irrotational flow of a perfect fluid, surface waves), heat flow, gravitation, and many other branches of science. Solutions of Laplace's equation are called *harmonic functions.*

B. Poisson's Equation

$$\Delta u \equiv f(x), \qquad x = (x_1, \ldots, x_n)$$

Here, the function $f(x)$ is given. This equation is found in many of the situations in which Laplace's equation appears, since the latter is a special case.

C. Helmholtz's Equation

$$\Delta u \pm \alpha^2 u = 0$$

This equation appears in the study of elastic waves, vibrating strings, bars and membranes, sound and acoustics, electromagnetic waves, and the operation of nuclear reactors.

D. The Heat (Diffusion) Equation

This equation is of the form:

$$u_t = a^2 \Delta u$$

where $u(x_1, \ldots, x_n, t)$ depends on the variable t(time) as well. It describes heat conduction or diffusion processes.

E. The Wave Equation

$$\Box u \equiv (1/c^2) - \Delta u = 0$$

This describes the propagation of a wave with velocity c. This equation governs most cases of wave propagation.

F. The Telegraph Equation

$$\Box u + \sigma u_t = 0$$

This applies to some types of wave propagation.

G. The Scalar Potential Equation

$$\Box u = f(x, t)$$

H. The Klein–Gordon Equation

$$\Box u + \mu^2 u = 0$$

I. Maxwell's Equations

$$\nabla \times \mathbf{H} = \sigma \mathbf{E} + \varepsilon\, \partial \mathbf{E}/\partial t$$
$$\nabla \times \mathbf{E} = -\mu\, \partial \mathbf{H}/\partial t$$

Here $\mathbf{E}$ and $\mathbf{H}$ are three-dimensional vector functions of position and time representing the electric and magnetic fields, respectively. This system of equations is used in electrodynamics.

J. The Cauchy–Riemann Equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

These equations describe the real and imaginary parts of an analytic function of a complex variable.

K. The Schrödinger Equation

$$-\frac{\hbar^2}{2m}\Delta\psi + V(x)\psi = i\hbar\frac{\partial \psi}{\partial t}$$

This equation describes the motion of a quantum mechanical particle as it moves through a potential field. The function $V(x)$ represents the potential energy, while the unknown function $\psi(x)$ is allowed to have complex values.

L. Minimal Surfaces

In three dimensions a surface $z = u(x, y)$ having the least area for a given contour satisfies the equation:

$$\left(1 + u_y^2\right)u_{xx} - 2u_x u_y u_{xy} + \left(1 + u_x^2\right)u_{yy} = 0$$

where $u_x = \partial u/\partial x$, etc.

M. The Navier–Stokes Equations

$$\frac{\partial u_j}{\partial t} + \sum_k \frac{\partial u_j}{\partial x_k}u_k + \frac{1}{\rho}\frac{\partial p}{\partial x_j} = \gamma\,\Delta u_j$$
$$\sum_k \frac{\partial u_k}{\partial x_k} = 0$$

This system describes viscous flow of an incompressible liquid with velocity components u_k and pressure p.

N. The Korteweg–Devries Equation

$$u_t + cuu_x + u_{xxx} = 0$$

This equation is used in the study of water waves.

IV. TYPES OF EQUATIONS

In describing partial differential equations, the following notations are helpful. Let $\mathbb{R}^n$ denote Euclidean n dimensional space, and let $\mathbf{x} = (x_1, \ldots, x_n)$ denote a point in $\mathbb{R}^n$. One can consider partial differential equations on various types of manifolds, but we shall restrict ourselves to $\mathbb{R}^n$. For a real- or complex-valued function $u(x)$ we shall use the following notation:

$$D_k u = \partial u / i\partial x_k, \qquad 1 \leqq k \leqq n$$

If $\mu = (\mu_1, \ldots, \mu_n)$ is a multi-index of nonnegative integers, we write:

$$x^\mu = x_1^{\mu_1} \ldots x_n^{\mu_n}, \qquad |\mu| = \mu_1 + \cdots + \mu_n$$
$$D^\mu = D_1^{\mu_1} \cdots D_n^{\mu_n}$$

Thus, D^μ is a partial derivative of order $|\mu|$.

A. Linear Equations

The most general linear partial differential equation of order m is

$$Au \equiv \sum_{|\mu| \leqq m} a_\mu(x) D^\mu u = f(x) \tag{1}$$

It is called linear because the operator A is linear, that is, satisfies:

$$A(\alpha u + \beta v) = \alpha Au + \beta Av$$

for all function u, v and all constant scalars α, β. If the equation cannot be put in this form, it is nonlinear. In Section III, the examples in Sections A–K are linear.

B. Nonlinear Equations

In general, nonlinear partial differential equations are more difficult to solve than linear equations. There may be no solutions possible, as is the case for the equation:

$$|\partial u/\partial x| + 1 = 0$$

There is no general method of attack, and only special types of equations have been solved.

1. Quasilinear Equation

A partial differential equation is called *quasilinear* if it is linear with respect to its derivatives of highest order. This means that if one replaces the unknown function $u(x)$ and all its derivatives of order lower than the highest by known functions, the equation becomes linear. Thus, a quasilinear equation of order m is of the form:

$$\sum_{|\mu|=m} a_\mu(x, u, Du, \ldots, D^{m-1}u)D^\mu u = f(x, u, Du, \ldots, D^{m-1}u) \quad (2)$$

where the coefficients a_μ depend only on x, u, and derivatives of u up to order $m-1$. Quasilinear equations are important in applications. In Section III, the examples in Sections L–N are quasilinear.

2. Semilinear Equation

A quasilinear equation is called *semilinear* if the coefficients of the highest-order derivatives depend only on x. Thus, a semilinear equation of order m is of the form:

$$Au \equiv \sum_{|\mu|=m} a_\mu(x)D^\mu u = f(x, u, Du, \ldots, D^{m-1}u) \quad (3)$$

where A is linear. Semilinear equations arise frequently in practice.

C. Elliptic Equations

The quasilinear equation (2) is called *elliptic* in a region $\Omega \subset \mathbb{R}^n$ if for every function $v(x)$ the only real vector $\xi = (\xi_1, \ldots, \xi_n) \varepsilon \mathbb{R}^n$ that satisfies:

$$\sum_{|\mu|=m} a_\mu(x, v, Dv, \ldots, D^{m-1}v)\xi^\mu = 0$$

is $\xi = 0$. It is called *uniformly elliptic* in Ω if there is a constant $c_0 > 0$ independent of x and v such that:

$$c_0|\xi|^m \leqq \left| \sum_{|\mu|=m} a_\mu(x, v, Dv, \ldots, D^{m-1}v)\xi^\mu \right|$$

where $|\xi|^2 = \xi_1^2 + \cdots \xi_n^2$. The equations in Sections A–C and J–L are elliptic equations or systems.

D. Parabolic Equations

When $m = 2$, the quasilinear equation (2) becomes:

$$\sum_{j,k} a_{jk}(x, u, Du)\frac{\partial^2 u}{\partial x_j \partial x_k} = f(x, u, Du) \quad (4)$$

We shall say that Eq. (4) is *parabolic* in a region $\Omega \subset \mathbb{R}^n$ if for every choice of the function $v(x)$, the matrix $\Lambda = (a_{jk}(x, v, Dv))$ has a vanishing determinant for each $x \in \Omega$. The equation in Section III.D is a parabolic equation.

E. HYPERBOLIC EQUATIONS

Equation (4) will be called *ultrahyperbolic* in Ω if for each $v(x)$ the matrix $\Lambda = (a_{jk}(x, v, Dv))$ has some positive, some negative, and no vanishing eigenvalues for each $x \in \Omega$. It will be called *hyperbolic* if all but one of the eigenvalues of Λ have the same sign and none of them vanish in Ω. The equations in Sections III.E–H are hyperbolic equations.

The only time Eq. (4) is neither parabolic nor ultrahyperbolic is when all the eigenvalues of Λ have the same sign with none vanishing. In this case, Eq. (4) is elliptic as described earlier.

F. Equations of Mixed Type

If the coefficients of Eq. (1) are variable, it is possible that the equation will be of one type in one region and of another type in a different region. A simple example is

$$\partial^2 u/\partial x^2 - x\partial^2 u/\partial y^2 = 0$$

in two dimensions. In the region $x > 0$ it is hyperbolic, while in the region $x < 0$ it is elliptic. (It becomes parabolic on the line $x = 0$.)

G. Other Types

The type of an equation is very important in determining what problems can be solved for it and what kind of solutions it will have. As we saw, some equations can be of different types in different regions. One can define all three types for higher order equations, but most higher order equations will not fall into any of the three categories.

V. PROBLEMS ASSOCIATED WITH PARTIAL DIFFERENTIAL EQUATIONS

In practice one is rarely able to determine the most general solution of a partial differential equation. Usually one looks for a solution satisfying additional conditions. One may wish to prescribe the unknown function and/or some of its derivatives on part or all of the boundary of the region

in question. We call this a *boundary-value problem.* If the equation involves the variable t (time) and the additional conditions are prescribed at some time $t = t_0$, we usually refer to it as an *initial-value problem.*

A problem associated with a partial differential equation is called *well posed* if:

1. A solution exists for all possible values of the given data.
2. The solution is unique.
3. The solution depends in a continuous way on the given data.

The reason for the last requirement is that in most cases the given data come from various measurements. It is important that a small error in measurement of the given data should not produce a large error in the solution. The method of measuring the size of an error in the given data and in the solution is a basic question for each problem. There is no standard method; it varies from problem to problem.

The kinds of problems that are well posed for an equation depend on the type of the equation. Problems that are suitable for elliptic equations are not suitable for hyperbolic or parabolic equations. The same holds true for each of the types. We illustrate this with several examples.

A. Dirichlet's Problem

For a region $\Omega \subset \mathbb{R}^n$, Dirichlet's problem for Eq. (2) is to prescribe u and all its derivatives up to order $\frac{1}{2}m - 1$ on the boundary $\partial\Omega$ of Ω. This problem is well posed only for elliptic equations. If $m = 2$, only the function u is prescribed on the boundary. If Ω is unbounded, one may have to add a condition at infinity.

It is possible that the Dirichlet problem is not well posed for a linear elliptic equation of the form (1) because 0 is an eigenvalue of the operator A. This means that there is a function $w(x) \not\equiv 0$ that vanishes together with all derivatives up to order $\frac{1}{2}m - 1$ on $\partial\Omega$ and satisfies $Aw = 0$ in Ω. Thus, any solution of the Dirichlet problem for Eq. (1) is not unique, for we can always add a multiple of w to it to obtain another solution. Moreover, it is easily checked that one can solve the Dirichlet problem for Eq. (1) only if:

$$\int_\Omega f(x)w(x)\,dx$$

is a constant depending on w and the given data. Thus, we cannot solve the Dirichlet problem for all values of the given data. When Ω is bounded, one can usually remedy the situation by considering the equation,

$$Au + \varepsilon u = f \tag{5}$$

in place of Eq. (1) for ε sufficiently small.

B. The Neumann Problem

As in the case of the Dirichlet problem, the Neumann problem is well posed only for elliptic operators. The Neumann problem for Eq. (2) in a region Ω is to prescribe on $\partial\Omega$ the normal derivatives of u from order $\frac{1}{2}m$ to $m - 1$. (In both the Dirichlet and Neumann problems, exactly $\frac{1}{2}m$ normal derivatives are prescribed on $\partial\Omega$. In the Dirichlet problem, the first $\frac{1}{2}m$ are prescribed [starting from the zeroth-order derivative—u itself]. In the Neumann problem the next $\frac{1}{2}m$ normal derivatives are prescribed.)

As in the case of the Dirichlet problem, the Neumann problem for a linear equation of the form (1) can fail to be well posed because 0 is an eigenvalue of A. Again, this can usually be corrected by considering Eq. (5) in place of Eq. (1) for ε sufficiently small.

When $m = 2$, the Neumann problem consists of prescribing the normal derivative $\partial u/\partial n$ of u on $\partial\Omega$. In the case of Laplace's or Poisson's equation, it is easily seen that 0 is indeed an eigenvalue if Ω is bounded, for then any constant function is a solution of:

$$\Delta w = 0 \quad \text{in} \quad \Omega, \qquad \partial w/\partial n = 0 \quad \text{on} \quad \partial\Omega \tag{6}$$

Thus, adding any constant to a solution gives another solution. Moreover, we can solve the Neumann problem:

$$\Delta u = f \quad \text{in} \quad \Omega, \qquad \frac{\partial u}{\partial n} = g \quad \text{on} \quad \partial\Omega \tag{7}$$

only if

$$\int_\Omega f(x)\,dx = \int_{\partial\Omega} g\,ds$$

Thus, we cannot solve Eq. (7) for all f and g. If Ω is unbounded, one usually requires that the solution of the Neumann problem vanish at infinity. This removes 0 as an eigenvalue, and the problem is well posed.

C. The Robin Problem

When $m = 2$, the Robin problem for Eq. (2) consists of prescribing:

$$Bu = \alpha\,\partial u/\partial n + \beta u \tag{8}$$

on $\partial\Omega$, where $\alpha(x)$, $\beta(x)$ are functions and $\alpha(x) \neq 0$ on $\partial\Omega$. If $\beta(x) \equiv 0$, this reduces to the Neumann problem. Again this problem is well posed only for elliptic equations.

D. Mixed Problems

Let B be defined by Eq. (8), and assume that $\alpha(x)^2 + \beta(x)^2 \neq 0$ on $\partial\Omega$. Consider the boundary value problem consisting of finding a solution of Eq. (4) in Ω and prescribing Bu on $\partial\Omega$. On those parts of $\partial\Omega$ where $\alpha(x) = 0$,

we are prescribing Dirichlet data. On those parts of $\partial\Omega$ where $\beta(x) = 0$, we are prescribing Neumann data. On the remaining sections of $\partial\Omega$ we are prescribing Robin data. This is an example of a mixed boundary-value problem in which one prescribes different types of data on different parts of the boundary. Other examples are provided by parabolic and hyperbolic equations, to be discussed later.

E. General Boundary-Value Problems

For an elliptic equation of the form (2) one can consider general boundary conditions of the form:

$$B_j u \equiv \sum_{|\mu| \leqq m_j} b_{j\mu} D^\mu u = g_j \quad \text{on} \quad \partial\Omega, \qquad 1 \leqq j \leqq m/2 \tag{9}$$

Such boundary-value problems can be well posed provided the operators B_j are independent in a suitable sense and do not "contradict" each other or Eq. (2).

F. The Cauchy Problem

For Eq. (2), the Cauchy problem consists of prescribing all derivatives of u up to order $m-1$ on a smooth surface S and solving Eq. (2) for u in a neighborhood of S. An important requirement for the Cauchy problem to have a solution is that the boundary conditions not "contradict" the equation on S. This means that the coefficient of $\partial^m u/\partial n^m$ in Eq. (2) should not vanish on S. Otherwise, Eq. (2) and the Cauchy boundary conditions,

$$\partial^k u/\partial n^k = g_k \quad \text{on} \quad S, k = 0, \ldots, n-1 \tag{10}$$

involve only the function g_k on S. This is sure to cause a contradiction unless f is severely restricted. When this happens we say that the surface S is *characteristic* for Eq. (2). Thus, for the Cauchy problem to have a solution without restricting f, it is necessary that the surface S be noncharacteristic. In the quasilinear case, the coefficient of $\partial^m u/\partial n^m$ in Eq. (2) depends on the g_k. Thus, the Cauchy data (10) will play a role in determining whether or not the surface S is characteristic for Eq. (2). The Cauchy–Kowalewski theorem states that for a noncharacteristic analytic surface S, real analytic Cauchy data g_k, and real analytic coefficients a_μ, f in Eq. (2), the Cauchy problem (2) and (10) has a unique real analytic solution in the neighborhood of S. This is true irrespective of the type of the equation. However, this does not mean that the Cauchy problem is well posed for all types of equations. In fact, the hypotheses of the Cauchy–Kowalewski theorem are satisfied for the Cauchy problem,

$$u_{xx} + u_{yy} = 0, \qquad y > 0$$
$$u(x, 0) = 0, \qquad u_y(x, 0) = n^{-1} \sin nx$$

and, indeed, it has a unique analytic solution:

$$u(x, y) = n^{-2} \sinh ny \sin nx$$

The function $n^{-1} \sin nx$ tends uniformly to 0 as $n \to \infty$, but the solution does not become small as $n \to \infty$ for $y \neq 0$. It can be shown that the Cauchy problem is well posed only for hyperbolic equations.

VI. METHODS OF SOLUTION

There is no general approach for finding solutions of partial differential equations. Indeed, there exist linear partial differential equations with smooth coefficients having no solutions in the neighborhood of a point. For instance, the equation:

$$\frac{\partial u}{\partial x_1} + i\frac{\partial u}{\partial x_2} + 2i(x_1 + ix_2)\frac{\partial u}{\partial x_3} = f(x_3) \tag{11}$$

has no solution in the neighborhood of the origin unless $f(x_3)$ is a real analytic function of x_3. Thus, if f is infinitely differentiable but not analytic, Eq. (11) has no solution in the neighborhood of the origin. Even when partial differential equations have solutions, we cannot find the "general" solution. We must content ourselves with solving a particular problem for the equation in question. Even then we are rarely able to write down the solution in closed form. We are lucky if we can derive a formula that will enable us to calculate the solution in some way, such as a convergent series or iteration scheme. In many cases, even this is unattainable. Then, one must be satisfied with an abstract theorem stating that the problem is well posed. Sometimes the existence theorem does provide a method of calculating the solution; more often it does not.

Now we describe some of the methods that can be used to obtain solutions in specific situations.

A. Separation of Variables

Consider a vibrating string stretched along the x axis from $x = 0$ and $x = \pi$ and fixed at its end points. We can assign the initial displacement and velocity. Thus, we are interested in solving the mixed initial and boundary value problem for the displacement $u(x, t)$:

$$\Box u = 0, \qquad 0 < x < \pi, \quad t > 0 \tag{12}$$
$$u(x, 0) = f(x),$$
$$u_t(x, 0) = g(x), \qquad 0 \leqq x \leqq \pi \tag{13}$$
$$u(0, t) = u(\pi, t) = 0, \qquad t \geqq 0 \tag{14}$$

We begin by looking for a solution of Eq. (12) of the form:

$$u(x, t) = X(x)T(t)$$

Such a function will be a solution of Eq. (12) only if:

$$\frac{T''}{T} = c^2 \frac{X''}{X} \tag{15}$$

Since the left-hand side of Eq. (15′) is a function of t only and the right-hand side is a function of x only, both sides are constant. Thus, there is a constant K such that $X''(x) = KX(x)$. If $K = \lambda^2 > 0$, X must be of the form:

$$X = Ae^{-\lambda x} + Be^{\lambda x}$$

For Eq. (14) to be satisfied, we must have

$$X(0) = X(\pi) = 0 \tag{15′}$$

This can happen only if $A = B = 0$. If $K = 0$, then X must be of the form:

$$x = A + Bx$$

Again, this can happen only if $A = B = 0$. If $K = -\lambda^2 < 0$, then X is of the form:

$$X = A \cos \lambda x + B \sin \lambda x$$

This can satisfy Eq. (15) only if $A = 0$ and $B \sin \lambda\pi = 0$. Thus, the only way that X should not vanish identically is if λ is an integer n. Moreover, T satisfies:

$$T'' + n^2 c^2 T = 0$$

The general solution for this is

$$T = A \cos nct + B \sin nct$$

Thus,

$$u(x, t) = \sin nx(A_n \cos nct + B_n \sin nct)$$

is a solution of Eqs. (12) and (14) for each integer n. However, it will not satisfy Eq. (13) unless $f(x)$, $g(x)$ are of a special form. The linearity of the operator $\Box$ allows one to add solutions of Eqs. (12) and (14). Thus,

$$u(x, t) = \sum_{n=1}^{\infty} \sin nx(A_n \cos nct + B_n \sin nct) \tag{16}$$

will be a solution provided the series converges. Moreover, it will satisfy Eq. (13) if:

$$f(x) = \sum_{n=1}^{\infty} A_n \sin nx,$$

$$g(x) = \sum_{n=1}^{\infty} ncB_n \sin nx$$

This will be true if $f(x)$, $g(x)$ are expandable in a Fourier sine series. If they are, then the coefficients A_n, B_n are given by:

$$A_n = \frac{2}{\pi} \int_0^{\pi} f(x) \sin nx \, dx,$$

$$B_n = \frac{2}{nc\pi} \int_0^{\pi} g(x) \sin nx \, dx$$

With these values, the series (16) converges and gives a solution of Eqs. (12)–(14).

B. Fourier Transforms

If we desire to determine the temperature $u(x, t)$ of a system in $\mathbb{R}^n$ with no heat added or removed and initial temperature given, we must solve:

$$u_t = a^2 \Delta u, \qquad x \varepsilon \mathbb{R}^n, \quad t > 0 \tag{17}$$

$$u(x, 0) = \varphi(x), \qquad x \varepsilon \mathbb{R}^n \tag{18}$$

If we apply the Fourier transform,

$$\hat{f}(\xi) = (2\pi)^{-n/2} \int e^{-i\xi x} f(x) \, dx \tag{19}$$

where $\xi x = \xi_1 x_1 + \cdots + \xi_n x_n$, we obtain:

$$\hat{u}_t(\xi, t) + a^2 |\xi|^2 \hat{u}(\xi, t) = 0$$

The solution satisfying Eq. (18) is

$$\hat{u}(\xi, t) = e^{-a^2|\xi|^2 t} \hat{\varphi}(\xi)$$

If we now make use of the inverse Fourier transform,

$$f(x) = (2\pi)^{-n/2} \int e^{i\xi x} \hat{f}(\xi) \, d\xi$$

we have

$$u(x, t) = \int K(x - y, t) \varphi(y) \, dy$$

where

$$K(x, t) = (2\pi)^{-n} \int e^{ix\xi - a^2|\xi|^2 t} \, d\xi$$

If we introduce the new variable,

$$\eta = at^{1/2}\xi - \tfrac{1}{2} i a^{-1} t^{-1/2} x$$

this becomes

$$(2\pi)^{-n} a^{-n} t^{-n/2} e^{-|x|^2/4a^2 t} \int e^{-|\eta|^2} \, d\eta$$

$$= (4\pi a^2 t)^{-n/2} e^{-|x|^2/4a^2 t}$$

This suggests that a solution of Eqs. (17) and (18) is given by:

$$u(x, t) = (4\pi a^2 t)^{-n/2} \int e^{-|x-y|^2/4a^2 t} \varphi(y) \, dy \tag{20}$$

It is easily checked that this is indeed the case if φ is continuous and bounded. However, the solution is not unique unless one places more restriction on the solution.

C. Fundamental Solutions, Green's Function

Let

$$K(x, y) = \frac{|x - y|^{2-n}}{(2-n)\omega_n} + h(x), \qquad n > 2,$$

$$K(x, y) = \frac{\log 4}{2\pi} + h(x), \qquad n = 2$$

where $\omega_n = 2\pi^{n/2}/\Gamma(\frac{1}{2}n)$ is the surface area of the unit sphere in $\mathbb{R}^n$ and $h(x)$ is a harmonic function in a bounded domain $\Omega \subset \mathbb{R}^n$ (i.e., $h(x)$ is a solution of $\Delta h = 0$ in Ω). If the boundary $\partial\Omega$ of Ω is sufficiently regular and $h \in C^2(\bar{\Omega})$, then Green's theorem implies for $y \in \Omega$:

$$u(y) = \int_\Omega K(x, y)\Delta u(x)\, dx + \int_{\partial\Omega} \left(u(x)\frac{\partial K(x, y)}{\partial n} - K(x, y)\frac{\partial u}{\partial n} \right) dS_x \quad (21)$$

for all $u \in C^2(\bar{\Omega})$. The function $K(x, y)$ is called a *fundamental solution* of the operator Δ. If, in addition, $K(x, y)$ vanishes for $x \in \partial\Omega$, it is called a *Green's function*, and we denote it by $G(x, y)$. In this case,

$$u(y) = \int_{\partial\Omega} u(x)\frac{\partial G(x, y)}{\partial n}\, dS_x, \qquad y \in \Omega \quad (22)$$

for all $u \in C^2(\bar{\Omega})$ that are harmonic in Ω. Conversely, this formula can be used to solve the Dirichlet problem for Laplace's equation if we know the Green's function for Ω, since the righthand side of Eq. (22) is harmonic in Ω and involves only the values of $u(x)$ on $\partial\Omega$. It can be shown that if the prescribed boundary values are continuous, then indeed Eq. (22) does give a solution to the Dirichlet problem for Laplace's equation.

It is usually very difficult to find the Green's function for an arbitrary domain. It can be computed for geometrically symmetric regions. In the case of a ball of radius R and center 0, it is given by:

$$G(x, y) = K(x, y) - (|y|/R)^{2-n} K(x, R^2|y|^{-2}y)$$

D. Hilbert Space Methods

Let

$$P(D) = \sum_{|\mu|=m} a_\mu D^\mu$$

be a positive, real, constant coefficient, homogeneous, elliptic partial differential operator of order $m = 2r$. This means that $P(D)$ has only terms of order m, and

$$c_0|\xi|^m \leqq P(\xi) \leqq C_0|\xi|^m, \qquad \xi \in \mathbb{R}^n \quad (23)$$

holds for positive constants c_0, C_0. We introduce the norm,

$$|v|_r = \left(\int |\xi|^m |\hat{v}(\xi)|^2\, d\xi \right)^{1/2}$$

for function v in $C_0^\infty(\Omega)$, the set of infinitely differentiable functions that vanish outside Ω. Here, Ω is a bounded domain in $\mathbb{R}^n$ with smooth boundary, and $\hat{v}(\xi)$ denotes the Fourier transform given by Eq. (19). By Eq. (23) we see that $(P(D)v, v)$ is equivalent to $|v|_r^2$ on $C_0^\infty(\Omega)$, where

$$(u, v) = \int_\Omega u(x)\overline{v(x)}\, dx$$

Let

$$a(u, v) = (u, P(D)v), \quad u, v \in C_0^\infty(\Omega) \quad (24)$$

If $u \in C^m(\bar{\Omega})$ is a solution of the Dirichlet problem,

$$P(D)u = f \quad \text{in} \quad \Omega \quad (25)$$

$$D^\mu u = 0 \quad \text{on} \quad \partial\Omega, \qquad |\mu| < r \quad (26)$$

then it satisfies

$$a(u, v) = (f, v), \qquad v \in C_0^\infty(\Omega) \quad (27)$$

Conversely, if $\partial\Omega$ is sufficiently smooth and $u \in C^m(\bar{\Omega})$ satisfies Eq. (27), then it is a solution of the Dirichlet problem, Eqs. (25) and (26). This is readily shown by integration by parts. Thus, one can solve Eqs. (25) and (26) by finding a function $u \in C^m(\bar{\Omega})$ satisfying Eq. (27). Since $a(u, v)$ is a scalar product, it would be helpful if we had a theorem stating that the expression (f, v) can be represented by the expression $a(u, v)$ for some u. Such a theorem exists (the Riesz representation theorem), provided $a(u, v)$ is the scalar product of a Hilbert space and

$$|(f, v)| \leqq Ca(v, v)^{1/2} \quad (28)$$

We can fit our situation to the theorem by completing $C_0^\infty(\Omega)$ with respect to the $|v|_r$ norm and making use of the fact that $a(v, v)^{1/2}$ and $|v|_r$ are equivalent on $C_0^\infty(\Omega)$ and consequently on the completion $H_0^r(\Omega)$. Moreover, inequality (28) follows from the Poincare inequality,

$$\|v\| \leqq M^r|v|_r, \qquad v \in C_0^\infty(\Omega) \quad (29)$$

which holds if Ω is contained in a cube of side length M. Thus, by Schwarz's inequality,

$$|(f, v)| \leqq \|f\|\,\|v\| \leqq \|f\|M^r|v|_r \leqq ca(v, v)^{1/2}$$

The Riesz representation theorem now tells us that there is a $u \in H_0^r(\Omega)$ such that Eq. (27) holds. If we can show that u is in $C^m(\bar{\Omega})$, it will follow that u is indeed a solution of the Dirichlet problem, Eqs. (25) and (26). As it stands now, u is only a *weak solution* of Eqs. (25) and (26). However,

it can be shown that if $\partial\Omega$ and f are sufficiently smooth, then u will be in $C^m(\bar{\Omega})$ and will be a solution of Eqs. (25) and (26).

The proof of the Poincare inequality (29) can be given as follows. It sufficies to prove it for $r = 1$ and Ω contained in the slab $0 < x_1 < M$. Since $v \in C_0^\infty(\Omega)$,

$$\begin{aligned} v(x_1, \ldots, x_n)^2 &= \left(\int_0^{x_1} v_{x1}(t, x_2, \ldots, x_n)\, dt \right)^2 \\ &\leqq x_1 \int_0^{x_1} v_{x1}(t, x_2, \ldots, x_n)^2\, dt \\ &\leqq M \int_0^{M} v_{x1}(t, x_2, \ldots, x_n)^2\, dt \end{aligned}$$

Thus,

$$\int_0^M v(x_1, \ldots, x_n)^2\, dx_1 \leqq M^2 \int_0^M v_{x1}(t, x_2, \ldots, x_n)^2\, dt$$

If we now integrate over $x_2, \ldots, x_n$, we obtain:

$$\|v\| \leqq M \|v_{x1}\|$$

But, by Parseval's identity,

$$\begin{aligned} \int |v_{x1}|^2\, dx &= \int |\hat{v}_{x1}|^2\, d\xi \\ &= \int \xi_1^2 |\hat{v}|^2\, d\xi \leqq \int |\xi|^2 |\hat{v}|^2\, d\xi = |v|_1^2 \end{aligned}$$

E. Iterations

An important method of solving both linear and nonlinear problems is that of *successive approximations*. We illustrate this method for the following Dirichlet problem:

$$\Delta u = f(x, u), \; x \in \Omega \tag{30}$$

$$u = 0 \quad \text{on} \quad \partial\Omega \tag{31}$$

We assume that the boundary of Ω is smooth and that $f(x, t) = f(x_1 \ldots, x_n, t)$ is differentiable with respect to all arguments. Also we assume that:

$$|f(x, t)| \leqq N, \qquad x \in \Omega, \quad -\infty < t < \infty \tag{32}$$

$$|\partial f(x, t)/\partial t| \leqq \psi(t), \quad x \in \Omega \tag{33}$$

where $\psi(t)$ is a continuous function.

First, we note that for every compact subset G of Ω there is a constant C such that:

$$\max_{\mathrm{G}} |\nabla v| \leqq C(\sup_{\Omega} |\Delta v| + \sup_{\Omega} |v|) \tag{34}$$

for all $v \in C^2(\Omega)$. Assume this for the moment, and let $w(x)$ be the solution of the Dirichlet problem,

$$\Delta w = -N \quad \text{in} \quad \Omega, \qquad w = 0 \quad \text{on} \quad \partial\Omega \tag{35}$$

It is clear that $w(x) \geqq 0$ in Ω. For otherwise it would have a negative interior mininum in Ω. At such a point, one has $\partial^2 w/\partial x_k^2 \geqq = 0$ for each k, and consequently. $\Delta w \geqq 0$ contradicting Eq. (35). Since $w \in C(\bar{\Omega})$, there is a constant C_t such that:

$$0 \leqq w(x) \leqq C_1, \qquad x \in \Omega$$

Let

$$K = \max_{|t| \leqq C_1} \psi(t)$$

Then, by Eq. (33):

$$|\partial f(x, t)/\partial t| \leqq K, \qquad |t| \leqq C_1 \tag{36}$$

Consequently,

$$f(x, t) - f(x, s) \leqq K(t - s) \tag{37}$$

when $-C_1 \leqq s \leqq t \leqq C_1$. We define a sequence $\{u_k\}$ of functions as follows. We take $u_0 = w$ and once u_{k-1} has been defined, we let u_k be the solution of the Dirichlet problem:

$$Lu_k \equiv \Delta u_k - K u_k = f(x, u_{k-1}) - K u_{k-1} \tag{38}$$

in Ω with $u_k = 0$ on $\partial\Omega$. The solution exists by the theory of linear elliptic equations. We show by induction that:

$$-w \leqq u_k \leqq u_{k-1} \leqq w \tag{39}$$

To see this for $k = 1$, note that:

$$L(u_1 - w) = f(x, w) - K w - \Delta w + K w \geqq 0$$

From this we see that $u_1 \leqq w$ in Ω. If $u_1 - w$ had an interior positive maximum in Ω, we would have

$$L(u_1 - w) = \Delta(u_1 - w) - K(u_1 - w) < 0$$

at such a point. Thus, $u_1 \leqq w$ in Ω. Also we note:

$$\Delta(u_1 + w) = f(x, w) + K(u_1 - w) + \Delta w \leqq K(u_1 - w)$$

This shows that $u_1 + w$ cannot have a negative minimum inside Ω. Hence, $u_1 + w \geqq 0$ in Ω, and Eq. (39) is verified for $k = 1$. Once we know it is verified for k, we note that:

$$\begin{aligned} L[u_{k+1} - u_k] &= f(x, u_k) - f(x, u_{k-1}) \\ &\quad - K(u_k - u_{k-1}) \geqq 0 \end{aligned}$$

by Eq. (37). Thus, $u_{k-1} \leqq u_k$ in Ω. Hence,

$$\begin{aligned} \Delta(u_{k+1} + w) &= f(x, u_k) - K u_k + K u_{k+1} \\ &\quad + \Delta w \leqq K(u_{k+1} - u_k) \end{aligned}$$

Again, we deduce from this that $u_{k+1} + w \geqq 0$ in Ω. Hence, Eq. (39) holds for $k + 1$ and consequently for all k. In particular, we see that the u_k are uniformly bounded in Ω, and by Eq. (38) the same is true of the functions Δu_k.

Hence, by Eq. (34), the first derivatives of the u_k are uniformly bounded on compact subsets of Ω. If we differentiate Eq. (38), we see that the sequence $\Delta(\partial u_k/\partial x_j)$ is uniformly bounded on compact subsets of Ω (here we make use of the continuous differentiability of f). If we now make use of Eq. (34) again, we see that the second derivatives of the u_k are uniformly bounded on compact subsets of Ω. Hence, by the Ascoli–Arzela theorem, there is a subsequence that converges together with its first derivatives uniformly on compact subsets of Ω. Since the sequence u_k is monotone, the whole sequence must converge to a continuous function u that satisfies $|u(x)| \leqq w(x)$. Hence, u vanishes on $\partial\Omega$. By Eq. (38), the functions Δu_k must converge uniformly on compact subsets, and by Eq. (34), the same must be true of the first derivatives of the u_k. From the differentiated Eq. (38) we see that the $\Delta(\partial u_k/\partial x_j)$ converge uniformly on bounded subsets and consequently the same is true of the second derivatives of the u_k by Eq. (34). Since the u_k converge uniformly to u in Ω and their second derivatives converge uniformly on bounded subsets, we see that $u \in C^2(\Omega)$ and $\Delta u_k \to \Delta u$. Hence,

$$\begin{aligned}\Delta u &= \lim \Delta u_k \\ &= \lim[f(x, u_{k-1}) + K(u_k - u_{k-1})] = f(x, u)\end{aligned}$$

and u is the desired solution.

It remains to prove Eq. (34). For this purpose we let $\varphi(x)$ be a function in $C_0^\infty(\Omega)$ which equals one on G. Then we have, by Eq. (21),

$$\varphi(y)v(y) = \int_\Omega K(x, y)\Delta(\varphi(x)v(x))\,dx$$

(the boundary integrals vanish because φ is 0 near $\partial\Omega$). Thus, if $y \in G$,

$$\begin{aligned}v(y) &= \int_\Omega K\{\varphi\Delta v + 2\nabla\varphi\cdot\nabla v + v\Delta\varphi\}\,dx \\ &= \int_\Omega \{\varphi K\Delta v - 2v\nabla K\cdot\nabla\varphi - vK\Delta\varphi\}\,dx\end{aligned}$$

by integration by parts. We note that $\nabla\varphi$ vanishes near the singularity of K. Thus, we may differentiate under integral sign to obtain:

$$\frac{\partial v(y)}{\partial y_j} = \int_\Omega \left\{\varphi\frac{\partial K}{\partial y_j}\Delta v - 2v\frac{\partial\nabla k}{\partial y_j}\cdot\nabla\varphi - v\frac{\partial k}{\partial y_j}\Delta\varphi\right\}dx$$

Consequently,

$$\begin{aligned}\left|\frac{\partial v}{\partial y_j}\right| &\leqq \sup_\Omega|\Delta v|\int_\Omega\left|\frac{\partial K}{\partial y_j}\right|dx \\ &\quad + \sup_\Omega|v|\int_\Omega\left\{|\nabla\varphi|\left|\frac{\partial K}{\partial y_j}\right| + |\Delta\varphi|\left|\frac{\partial K}{\partial y_j}\right|K\right\}dx\end{aligned}$$

and all of the integrals are finite. This gives Eq. (34), and the proof is complete.

F. Variational Methods

In many situations, methods of the calculus of variations are useful in solving problems for partial differential equations, both linear and nonlinear. We illustrate this with a simple example. Suppose we wish to solve the problem,

$$-\sum\frac{\partial}{\partial x_k}\left[p_k(x)\frac{\partial u(x)}{\partial x_k}\right] + q(x)u(x) = 0 \quad \text{in} \quad \Omega \tag{40}$$

$$u(x) = g(x) \quad \text{on} \quad \partial\Omega \tag{41}$$

Assume that $p_k(x) \geqq c_0$, $q(x) \geqq c_0$, $c_0 > 0$ for $x \in \Omega$, Ω bounded, $\partial\Omega$ smooth, and that g is in $C^1(\partial\Omega)$. We consider the expression,

$$a(u, v) = \int_\Omega\left\{\frac{1}{2}\sum p_k(x)\frac{\partial u(x)}{\partial x_k}\frac{\partial v(x)}{\partial x_k} + q(x)u(x)v(x)\right\}dx$$

and put $a(u) = a(u, u)$. If $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$, satisfies, Eq. (41), and

$$a(u) \leqq a(w) \tag{42}$$

for all w satisfying Eq. (41), then it is readily seen that u is a solution of Eq. (40). Let v be a smooth function which vanishes on $\partial\Omega$. Then, for any scalar β,

$$a(u) \leqq a(u + \beta v) = a(u) + 2\beta a(u, v) + \beta^2 a(v)$$

and, consequently,

$$2\beta a(u, v) + \beta^2 a(v)^2 \geqq 0$$

for all β. This implies $a(u, v) = 0$. Integration by parts now yields Eq. (40). The problem now is to find a u satisfying Eq. (42). We do this as follows. Let H denote the Hilbert space obtained by completing $C^2(\bar{\Omega})$ with respect to the norm $a(w)$, and let H_0 be the subspace of those functions in H that vanish on $\partial\Omega$. Under the hypotheses given it can be shown that H_0 is a closed subspace of H. Let w be an element of H satisfying Eq. (41), and take a sequence $\{v_k\}$ of functions in H_0 such that:

$$a(w - v_k) \to d = \inf_{v\in H_0} a(w - v)$$

The parallelogram law for Hilbert space tells us that:

$$\begin{aligned}&a(2w - v_j - v_k) + a(v_j - v_k) \\ &\quad = 2a(w - v_j) + 2a(w - v_k)\end{aligned}$$

But,

$$a(2w - v_j - v_k) = 4a\left(w - \tfrac{1}{2}(v_j + u_k)\right) \geqq 4d$$

Hence,

$$4d + a(v_j - v_k) \leqq 2a(w - v_j) + 2a(w - v_k) \to 4d$$

Thus,

$$a(v_j - v_k) \to 0$$

and $\{v_k\}$ is a Cauchy sequence in H_0. Since H_0 is complete, there is a $v \in H_0$ such that $a(v_k - v) \to 0$. Put $u = w - v$. Then clearly u satisfies Eqs. (41) and (42), hence $a(u, v) = 0$ for all $v \in H_0$. If u is in $C^2(\Omega)$, then it will satisfy Eq. (40) as well. The rub is that functions in H need not be in $C^2(\Omega)$. However, the day is saved if the $p_k(x)$ and $q(x)$ are sufficiently smooth. For then one can show that functions $u \in H$ which satisfy $a(u, v) = 0$ for all $v \in H_0$ are indeed in $C^2(\Omega)$.

G. Critical Point Theory

Many partial differential equations and systems are the Euler–Lagrange equations corresponding to a real valued functional G defined on some Hilbert space E. Such a functional usually represents the energy or related quantity of the physical system governed by the equations. In such a case, solutions of the partial differential equations correspond to critical points of the functional. For instance, if the equation to be solved is

$$-\Delta u(x) = f(x, u), x \in \Omega \subset \mathbb{R}^n \tag{43}$$

then solutions of Eq. (43) are the critical points of the functional,

$$G(u) = \|\nabla u\|^2 - 2\int_\Omega F(x, u)\, dx$$

where the norm is that of $L^2(\Omega)$ and

$$F(x, t) = \int_0^t f(x, s)\, ds$$

The history of this approach can be traced back to the calculus of variations in which equations of the form:

$$G'(u) = 0 \tag{44}$$

are the Euler–Lagrange equations of the functional G. The original method was to find maxima or minima of G by solving Eq. (44) and then show that some of the solutions are extrema. This approach worked well for one-dimensional problems. In this case, it is easier to solve Eq. (44) than it is to find a maximum or minimum of G. However, in higher dimensions it was realized quite early that it is easier to find maxima and minima of G than it is to solve Eq. (44). Consequently, the tables were turned, and critical point theory was devoted to finding extrema of G. This approach is called the *direct method* in the calculus of variations. If an extremum point of G can be identified, it will automatically be a solution of Eq. (44).

The simplest extrema to find are global maxima and minima. For such points to exist one needs G to be semi-bounded. However, this in itself does not guarantee that an extremum exists. All that can be derived from the semi-boundedness is the existence of a sequence $\{u_k\} \subset E$ such that:

$$G(u_k) \to c, G'(u_k) \to 0 \tag{45}$$

where c is either the supremum or infimum of G. The existence of such a sequence does not guarantee that an extremum exists. However, it does so if one can show that the sequence has a convergent subsequence. This condition is known as the *Palais–Smale condition*. A sequence satisfying Eq. (45) is called a *Palais–Smale sequence.*

Recently, researchers have discovered several sets of sufficient conditions on G which will guarantee the existence of a Palais–Smale sequence. If, in addition, the Palais–Smale condition holds, then one obtains a critical point. These conditions can apply to functionals which are not semibounded. Points found by this method are known as *mountain pass points* due to the geometrical interpretation that can be given. The situation can be described as follows. Suppose Q is an open set in E and there are two points e_0, e_1 such that $e_0 \in Q$, $e_1 \notin \bar{Q}$, and

$$G(e_0), G(e_1) \le b_0 = \inf_{\partial Q} G \tag{46}$$

Then there exists a Palais–Smale sequence, Eq. (45), with c satisfying:

$$b_0 \le c < \infty \tag{47}$$

When G satisfies Eq. (46) we say that it exhibits mountain pass geometry.

It was then discovered that other geometries (i.e., configurations) produce Palais–Smale sequences, as well. Consider the following situation. Assume that

$$E = M \oplus N$$

is a decomposition of E into the direct sum of closed subspaces with

$$\dim N < \infty$$

Suppose there is an $R > 0$ such that

$$\sup_{N \cap \partial B_R} G \le b_0 = \inf_M G$$

Then, again, there is a sequence satisfying Eqs. (45) and (47). Here, B_R denotes the ball of radius R in E and ∂B_R denotes its boundary.

In applying this method to solve Eq. (43), one discovers that the asymptotic behavior of $f(x, t)$ at ∞ plays an important role. One can consider several possibilities. When

$$\lim_{|t| \to \infty} \sup |f(x, t)/t| = \infty \tag{48}$$

we say that problem (43) is *superlinear*. Otherwise, we call it *sublinear*. If

$$f(x,t)/t \to b_{\pm}(x) \ \ as \ \ t \to \pm\infty \tag{49}$$

and the $b_{\pm}$ are different, we say that it has a nonlinearity at ∞. An interesting special case of Eq. (49) is when the $b_{\pm}(x)$ are constants. If $b_{-}(x) \equiv a$ and $b_{+}(x) \equiv b$, we say that (a, b) is in the Fučík spectrum $\sum$ if:

$$-\Delta u = bu^{+} - au^{-}$$

has a nontrivial solution, where $u^{\pm} = \max\{\pm u, 0\}$. Because of its importance in solving Eq. (43), we describe this spectrum. It has been shown that emanating from each eigenvalue λ_ℓ of $-\Delta$, there are curves $\mu_\ell(a)$, $\nu_{\ell-1}(a)$ (which may coincide) which are strictly decreasing at least in the square $S = [\lambda_{\ell-1}, \lambda_{\ell+1}]^2$ and such that $(a, \mu_\ell(a))$ and $(a, \nu_{\ell-1}(a))$ are in $\sum$ in the square S. Moreover, the regions $b > \mu_\ell(a)$ and $b < \nu_{\ell-1}(a)$ are free of $\sum$ in S. These curves are known exactly only in the one-dimensional case (ordinary differential equations). In higher dimensions it is not known in general how many curves of $\sum$ emanate from each eigenvalue. It is known that there is at least one (when $\mu_\ell(a)$ and $\nu_{\ell-1}(a)$ coincide). If there are two or more curves emanating from an eigenvalue, the status of the region between them is unknown in general. If the eigenvalue is simple, there are at most two curves of $\sum$ emanating from it, and any region between them is not in $\sum$. On the other hand, examples are known in which many curves of $\sum$ emanate from a multiple eigenvalue. In the higher dimensional case, the curves have not been traced asymptotically.

If

$$f(x,t)/t \to \lambda_\ell \ \ as \ \ |t| \to \infty \tag{50}$$

where λ_ℓ is one of the eigenvalues of $-\Delta$, we say that Eq. (43) has asymptotic resonance. One can distinguish several types. One can have the situation:

$$f(x,t) = \lambda_\ell t + p(x,t) \tag{51}$$

where $p(x,t) = o(|t|^{\beta})$ as $|t| \to \infty$ for some $\beta < 1$. Another possibility is when $p(x,t)$ satisfies:

$$|p(x,t)| \le V(x) \in L^2(\Omega) \tag{52}$$

and

$$p(x,t) \to p_{\pm}(x) \ \ a.e. \ \ as \ \ t \to \pm\infty$$

A stronger form occurs when

$$p(x,t) \to 0 \ \ as \ \ |t| \to \infty \tag{53}$$

and

$$|P(x,t)| \le W(x) \in L^1(\Omega) \tag{54}$$

where

$$P(x,t) := \int_0^t p(x,s)\,ds$$

This type of problem is more difficult to solve; it is called *strong resonance*. Possible situations include:

$$P(x,t) \to P_{\pm}(x) \ \ as \ \ t \to \pm\infty \tag{55}$$

and

$$P(x,t) \to P_0(x) \ \ as \ \ |t| \to \infty \tag{56}$$

What is rather surprising is that the stronger the resonance, the more difficult it is to solve Eq. (43). There is an interesting connection between Eq. (43) and nonlinear eigenvalue problems of the form:

$$G'(u) = \beta u \tag{57}$$

for functionals. This translates into the problem

$$-\Delta u = \lambda f(x,u) \tag{58}$$

for partial differential equations. It can be shown that there is an intimate relationship between Eqs. (44) and (57) (other than the fact that the former is a special case of the latter). In fact, the absence of a certain number of solutions of Eq. (44) implies the existence of a rich family of solutions of Eq. (57) on all spheres of sufficiently large radius, and vice versa. The same holds true for Eqs. (43) and (58).

H. Periodic Solutions

Many problems in partial differential equations are more easily understood if one considers periodic functions. Let

$$Q = \left\{x \in \mathbb{R}^n : 0 \le x_j \le 2\pi\right\}$$

be a cube in $\mathbb{R}^n$. By this we mean that Q consists of those points $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ such that each component x_j satisfies $0 \le x_j \le 2\pi$. Consider n-tuples of integers $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{Z}^n$, where each $\mu_j \ge 0$, and write $\mu x = \mu_1 x_1 + \cdots + \mu_n x_n$. For $t \in \mathbb{R}$ we let H_t be the set of all series of the form:

$$u = \sum a_\mu e^{i\mu x} \tag{59}$$

where the a_μ are complex numbers satisfying:

$$\alpha_{-\mu} = \bar{\alpha}_\mu$$

(this is done to keep the functions u real valued), and

$$\|u\|_t^2 = (2\pi)^n \sum (1+\mu^2)^t |\alpha_\mu|^2 < \infty \tag{60}$$

where $\mu^2 = \mu_1^2 + \cdots + \mu_n^2$. It is not required that the series (59) converge in anyway, but only that Eq. (60) hold. If

$$u = \sum a_\mu e^{i\mu x}, \qquad v = \sum \beta_\mu e^{i\mu x}$$

are members of H_t, we can introduce the scalar product:

$$(u, v)_t = (2\pi)^n \sum (1 + \mu^2)^t \alpha_\mu \beta_{-\mu} \tag{61}$$

With this scalar product, H_t becomes a Hilbert space. If

$$f(x) = \sum \gamma_\mu e^{i\mu x} \tag{62}$$

we wish to solve

$$-\Delta u = f \tag{63}$$

In other words, we wish to solve:

$$\sum \mu^2 \alpha_\mu e^{i\mu x} = \sum \gamma_\mu e^{i\mu x}$$

This requires:

$$\mu^2 \alpha_\mu = \gamma_\mu \quad \forall \mu$$

In order to solve for α_μ, we must have

$$\gamma_0 = 0 \tag{64}$$

Hence, we cannot solve for all f. However, if Eq. (64) holds, we can solve Eq. (63) by taking:

$$\alpha_\mu = \gamma_\mu / \mu^2 \ \ when \ \ \mu \neq 0 \tag{65}$$

On the other hand, we can take α_0 to be any number we like, and u will be a solution of Eq. (63) as long as it satisfies Eq. (65). Thus, we have Theorem 0.1:

Theorem 0.1. *If f, given by Eq. (62), is in H_t and satisfies Eq. (64), then Eq. (63) has a solution $u \in H_{t+2}$. An arbitrary constant can be added to the solution.*

If we wish to solve:

$$-(\Delta + \lambda)u = f \tag{66}$$

where $\lambda \in \mathbb{R}$ is any constant, we want $(\mu^2 - \lambda)\alpha_\mu = \gamma_\mu$, or

$$\alpha_\mu = \gamma_\mu / (\mu^2 - \lambda) \ \ when \ \ \mu^2 \neq \lambda \tag{67}$$

Therefore, we must have

$$\gamma_\mu = 0 \ \ when \ \ \mu^2 = \lambda \tag{68}$$

Not every λ can equal some μ^2. This is certainly true if $\lambda < 0$ or if $\lambda \notin \mathbb{Z}$. But even if λ is a positive integer, there may be no μ such that:

$$\lambda = \mu^2 \tag{69}$$

For instance, if $n = 3$ and $\lambda = 7$ or 15, there is no μ satisfying Eq. (69). Thus, there is a subset:

$$0 = \lambda_0 < \lambda_1 < \cdots < \lambda_k < \cdots$$

of the positive integers for which there are n-tuples μ satisfying Eq. (69). For $\lambda = \lambda_k$, we can solve Eq. (66) only if Eq. (68) holds. In that case, we can solve by taking α_μ to be given by Eq. (67) when Eq. (69) does not hold, and to be given arbitrarily when Eq. (69) does hold. On the other hand, if λ is not equal to any λ_k, then, Eq. (69) never holds, and we can solve Eq. (66) by taking the α_μ to satisfy Eq. (67). Thus, we have Theorem 0.2:

Theorem 0.2. *There is a sequence $\{\lambda_k\}$ of nonnegative integers tending to $+\infty$ with the following properties. If $f \in H_t$ and $\lambda \neq \lambda_k$ for every k, then there is a unique solution $u \in H_{t+2}$ of Eq. (66). If $\lambda = \lambda_k$ for some k, then one can solve Eq. (66) only if f satisfies Eq. (68). The solution is not unique; there is a finite number of linearly independent periodic solutions of:*

$$(\Delta + \lambda_k)u = 0 \tag{70}$$

which can be added to the solution.

The values λ_k for which Eq. (70) has a nontrivial solution (i.e., a solution which is not $\equiv 0$) are called *eigenvalues*, and the corresponding nontrivial solutions are called *eigen functions.*

To analyze the situation a bit further, suppose $\lambda = \lambda_k$ for some k, and $f \in H_t$ is given by Eq. (62) and satisfies Eq. (68). If $v \in H_t$ is given by:

$$v = \sum \beta_\mu e^{i\mu x} \tag{71}$$

then

$$(f, v)_t = (2\pi)^n \sum (1 + \mu^2)^t \gamma_\mu \beta_{-\mu}$$

by Eq. (61). Hence, we have

$$(f, v)_t = 0$$

for all $v \in H_t$ satisfying Eq. (71) and

$$\beta_\mu = 0 \ \ when \ \ \mu^2 \neq \lambda_k \tag{72}$$

On the other hand,

$$(\Delta + \lambda_k)v = \sum (\lambda_k - \mu^2)\beta_\mu e^{i\mu x} \tag{73}$$

Thus, v is a solution of Eq. (70) if and only if it satisfies Eq. (72). Conversely, if $(f, v)_t = 0$ for all v satisfying Eqs. (71) and (72), then f satisfies Eq. (68). Combining these, we have Theorem 0.3:

Theorem 0.3. If $\lambda = \lambda_k$ for some k, *then there is a solution $u \in H_{t+2}$ of Eq. (66) if and only if*:

$$(f, v)_t = 0 \tag{74}$$

for all $v \in H_t$ satisfying:

$$(\Delta + \lambda_k)v = 0 \tag{75}$$

Moreover, any solution of Eq. (75) can be added to the solution of Eq. (66).

What are the solutions of Eq. (75)? By Eq. (73), they are of the form:

$$v = \sum_{\mu^2=\lambda_k} \beta_\mu e^{i\mu x} = \sum_{\mu^2=\lambda_k} [a_\mu \cos \mu x + b_\mu \sin \mu x]$$

where we took $\beta_\mu = \alpha_\mu - ib_\mu$.Thus, Eq. (74) becomes:

$$(f, \cos \mu x)_t = 0, (f, \sin \mu x)_t = 0 \ \ when \ \ \mu^2 = \lambda_k$$

The results obtained for periodic solutions are indicative of almost all regular boundary-value problems.

I. Sobolev's Inequalities

Certain inequalities are very useful in solving problems in partial differential equations. The following are known as the Sobolev inequalities.

Theorem 0.4. *For each* $p \geq 1, q \geq 1$ *satisfying*:

$$\frac{1}{p} \leq \frac{1}{q} + \frac{1}{n} \tag{76}$$

there is a constant C_{pq} *such that:*

$$|u|_q \leq C_{pq}(|\nabla u|_p + |u|_p), \qquad u \in C^1(Q) \tag{77}$$

where

$$|u|_q = \left(\int_Q |u|^q \, dx \right)^{1/q} \tag{78}$$

and

$$|\nabla u| = \left(\sum_{k=1}^{n} \left| \frac{\partial u}{\partial x_k} \right|^2 \right)^{1/2} \tag{79}$$

As a corollary, we have Corollary 0.1:

Corollary 0.1. $|u|_q \leq C_q ||u||_1, \ u \in H_1, \ where \ 1 \leq q \leq 2^* := 2n/(n-2)$.

SEE ALSO THE FOLLOWING ARTICLES

CALCULUS • DIFFERENTIAL EQUATIONS, ORDINARY • GREEN'S FUNCTIONS

BIBLIOGRAPHY

Bers, L., John, F., and Schechter, M. (1979). "Partial Differential Equations," American Mathematical Society, Providence, RI.

Courant, R., and Hilbert, D. (1953, 1962). "Methods of Mathematical Physics: I, II," Wiley–Interscience, New York.

Gilbarg, D., and Trudinger, N. S. (1983). "Elliptic Partial Differential Equations of Second Order," Springer-Verlag, New York.

Gustofsen, K. E. (1987). "Partial Differential Equations and Hilbert Space Methods," Wiley, New York.

John, F. (1978). "Partial Differential Equations," Springer-Verlag, New York.

Lions, J. L., and Magenes, E. (1972). "Non-Homogeneous Boundary Value Problems and Applications," Springer-Verlag, New York.

Powers, D. L. (1987). "Boundary Value Problems," Harcourt Brace Jovanovich, San Diego, CA.

Schechter, M. (1977). "Modern Methods in Partial Differential Equations: An Introduction," McGraw–Hill, New York.

Schechter, M. (1986). "Spectra of Partial Differential Operators," North-Holland, Amsterdam.

Treves, F. (1975). "Basic Linear Partial Differential Equations," Academic Press, New York.

Zauderer, E. (1983). "Partial Differential Equations of Applied Mathematics," Wiley, New York.



Discrete Mathematics and Combinatorics

Douglas R. Shier
Clemson University

I. Nature of Combinatorics
II. Basic Counting Techniques
III. Recurrence Relations and Generating Functions
IV. Inclusion–Exclusion Principle
V. Existence Problems

GLOSSARY

Algorithm Systematic procedure or prescribed series of steps followed in order to solve a problem.
Binary Pertaining to the digits 0 and 1.
Event Set of occurrences defined with respect to some probabilistic process.
Identity Mathematical equation that always holds.
Integers The numbers 0, 1, 2, . . . and their negatives.
List Ordered sequence of elements.
Mutually exclusive Events that cannot occur simultaneously.
Prime Integer greater than 1 that cannot be evenly divided by any integer other than itself and 1.
Set Unordered collection of elements.
String Ordered sequence of letters taken from some alphabet.
Universal set Set that contains all elements relevant to the current discussion.

COMBINATORICS is a branch of discrete mathematics that involves the study of arrangements of various objects. Typically, the focus of combinatorics is on determining whether arrangements can be found that satisfy certain properties or on counting all possible arrangements of such objects. While the roots of combinatorics extend back several thousands of years, its relevance to modern science and engineering is increasingly evident.

I. NATURE OF COMBINATORICS

Combinatorics constitutes a rapidly growing area of contemporary mathematics and is one with an enviable repertoire of applications to areas as diverse as biology, chemistry, physics, engineering, communications, cryptography, and computing. Of particular significance is its symbiotic relationship to the concerns and constructs of computer science. On the one hand, the advent of high-speed computers has facilitated the detailed study of existing combinatorial patterns as well as the discovery of new arrangements. On the other hand, the design and analysis of computer algorithms frequently require the insights and tools of combinatorics. It is not at all surprising, then, that computer science, which is ultimately concerned with

the manipulation of finite sets of symbols (e.g., strings of binary digits), and combinatorial mathematics, which provides tools for analyzing such patterns of symbols, have rapidly achieved prominence together. Moreover, since the symbols themselves can be abstract objects (rather than simply numerical quantities), combinatorics supports the more abstract manipulations of symbolic mathematics and symbolic computer languages.

Combinatorics is at heart a problem-solving discipline that blends mathematical techniques and concepts with a necessary touch of ingenuity. In order to emphasize this dual nature of combinatorics, the sections that follow will first present certain fundamental combinatorial principles and then illustrate their application through a number of diverse examples. Specifically, Sections II–IV provide an introduction to some powerful techniques for counting various combinatorial arrangements, and Section V examines when certain patterns can be guaranteed to exist.

II. BASIC COUNTING TECHNIQUES

A. Fundamental Rules of Sum and Product

Two deceptively simple, but fundamentally important, rules allow the counting of complex patterns by decomposition into simpler patterns. The first such principle states, in essence, that if we slice a pie into two nonoverlapping portions, then indeed the whole (pie) is equal to the sum of its two parts.

Rule of Sum. Suppose that event E can occur in m different ways, that event F can occur in n different ways, and that the two events are mutually exclusive. Then, the compound event where at least one of the two events happens can occur in $m + n$ ways.

The second principle indicates the number of ways that a menu of choices (one item chosen from E, another item chosen from F) can be selected.

Rule of Product. Suppose that event E can occur in m different ways and that subsequently event F can occur in n different ways. Then, a choice from E followed by a choice from F can be made in $m \times n$ ways.

EXAMPLE 1. A certain state anticipates a total of 2,500,000 registered vehicles within the next ten years. Can the current system of license plates (consisting of six digits) accommodate the expected number of vehicles? Should there instead be a change to a proposed new system consisting of two letters followed by four digits?

Solution. To analyze the current situation, there are ten possibilities (0–9) for each of the six digits, so application of the product rule yield $10 \times 10 \times 10 \times 10 \times 10 \times 10 = 1{,}000{,}000$ possibilities, not enough to accommodate the expected number of vehicles. By contrast, the proposed new system allows (again by the product rule) $26 \times 26 \times 10 \times 10 \times 10 \times 10 = 6{,}760{,}000$ possibilities, more than enough to satisfy the anticipated demand.

EXAMPLE 2. DNA (deoxyribonucleic acid) consists of a chain of nucleotide bases (adenine, cytosine, guanine, thymine). How many different three-base sequences are possible?

Solution. For each of the three positions in the sequence, there are four possibilities for the base, so (by the product rule) there are $4 \times 4 \times 4 = 64$ such sequences.

EXAMPLE 3. In a certain computer programming language, each identifier (variable name) consists of either one or two alphanumeric characters (A–Z, 0–9), but the first character must be alphabetic (A–Z). How many different identifier names are possible in this language?

Solution. In this case, analysis of the compound event can be broken into counting the possibilities for event E, a single-character identifier, and for event F, a two-character identifier. The number of possibilities for E is 26, whereas (by the product rule) the number of possibilities for F is $26 \times (26 + 10) = 936$. Since the two events E and F are mutually exclusive, the total number of distinct identifiers is $26 + 936 = 962$.

B. Permutations and Combinations

In the analysis of combinatorial problems, it is essential to recognize when order is important in the arrangement and when it is not. To emphasize this distinction, the set $X = [x_1, x_2, \ldots, x_n]$ consists of n elements x_i, assembled without regard to order, whereas the list $X = [x_1, x_2, \ldots, x_n]$ contains elements arranged in a prescribed order.

In the previous examples, the order of arrangement was clearly important so lists were implicitly being counted. More generally, arrangements of objects into a list are referred to as *permutations*. For example, the objects a, b, c can be arranged into the following permutations: $[a, b, c]$, $[a, c, b]$, $[b, a, c]$, $[b, c, a]$, $[c, a, b]$, $[c, b, a]$. By the product rule, n distinct objects can be arranged into:

$$n! = n \times (n - 1) \times (n - 2) \times \cdots \times 2 \times 1$$

different permutations. (The symbol $n!$, or n factorial, denotes the product of the first n positive integers.) A permutation of size k is a list with k elements chosen from the n given objects, and there are exactly

$$P(n, k) = n \times (n - 1) \times (n - 2) \times \cdots \times (n - k + 1)$$

such permutations.

EXAMPLE 4. In a manufacturing plant, a particular product is fabricated by processing in turn on four different machines. If any processing sequence using all four machines is permitted, how many different processing orders are possible? How many processing orders are there if only two machines from the four need to be used?

Solution. Each processing order corresponds to a permutation of the four machines, so there are $P(4, 4) = 4! = 24$ different orders. If processing on any two machines is allowable then there are $P(4, 2) = 4 \times 3 = 12$ different orders.

When the order of elements occurring in the arrangement is not pertinent, then a way of arranging k objects, chosen from n distinct objects, is called a *combination* of size k. For example, the objects a, b, c can be arranged into the following combinations, or sets, of size 2: $\{a, b\}, \{a, c\}, \{b, c\}$. The number of combinations of size k from n objects is given by the formula:

$$C(n, k) = P(n, k)/k!$$

EXAMPLE 5. A group of ten different blood samples is to be split into two batches, each consisting of five "pooled" samples. Further chemical analysis will then be performed on the two batches. In how many ways can the samples be split in this fashion?

Solution. Any division of the samples $S_1, S_2, \ldots, S_{10}$ into the two batches can be uniquely identified by those samples belonging to the first batch. For example, $\{S_1, S_2, S_5, S_6, S_8\}$ defines one such division. Since the order of samples within each batch is not important, there are $C(10, 5) = 252$ ways to divide the original samples.

EXAMPLE 6. Suppose that 12 straight lines are drawn on a piece of paper, with no two lines being parallel and no three meeting at a single point. How many different triangles are formed by these lines?

Solution. Any three lines form a triangle since no lines are parallel. As a result, there are as many triangles as choices of three lines selected from the 12, giving $C(12, 3) = 220$ such triangles.

EXAMPLE 7. How many different solutions are there in nonnegative integers x_i to the equation $x_1 + x_2 + x_3 + x_4 = 8$?

Solution. We can view this problem as an equivalent one in which eight balls are placed into four numbered boxes. For example, the solution $x_1 = 2, x_2 = 3, x_3 = 2, x_4 = 1$ corresponds to placing 2, 3, 2, 1 balls into boxes 1, 2, 3, 4. This solution can also be represented by the string * * | * * * | * * | * which shows the number of balls residing in the four boxes. The number of solutions is then the number of ways of constructing a string of 11 symbols (eight stars and three bars); namely, we can select the three bars in $C(11, 3) = 165$ ways.

C. Binomial Coefficients

Ways of arranging objects can also be viewed from an algebraic perspective. To understand this correspondence, consider the product of n identical factors $(1 + x)$, namely:

$$(1 + x)^n = (1 + x)(1 + x) \cdots (1 + x)$$

The coefficient of x^k in the expansion of this product is juh the number of ways to select the symbol x from exactly k of the factors. However, the number of ways to select these k factors from the n available is $C(n, k)$, meaning that:

$$(1 + x)^n = C(n, 0) + C(n, 1)x + C(n, 2)x^2 + \cdots + C(n, n)x^n \quad (1)$$

Because the coefficients $C(n, k)$ arise in this way from the expansion of a two-term expression, they are also referred to as *binomial coefficients*. These coefficients can be convenienthvaced in a triangular array, called *Pascal's triangle*, as shown in Fig. 1. Row n of Pascal's triangle contains the values $C(n, 0), C(n, 1), \ldots, C(n, n)$. Several patterns are apparent from this figure. First, the binomial coefficients are symmetrically placed within each row: namely, $C(n, k) = C(n, n - k)$. Second, the coefficient appearing in any row equals the sum of the two coefficients appearing in the previous row just to the left and to the right. For example, in row 5 the third entry, 10, is the sum of the second and third entries, 4 and 6, from the previous row. In general, the binomial coefficients satisfy the identity:

$$C(n, k) = C(n - 1, k - 1) + C(n - 1, k)$$

The binomial coefficients satisfy a number of other interesting and useful identities. To illustrate one way of

n = 1: 1 1

n = 2: 1 2 1

n = 3: 1 3 3 1

n = 4: 1 4 6 4 1

n = 5: 1 5 10 10 5 1

n = 6: 1 6 15 20 15 6 1

FIGURE 1 Arrangement of binomial coefficients $C(n, k)$ in Pascal's triangle.

discovering such identities, formally substitute the value $x = 1$ into both sides of Eq. (1), yielding the identity:

$$2^n = C(n, 0) + C(n, 1) + C(n, 2) + \cdots + C(n, n)$$

In other words, the binomial coefficients for n must sum to the value 2^n. If instead, the value $x = -1$ is formally substituted into Eq. (1), the following identity results:

$$0 = C(n, 0) - C(n, 1) + C(n, 2) - \cdots + (-1)^n C(n, n)$$

This simply states that the alternating sum of the binomial coefficients in any row of Fig. 1 must be zero.

A final identity involving the numbers in Fig. 1 concerns a string of coefficients progressing from the left-hand border along a downward sloping diagonal to any other entry in the figure. Then, the sum of these coefficients will be found as the value just below and to the left of the last such entry. For instance, the sum $C(2, 0) + C(3, 1) + C(4, 2) + C(5, 3) = 1 + 3 + 6 + 10 = 20$ is indeed the same as the binomial coefficient $C(6, 3)$. In general, this observation can be expressed as the identity:

$$C(n + k + 1, k) = C(n, 0) + C(n + 1, 1) + C(n + 2, 2) + \cdots + C(n + k, k)$$

This identity can be given a pleasant combinatorial interpretation, namely, consider selecting k items (without regard for order) from a total of $n + k + 1$ items, which can be done in $C(n + k + 1, k)$ ways. In any such selection, there will be some item number r so that items $1, 2, \ldots, r$ are selected but $r + 1$ is not. This then leaves $k - r$ items to be selected from the remaining $n + k + 1 - (r + 1) = n + k - r$ items, which can be done in $C(n + k - r, k - r)$ ways. Since the cases $r = 0, 1, \ldots, k$ are mutually exclusive, the sum rule shows the total number of selections is also equal to $C(n + k, k) + C(n + k - 1, k - 1) + \cdots + C(n + 1, 1) + C(n, 0)$. Thus, by counting the same group of objects in two different ways, one can verify the above identity. This technique of "double counting" provides a powerful tool applicable to a number of other combinatorial problems.

D. Discrete Probability

Probability theory is an important area of mathematics in which combinatorics plays an essential role. For example, if there are only a finite number of outcomes $S_1, S_2, \ldots, S_m$ to some process, the ability to count the number of occurrences of S_i provides valuable information on the likelihood that outcome S_i will in fact be observed. Indeed, many phenomena in the physical sciences are governed by probabilistic rather than deterministic laws; therefore, one must generally be content with assessing the probability that certain desirable (or undesirable) outcomes will occur.

EXAMPLE 8. What is the probability that a hand of five cards, dealt from a shuffled deck of cards, contains at least three aces?

Solution. The population of 52 cards can be conveniently partitioned into set A of the 4 aces and set N of the 48 non-aces. In order to obtain exactly three aces, the hand must contain three cards from set A and two cards from set N, which can be achieved in $C(4, 3)\,C(48, 2) = 4512$ ways. To obtain exactly 4 aces, the hand must contain four cards from A and one card from N, which can be achieved in $C(4, 4)\,C(48, 1) = 48$ ways. Since the total number of possible hands of five cards chosen from the 52 is $C(52, 5) = 2{,}598{,}960$, the probability of the required hand is $(4512 + 48)/2{,}598{,}960 = 0.00175$, indicating a rate of occurrence of less than twice in a thousand.

EXAMPLE 9. In a certain state lottery, six winning numbers are selected from the numbers $1, 2, \ldots, 40$. What are the odds of matching all six winning numbers? What are the odds of matching exactly five? Exactly four?

Solution. The number of possible choices is the number of ways of selecting six numbers from the 40, or $C(40, 6) = 3{,}838{,}380$. Since only one of these is the winning selection, the odds of matching all six numbers is $1/3{,}838{,}380$. To match exactly five of the winning numbers, there are $C(6, 5) = 6$ ways of selecting the five matching numbers and $C(34, 1) = 34$ ways of selecting a nonmatching number, giving (by the product rule) $6 \times 34 = 204$ ways, so the odds are $204/3{,}838{,}380 = 17/319{,}865$ for matching five numbers. To match exactly four winning numbers, there are $C(6, 4) = 15$ ways of selecting the four matching numbers and $C(34, 2) = 561$ ways of selecting the nonmatching numbers, giving $15 \times 561 = 8415$ ways, so the odds are $8415/3{,}838{,}380 = 561/255{,}892$ (or approximately 0.0022) of matching four numbers.

EXAMPLE 10. An alarm system is constructed from five identical components, each of which can fail (independently of the others) with probability q. The system is designed with a certain amount of redundancy so that it functions whenever at least three of the components are working. How likely is it that the entire system functions?

Solution. There are two states for each individual component, either good or failed. The state of the system can be represented by a binary string $x_1x_2x_3x_4x_5$, where x_i is 1 if component i is good and is 0 if it fails. A functioning state for the system thus corresponds to a binary string having at most two zeros. The number of states with exactly two zeros is $C(5, 2) = 10$, so the probability

of two failed and three good components is $10q^2(1-q)^3$. Similarly, the probability of exactly one failed component is $C(5,1)q^1(1-q)^4 = 5q(1-q)^4$, and the probability of no failed components is $C(5,0)(1-q)^5 = (1-q)^5$. Altogether, the probability that the system functions is given by $10q^2(1-q)^3 + 5q(1-q)^4 + (1-q)^5$. For example, when $q = 0.01$, the system will operate with probability 0.99999015 and thus fail with probability 0.00000985; this shows how adding redundancy to a system composed of unreliable components (1% failure rate) produces a highly reliable system (0.001% failure rate).

III. RECURRENCE RELATIONS AND GENERATING FUNCTIONS

A. Recurrence Relations and Counting Problems

Not all counting problems can be solved as readily and as directly as in Section II. In fact, the best way to solve specific counting problems is often to solve instead a more general, and presumably more difficult, problem. One technique for doing this involves the use of recurrence relations.

Recall that the binomial coefficients satisfy the relation:

$$C(n,k) = C(n-1,k-1) + C(n-1,k)$$

Such an expression shows how the value $C(n,k)$ can be calculated from certain "prior" values $C(n-1,k-1)$ and $C(n-1,k)$. This type of relation is termed a recurrence relation, since it enables any specific value in the sequence to be obtained from certain previously calculated values.

EXAMPLE 11. How many strings of eight binary digits contain no consecutive pair of zeros?

Solution. It is easy to find the number f_1 of such strings of length 1, since the strings "0" and "1" are both acceptable, yielding $f_1 = 2$. Also, the only forbidden string of length 2 is "00" so $f_2 = 3$. There are three forbidden strings of length 3 ("001," "100," and "000"), whereupon $f_3 = 5$. At this point, it becomes tedious to calculate subsequent values directly, but they can be easily found by noticing that a certain recurrence relation governs the sequence f_n. In an acceptable string of length n, either the first digit is a 1 or it is a 0. In the former case, the remaining digits can be any acceptable string of length $n-1$ (and there are f_{n-1} of these). In the latter case, the second digit must be a 1 and then the remaining digits must form an acceptable string of length $n-2$ (there are f_{n-2} of these). These observations provide the recurrence relation:

$$f_n = f_{n-1} + f_{n-2} \qquad (2)$$

Using the initial conditions $f_1 = 2$ and $f_2 = 3$, the values $f_3, f_4, \ldots, f_8$ can be calculated in turn by substitution into Eq. (2):

$$f_3 = 5, \qquad f_4 = 8, \qquad f_5 = 13,$$
$$f_6 = 21, \qquad f_7 = 34, \qquad f_8 = 55$$

Therefore, 55 binary strings of length 8 have the desired property.

In this problem it was clearly expedient to solve the general problem by use of a recurrence relation that stressed the interdependence of solutions to related problems. The particular sequence obtained for this problem, $[1, 2, 3, 5, 8, 13, 21, 34, \ldots]$, with $f_0 = 1$ added for convenience, is called the *Fibonacci sequence*, and it arises in numerous problems of mathematics as well as biology, physics, and computer science.

EXAMPLE 12. Suppose that ten straight lines are drawn in a plane so that no two lines are parallel and no three intersect at a single point. Into how many different regions will the plane be divided by these lines?

Solution. As seen in Fig. 2, the number of regions created by one straight line is $f_1 = 2$, the number created by two lines is $f_2 = 4$, and the number created by three lines is $f_3 = 7$. The picture becomes excessively complicated with more added lines, so it is prudent to seek a general solution for f_n, the number of regions created by n lines in the plane. Suppose that $n-1$ lines have already been drawn and that line n is now added. Because the lines are all mutually nonparallel, line n must intersect each existing line exactly once. These $n-1$ intersection points divide the new line into n segments and each segment serves to subdivide an existing region into two regions. Thus, the n segments increase the number of regions by exactly n, producing the recurrence relation:

$$f_n = f_{n-1} + n$$

Given the initial condition $f_1 = 2$, application of this recurrence relation yields the values $f_2 = f_1 + 2 = 4$ and $f_3 = f_2 + 3 = 7$, as previously verified. In fact, such a recurrence relation can be explicitly solved, giving:

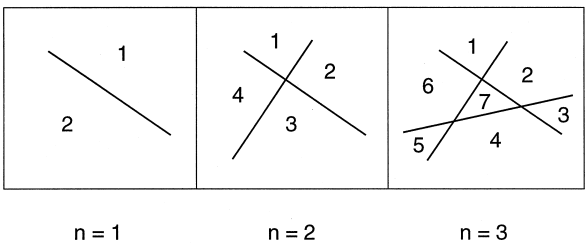


FIGURE 2 Number of regions created by the placement of n lines.

$$f_n = (n^2 + n + 2)/2$$

indicating that there are $f_{10} = 112/2 = 56$ regions bounded by ten lines.

In representing mathematical expressions, the placement of parentheses can be crucial. For example, $((x - y) - z)$ does not in general give the same result as $(x - (y - z))$. Moreover, the placement of parentheses must be syntactically valid as in $((x - y) - z)$ or $(x - (y - z))$, whereas $(x) - y) - (z$ is not syntactically valid.

EXAMPLE 13. How many valid ways are there of parenthesizing an expression involving n variables?

Solution. Notice that there is only one valid form for a single variable x, namely (x), giving $f_1 = 1$; similarly, it can be verified that $f_2 = 1$ and $f_3 = 2$. More generally, suppose that there are n variables. Then any valid form must be expressible as (AB), where A and B are themselves valid forms. Notice that if A contains k variables, then B must contain $n - k$ variables, so in this case the f_k valid forms for A can be combined with the f_{n-k} valid forms for B to yield $f_k f_{n-k}$ valid forms for the whole expression. Since the different possibilities for k ($k = 1, 2, \ldots, n - 1$) are mutually exclusive, we obtain the recurrence relation:

$$f_n = f_1 f_{n-1} + f_2 f_{n-2} + \cdots + f_{n-2} f_2 + f_{n-1} f_1$$

This equation can be explicitly solved for f_n in terms of binomial coefficients:

$$f_n = C(2n - 2, n - 1)/n$$

The numbers in this sequence, $[1, 1, 2, 5, 14, 42, 132, \ldots]$, are called *Catalan numbers* (Eugène Catalan, 1814–1894), and they occur with some regularity as solutions to a variety of combinatorial problems (such as triangulating convex polygons, counting voting sequences, and constructing rooted binary trees).

B. Generating Functions and Counting Problems

The examples in Part A of this section serve to illustrate the theme that solving a specific problem can frequently be aided by relating the given problem to other, often simpler, problems of the same type. For example, a problem of size n might be related to a problem of size $n - 1$ and to another problem of size $n - 3$. Another way of pursuing such interrelationships among problems of different sizes is through use of a generating function. As a matter of fact, this concept has already been previewed in studying the binomial coefficients. Specifically, Eq. (1) shows that $f(x) = (1 + x)^n$ can be viewed as a generating function for the binomial coefficients $C(n, k)$:

$$f(x) = C(n, 0) + C(n, 1)x + C(n, 2)x^2 + \cdots + C(n, n)x^n$$

The variable x simply serves as a formal symbol and its exponents represent placeholders for carrying the coefficient information. More generally, a generating function is a polynomial in the variable x:

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \cdots$$

and it serves as a template for studying the sequence of coefficients $[a_0, a_1, a_2, \ldots, a_n, \ldots]$.

Recall that the binomial coefficients $C(n, k)$ count the number of combinations of size k derived from a set $\{1, 2, \ldots, n\}$ of n elements. In this context, the generating function $f(x) = (1 + x)^n$ for the binomial coefficients can be developed by the following reasoning. At each step $k = 1, 2, \ldots, n$ a decision is made as to whether or not to include element k in the current combination. If x^0 is used to express exclusion and x^1 inclusion, then the factor $(x^0 + x^1) = (1 + x)$ at step k compactly encodes these two choices. Since each element k presents the same choices (exclude/include), the product of the n factors $(1 + x)$ produces the desired enumerator $(1 + x)^n$. This reasoning applies more generally to cases where the individual choices at each step are not identical, so the factors need not all be the same (as in the case of the binomial coefficients). The following examples give some idea of the types of problems that can be addressed through the use of generating functions.

EXAMPLE 14. In how many different ways can change for a dollar be given, using only nickels, dimes, and quarters?

Solution. The choices for the number of nickels to use can be represented by the polynomial:

$$(1 + x^5 + x^{10} + \cdots) = (1 - x^5)^{-1}$$

where x^i signifies that exactly i cents worth of nickels are used. Similarly, the choices for dimes are embodied in:

$$(1 + x^{10} + x^{20} + \cdots) = (1 - x^{10})^{-1}$$

and the choices for quarters in:

$$(1 + x^{25} + x^{50} + \cdots) = (1 - x^{25})^{-1}$$

Multiplying together these three polynomials produces the required generating function:

$$f(x) = (1 - x^5)^{-1}(1 - x^{10})^{-1}(1 - x^{25})^{-1}$$

The coefficient of x^n in the expanded form of this generating function indicates the number of ways of making change for n cents. In particular, there are 29 different ways of making change for a dollar ($n = 100$). This can be verified by using a symbolic algebra package such as *Mathematica* or *Maple*.

A *partition* of the positive integer n is a set of positive integers (or "parts") that together sum to n. For example,

the number 4 has the five partitions: {1, 1, 1, 1}, {1, 1, 2}, {1, 3}, {2, 2}, and {4}.

EXAMPLE 15. How many partitions are there for the integer n?

Solution. The choices for the number of ones to include as parts is represented by the polynomial:

$$(1 + x + x^2 + \cdots) = (1 - x)^{-1}$$

where the x^i term means that 1 is to appear i times in the partition. Similarly, the choices for the number of twos to include is given by:

$$(1 + x^2 + x^4 + \cdots) = (1 - x^2)^{-1}$$

the choices for the number of threes is given by:

$$(1 + x^3 + x^6 + \cdots) = (1 - x^3)^{-1}$$

and so forth. Therefore, the number of partitions of n can be found as the coefficient of x^n in the generating function:

$$f(x) = (1 - x)^{-1}(1 - x^2)^{-1}(1 - x^3)^{-1} \cdots$$

EXAMPLE 16. Find the number of partitions of the integer n into *distinct* parts.

Solution. Since the parts must be distinct, the choices for any integer i are whether to include it (x^i) or not (x^0) in the given partition. As a result, the generating function for this problem is

$$f(x) = (1 + x)(1 + x^2)(1 + x^3) \cdots$$

For example, the coefficient of x^8 in the expansion of $f(x)$ is found to be 6, meaning that there are six partitions of 8 into distinct parts: namely, {8}, {1, 7}, {2, 6}, {3, 5}, {1, 2, 5}, {1, 3, 4}.

IV. INCLUSION–EXCLUSION PRINCIPLE

Another important counting technique is based on the idea of successively adjusting an initial count through systematic additions and subtractions that are guaranteed to produce a correct final answer. This technique, called the *inclusion–exclusion principle*, is applicable to many instances where direct counting would be impractical.

As a simple example, suppose we wish to count the number of elements that are *not* in some subset A of a given universal set U. Then, the required number of elements equals the total number of elements in U, denoted by $N = N(U)$, minus the number of elements in A, denoted by $N(A)$. Expressed in this notation,

$$N(A') = N - N(A)$$

where $A' = U - A$ designates the set of elements in U that do not appear in A. Figure 3a depicts this relation using a Venn diagram (John Venn, 1834–1923), in which the enclosing rectangle represents the set U, the inner circle represents A, and the shaded portion represents A'. The quantity $N(A')$ is thus obtained by excluding $N(A)$ elements from N.

EXAMPLE 17. The letters a, b, c, d, e are used to form five-letter words, using each letter exactly once. How many words do *not* contain the sequence *bad*?

Solution. The universe here consists of all words, or permutations, formed from the five letters, so there are $N = 5! = 120$ words in total. Set A consists of all such words containing *bad*. By treating these three letters as a new "megaletter" x, the set A equivalently contains all words formed from x, c, e so $N(A) = 3! = 6$. The number of words not containing *bad* is then $N - N(A) = 120 - 6 = 114$.

Figure 3b shows the situation for two sets, A and B, contained in the universal set U. As this figure suggests,

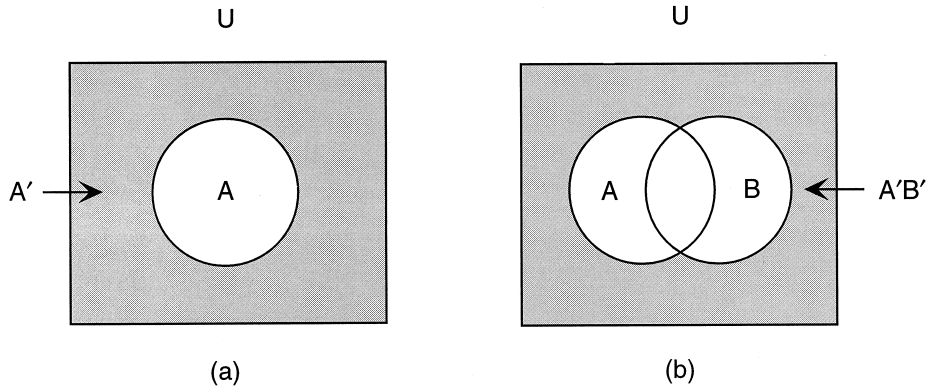


FIGURE 3 Venn diagrams relative to a universal set U. (a) Sets A and A'; (b) Sets A, B, and $A'B'$.

the number of elements in either A or B (or both) is then the number of elements in A plus the number of elements in B, minus the number of elements in both:

$$N(A \cup B) = N(A) + N(B) - N(AB) \qquad (3)$$

Here $A \cup B$ denotes the elements either in A or in B, or in both, whereas AB denotes the elements in both A and B. Since the sum $N(A) + N(B)$ counts the elements of AB twice rather than just once, $N(AB)$ is subtracted to remedy the situation. The number of elements $N(A'B')$ in neither A nor B is $N - N(A \cup B)$, thus an alternative form of Eq. (3) is

$$N(A'B') = N - [N(A) + N(B)] + N(AB) \qquad (4)$$

This form shows how terms are alternately included and excluded to produce the desired result.

EXAMPLE 18. In blood samples obtained from 50 patients, laboratory tests show that 20 patients have antibodies to type A bacteria, 29 patients have antibodies to type B bacteria, and 8 patients have antibodies to both types. How many patients have antibodies to neither of the two types of bacteria?

Solution. The given data state that $N = 50$, $N(A) = 20$, $N(B) = 29$, and $N(AB) = 8$. Therefore, by Eq. (4):

$$N(A'B') = 50 - [20 + 29] + 8 = 9$$

meaning that nine patients are immune to neither type of bacteria.

The foregoing equations generalize in a natural way to three sets A, B, and C:

$$N(A \cup B \cup C) = N(A) + N(B) + N(C) - [N(AB) + N(AC) + N(BC)] + N(ABC) \qquad (5)$$

$$N(A'B'C') = N - [N(A) + N(B) + N(C)] + [N(AB) + N(AC) + N(BC)] - N(ABC) \qquad (6)$$

In each of these forms, the final result is obtained by successive inclusions and exclusions, thus justifying these as manifestations of the inclusion–exclusion principle.

EXAMPLE 19. An electronic assembly is comprised of components 1, 2, and 3 and functions only when at least two components are working. If all components fail independently of one another with probability q, what is the probability that the entire assembly functions?

Solution. Let A denote the event in which components 1 and 2 work, let B denote the event in which components 1 and 3 work, and let C denote the event in which components 2 and 3 work. Of interest is the probability that at least one of the events A, B, or C occurs. The analogous form of Eq. (5) in the probabilistic case is

$$Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C) - [Pr(AB) + Pr(AC) + Pr(BC)] + Pr(ABC)$$

Here, $Pr(A)$ is the probability of event A occurring, $Pr(AB)$ is the probability of event AB occurring, and so forth. Notice that $Pr(A) = Pr(B) = Pr(C) = (1-q)^2$ and $Pr(AB) = Pr(AC) = Pr(BC) = Pr(ABC) = (1-q)^3$, so the assembly functions with probability:

$$Pr(A \cup B \cup C) = 3(1-q)^2 - 3(1-q)^3 + (1-q)^3 = 3(1-q)^2 - 2(1-q)^3$$

Two positive integers are called *relatively prime* if the only positive integer evenly dividing both is the number 1. For example, 7 and 15 are relatively prime, whereas 6 and 15 are not (they share the common divisor 3).

EXAMPLE 20. How many positive integers not exceeding 60 are relatively prime to 60?

Solution. The appropriate universe here is $U = \{1, 2, \ldots, 60\}$ and the (prime) divisors of $N = 60$ are 2, 3, 5. Relative to U, let A be the set of integers divisible by 2, let B be the set of integers divisible by 3, and let C be the set of integers divisible by 5. The problem here is to calculate $N(A'B'C')$, the number of integers that share no divisors with 60. Because every other integer is divisible by 2, we have $N(A) = 60/2 = 30$. Similarly, $N(B) = 60/3 = 20$ and $N(C) = 60/5 = 12$. Because any integer divisible by both 2 and 3 must also be divisible by 6, we have $N(AB) = 60/6 = 10$. Likewise, $N(AC) = 60/10 = 6$, $N(BC) = 60/15 = 4$, and $N(ABC) = 60/30 = 2$. Substituting these values into Eq. (6) gives:

$$N(A'B'C') = 60 - [30 + 20 + 12] + [10 + 6 + 4] - 2 = 16$$

so there are 16 positive numbers not exceeding 60 that are relatively prime to 60.

EXAMPLE 21. Each package of a certain product contains one of three possible prizes. How likely is it that a purchaser of five packages of the product will get at least one of each prize?

Solution. The number of possible ways in which this event can occur will first be calculated, after which a probabilistic statement can be deduced. Let the prizes be denoted a, b, c, and let the contents of the five packages be represented by a string of five letters (where repetition is allowed). For example, the string *bacca* would represent one possible occurrence. Define A to be the set of all such strings that do not include a. In similar fashion, define B (respectively, C) to be the set of strings that do not include b (respectively, c). It is required then to calculate

$N(A'B'C')$, the number of instances in which a, b, and c all occur. The approach will be to use the inclusion–exclusion relation, Eq. (6), to calculate this quantity. Here, N is the number of strings of five letters over the alphabet $\{a, b, c\}$, so (by the product rule) $N = 3^5 = 243$. Also, $N(A)$ is the number of strings from the alphabet $\{b, c\}$, whereupon $N(A) = 2^5 = 32$. By similar reasoning, $N(B) = N(C) = 32$, $N(AB) = N(AC) = N(BC) = 1^5 = 1$, and $N(ABC) = 0$. As a result,

$$N(A'B'C') = 243 - 3(32) + 3(1) - 0 = 150$$

In summary, the total number of possible strings is 243 and all three prizes are obtained in 150 of these cases. If all strings are equally likely (i.e., any package is just as likely to contain each prize), then the required probability is $150/243 = 0.617$, indicating a better than 60% chance of obtaining three different prizes in just five packages. A similar analysis shows that for six packages, the probability of obtaining all three prizes increases to 0.741; for seven packages, there is a 0.826 probability.

V. EXISTENCE PROBLEMS

A. Pigeonhole Principle

In certain combinatorial problems, it may be exceedingly difficult to count the number of arrangements of a prescribed type. In fact, it might not even be clear that any such arrangement actually exists. What is needed then is some guiding mathematical assurance that configurations of the desired type do indeed exist. One such principle is the so-called *pigeonhole principle*. While its statement is overwhelmingly self-evident, its applications range from the simplest to the most challenging problems in combinatorics.

The pigeonhole principle states that if there are more than k objects (or pigeons) to be placed in k locations (or pigeonholes), then some location must house two (or possibly more) such objects. A simple illustration of this principle assures that at least two residents of a town with 400 inhabitants have the same birthday. Here, the objects are the residents and the locations are the 356 possible birthdays. Since there are more objects than locations, some location must contain at least two objects, meaning that some two residents (or more) must share the same birthday. Notice that this principle only guarantees the *existence* of two such residents; it does not give any information about finding them.

EXAMPLE 22. In any selection of five different elements from $\{1, 2, \ldots, 8\}$, there must be some pair of selected elements that sum to 9.

Solution. Here the objects are the numbers $1, 2, \ldots, 8$ and the locations are the four sets $A_1 = \{1, 8\}$, $A_2 = \{2, 7\}$, $A_3 = \{3, 6\}$, and $A_4 = \{4, 5\}$. Notice that for each of these sets its two elements sum to 9. According to the pigeonhole principle, placing five numbers into these four sets results in some set A_i having both its elements selected, so the sum of these two elements (by the construction of set A_i) must equal 9.

EXAMPLE 23. In a room with $n \geq 2$ persons, there must be two persons having exactly the same number of friends in the room.

Solution. The number of possible friendships for any given person ranges from 0 to $n - 1$. However, if $n - 1$ occurs then that person is a friend of everyone else, and (assuming that friendship is a mutual relation) no other person can be without friends. Thus, both 0 and $n - 1$ cannot simultaneously occur in a group of n persons. If $1, 2, \ldots, n - 1$ are the possible numbers of friendships, then using these $n - 1$ numbers as locations for the n persons (objects), the pigeonhole principle assures that some number in $\{1, 2, \ldots, n - 1\}$ must appear twice. A similar result can be established for the case $\{0, 1, \ldots, n - 2\}$, demonstrating there must always be at least two persons having the same number of friends in the room.

Two strings $x = x_1 x_2 \cdots x_n$ and $y = y_1 y_2 \cdots y_m$ over the alphabet $\{a, b, \ldots, z\}$ are said to be disjoint if they share no common letter and are said to overlap otherwise.

EXAMPLE 24. In any collection of at least six strings, there must either be three strings that are mutually disjoint or three strings that mutually overlap.

Solution. For a given string x, let the $k \geq 5$ other strings in the collection be divided into two groups, D and O; D consists of those strings that are disjoint from x, and O consists of those that overlap with x. By a generalization of the pigeonhole principle, one of these two sets must contain at least three elements. Suppose that it is set D. Then, either D contains three mutually overlapping strings or it contains two disjoint strings y and z. In the first case, these three strings satisfy the stated requirements. In the second case, the elements x, y, z are all mutually disjoint, so again the requirements are met. A similar argument can be made if O is the set containing at least three elements. In any event, there will either be three mutually disjoint strings or three mutually overlapping strings.

This last example is a special case of *Ramsey's theorem* (Frank Ramsey, 1903–1930), which guarantees that if there are enough objects then configurations of certain types will always be guaranteed to exist. Not only does this theorem (which generalizes the pigeonhole principle) produce some very deep combinatorial results, but it has also

been applied to problems arising in geometry, the design of communication networks, and information retrieval.

B. Combinatorial Designs

Combinatorial designs involve ways of arranging objects into various groups in order to meet specified requirements. Such designs find application in the planning of statistical experiments as well as in other areas of mathematics (number theory, coding theory, geometry, and algebra).

As one illustration, suppose that an experiment is to be designed to test the effects of five different drugs using five different subjects. One clear requirement is that each subject should receive all five drugs, since otherwise the results could be biased by variation among the subjects. Each drug is to be administered for one week, so that at the end of five weeks the experiment will be completed. However, the order in which drugs are administered could also have an effect on their observed potency, so it is also desirable for all drugs to be represented on any given week of the experiment. One way of designing such an experiment is depicted in Fig. 4, which shows one source of variation—the subjects ($S_1, S_2, \ldots, S_5$)—appearing along the rows and the other source of variation—the weeks ($W_1, W_2, \ldots, W_5$)—appearing along the columns. The entries within each row show the order in which the drugs ($A, B, \ldots, E$) are administered to each subject on a weekly basis. Such an arrangement is termed a *Latin square*, since the five treatments (drugs) appear exactly once in each row and exactly once in each column.

Figure 4 clearly demonstrates the existence of a 5×5 Latin square; more generally, Latin squares of size $n \times n$ exist for each value of $n \geq 1$. There are also occasions when it is desirable to superimpose certain pairs of $n \times n$ Latin squares. An example of this arises in testing the effects of n types of fertilizer and n types of insecticide on the yield of a particular crop. Suppose that a field on which the crop is grown is divided into an $n \times n$ grid of plots. In order to minimize vertical and horizontal variations in the composition and drainage properties of the soil, each fertilizer should appear on exactly one plot in each "row" and exactly one plot in each "column" of the grid. Likewise, each insecticide should appear once in each row and once in each column of the grid. In other words, a Latin square design should be used for each of the two treatments. Figure 5a shows a Latin square design for four fertilizer types (A, B, C, D), and Fig. 5b shows another Latin square design for four insecticide types (a, b, c, d).

In addition, the fertilizer and insecticide treatments can themselves interact, thus an ideal design would ensure that each of the n^2 possible combinations of the n fertilizers and n insecticides appear together once. Figure 5c shows that the two Latin squares in Figs. 5a and b (when superimposed) have this property; namely, each fertilizer–insecticide pair occurs exactly once on a plot. Such a pair of Latin squares is called *orthogonal*.

A pair of orthogonal $n \times n$ Latin squares need not exist for all values of $n \geq 2$. However, it has been proved that the only exceptions occur when $n = 2$ and $n = 6$. In all other cases, an orthogonal pair can be constructed.

Latin squares are special instances of *complete* designs, since every treatment appears in each row and in each

	W_1	W_2	W_3	W_4	W_5
S_1	A	B	C	D	E
S_2	B	C	D	E	A
S_3	C	D	E	A	B
S_4	D	E	A	B	C
S_5	E	A	B	C	D

FIGURE 4 Latin square design for drug treatments ($A, \ldots, E$) applied to subjects ($S_1, \ldots, S_5$) by week ($W_1, \ldots, W_5$).

A	B	C	D
C	D	A	B
D	C	B	A
B	A	D	C

(a)

a	b	c	d
d	c	b	a
b	a	d	c
c	d	a	b

(b)

Aa	Bb	Cc	Dd
Cd	Dc	Ab	Ba
Db	Ca	Bd	Ac
Bc	Ad	Da	Cb

(c)

FIGURE 5 Orthogonal Latin squares. (a) Latin square design for fertilizers ($A, \ldots, D$); (b) Latin square design for insecticides ($a, \ldots, d$); (c) superimposed Latin squares.

column. Another useful class of combinatorial designs is one in which not all treatments appear within each test group. Such *incomplete* designs are especially relevant when the number of treatments is large relative to the number of tests that can be performed on an experimental unit.

As an example of an incomplete design, consider an experiment in which subjects are to compare $v = 7$ brands of soft drink $(A, B, \ldots, G)$. For practical reasons, every subject is limited to receiving $k = 3$ types of soft drink. Moreover, to ensure fairness in the representation of the various beverages, each soft drink should be tasted by the same number $r = 3$ of subjects, and each pair of soft drinks should appear together the same number $\lambda = 1$ of times. It turns out that such a design can be constructed using $b = 7$ subjects, with the soft drinks compared by each subject i given by the set B_i below:

$$B_1 = \{A, B, D\}; \qquad B_2 = \{A, C, F\};$$
$$B_3 = \{A, E, G\}; \qquad B_4 = \{B, C, G\};$$
$$B_5 = \{B, E, F\}; \qquad B_6 = \{C, D, E\};$$
$$B_7 = \{D, F, G\}$$

The sets B_i are referred to as *blocks*, and such a design is termed a (b, v, r, k, λ) *balanced incomplete block design*. In any such design, the parameters b, v, r, k, λ must satisfy the following conditions:

$$bk = vr, \qquad \lambda(v - 1) = r(k - 1)$$

In the previous example, these relations hold since $7 \times 3 = 7 \times 3$ and $1 \times 6 = 3 \times 2$. While the above conditions must hold for any balanced incomplete block design, there need not exist a design corresponding to every set of parameters satisfying these conditions.

SEE ALSO THE FOLLOWING ARTICLES

COMPUTER ALGORITHMS • PROBABILITY

BIBLIOGRAPHY

Bogart, K. P. (2000). "Introductory Combinatorics," 3rd ed. Academic Press, San Diego, CA.

Cohen, D. I. A. (1978). "Basic Techniques of Combinatorial Theory," Wiley, New York.

Grimaldi, R. P. (1999). "Discrete and Combinatorial Mathematics," 4th ed. Addison-Wesley, Reading, MA.

Liu, C. L. (1985). "Elements of Discrete Mathematics," 2nd ed. McGraw–Hill, New York.

McEliece, R. J., Ash, R. B., and Ash, C. (1989). "Introduction to Discrete Mathematics," Random House, New York.

Roberts, F. S. (1984). "Applied Combinatorics," Prentice Hall, Englewood Cliffs, NJ.

Rosen, K. H. (1999). "Discrete Mathematics and Its Applications," 4th ed. McGraw–Hill, New York.

Rosen, K. H., ed. (2000). "Handbook of Discrete and Combinatorial Mathematics," CRC Press, Boca Raton, FL.

Tucker, A. (1995). "Applied Combinatorics," 3rd ed. John Wiley & Sons, New York.



Distributed Parameter Systems

N. U. Ahmed
University of Ottawa

I. System Models
II. Linear Evolution Equations
III. Nonlinear Evolution Equations and Differential Inclusions
IV. Recent Advances in Infinite-Dimensional Systems and Control

GLOSSARY

Banach space Normed vector space in which every Cauchy sequence has a limit; normed space complete with respect to the norm topology.

Cauchy sequence Sequence $\{x_n\} \in X$ is called a *Cauchy sequence* if $\lim_{n,m\to\infty} \|x_n - x_m\| = 0$; an element $x \in X$ is its limit if $\lim_{n\to\infty} \|x_n - x\| = 0$.

Normed vector space $X \equiv (X, \|\cdot\|)$ Linear vector space X furnished with a measure of distance between its elements $d(x, y) = \|x - y\|$, $\|x\| = d(0, x)$ satisfying: (a) $\|x\| \geq 0$, $\|x\| = 0$ iff $x = 0$, (b) $\|x + y\| \leq \|x\| + \|y\|$, (c) $\|\alpha x\| = |\alpha|\, \|x\|$ for all $x, y \in X$, and α scalar.

Reflexive Banach space Banach space X is reflexive if it is equivalent to (isomorphic to, indistinguishable from) its bidual $(X^*)^* \equiv X^{**}$.

Strictly convex normed space Normed space X is said to be strictly convex if, for every pair $x, y \in X$ with $x \neq y$, $\|\frac{1}{2}(x + y)\| < 1$ whenever $\|x\| \leq 1$, $\|y\| \leq 1$.

Uniformly convex normed space Normed space X is said to be uniformly convex if, for every $\varepsilon > 0$, there exists a $\delta > 0$, such that $\|x - y\| > \varepsilon$ implies that $\|\frac{1}{2}(x + y)\| < 1 - \delta$; for $\|x\| \leq 1$, $\|y\| \leq 1$; a uniformly convex Banach space is reflexive.

Weak convergence and weak topology Sequence $\{x_n\} \in X$ is said to be weakly convergent to x if, for every $x^* \in X^*$, $x^*(x_n) \to x^*(x)$ as $n \to \infty$; the notion of weak convergence induces a topology different from the norm topology and is called the weak topology; the spaces L_p, $1 \leq p < \infty$, are weakly (sequentially) complete.

Weak* convergence and *w-topology** Sequence $\{x_n^*\} \in X^*$ is said to be w^*-convergent to x^* if, for every $x \in X$, $x_n^*(x) \to x^*(x)$ as $n \to \infty$; the corresponding topology is called *w^*-topology*.

Weak and weak* compactness Set K in a Banach space X is said to be weakly (sequentially) compact if every sequence $\{x_n\} \in K$ has a subsequence $\{x_{n_k}\}$ and an $x_0 \in K$ such that, for each $x^* \in X^*$, $x^*(x_{n_k}) \to x^*(x_0)$; weak* compactness is defined similarly by interchanging the rolls of X^* and X.

DISTRIBUTED PARAMETER SYSTEMS are those whose state evolves with time distributed in space. More precisely, they are systems whose temporal evolution of state can be completely described only by elements of ∞-dimensional spaces. Such systems can be described by

partial differential equations, integro-partial differential equations, functional equations, or abstract differential or integro-differential equations on topological spaces.

I. SYSTEM MODELS

Mysteries of the universe have baffled scientists, poets, philosphers, and even the prophets. Scientists have tried to learn the secrets of nature by building and studying mathematical models for various systems. It is a difficult task to give a precise definition of what is meant by a *system*. But, roughly speaking, you may think of any physical or abstract entity that obeys certain systematic laws, deterministic or probabilistic, as being a system.

Systems can be classified into two major categories: (a) lumped parameter systems, and (b) distributed parameter systems. Lumped parameter systems are those whose temporal evolution can be described by a finite number of variables. That is, they are systems having finite degrees of freedom and they are governed by ordinary differential or difference equations in a finite-dimensional space. On the other hand, distributed parameter systems are those whose temporal evolution can be described only by elements of an ∞-dimensional space called the *state space*. These systems are governed by partial differential equations or functional differential equations or a combination of partial and ordinary differential equations.

A. Examples of Physical Systems

In this section we present a few typical examples of distributed parameter systems. The Laplace equation in $\mathbb{R}^n$ is given by:

$$\Delta\phi \equiv \sum_{i=1}^{n} \frac{\partial^2 \phi}{\partial x_i^2} = 0 \tag{1}$$

For $n = 3$, this equation is satisfied by the velocity potential of an irrotational incompressible flow, by the gravitational field outside the attracting masses, by electrostatic and magnetostatic fields, and also by the temperature of a body in thermal equilibrium. In addition to its importance on its own merit, the Laplacian is also used as a basic operator in diffusion and wave propagation problems. For example, the temperature of a body is governed by the so-called heat equation:

$$\frac{\partial T}{\partial t} = k\Delta T + f(t, x), \qquad (t, x) \in I \times \Omega, \qquad \Omega \subset \mathbb{R}^3 \tag{2}$$

where k is the thermal conductivity of the material and f is an internal source of heat. The classical wave equation in R^n is given by:

$$\frac{\partial^2 \psi}{\partial t^2} = c^2 \Delta\psi \tag{3}$$

where c denotes the speed of propagation. It may represent the displacement of a vibrating string, or propagation of acoustic or light waves and surface waves on shallow water.

Maxwell's equation, for a nonconducting medium with permeability μ and permittivity ε describing the temporal evolution of electric and magnetic field vectors H and E, respectively, in $\mathbb{R}_3$, is given by:

$$\begin{aligned} \mu\frac{\partial H}{\partial t} + \nabla \times E &= 0, \\ \varepsilon\frac{\partial E}{\partial t} - \nabla \times H &= 0 \\ \operatorname{div} E \equiv \nabla \cdot E &= 0, \\ \operatorname{div} H \equiv \nabla \cdot H &= 0 \end{aligned} \tag{4}$$

Under the assumptions of small displacement, the classical elastic waves in $\mathbb{R}^3$ are governed by the system of equations,

$$\begin{aligned} \rho\frac{\partial^2 y_i}{\partial t^2} &= \mu\,\Delta y_i + (\lambda + \mu)\frac{\partial}{\partial x_i}(\operatorname{div} y) \\ i &= 1, 2, 3, \\ y &= (y_1, y_2, y_3) \\ y_i &= y_i(t, x_1, x_2, x_3) \end{aligned} \tag{5}$$

where y represents the displacement of the elastic body from its unstrained configuration, ρ is the mass density, and λ, μ, are Lamé constants. In $\mathbb{R}^3$, the state of a single particle of mass m subject to a field of potential $v = v(t, x_1, x_2, x_3)$ is given by the Schrödinger equation:

$$i\hbar\frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m}\Delta\psi + v\psi \tag{6}$$

where $\hbar = 2\pi h$ is Planck's constant.

In recent years the nonlinear Schrödinger's equation has been used to take into account nonlinear interaction of particles in a beam by replacing $v\psi$ by a suitable nonlinear function $g(t, x, \psi)$.

The equation for an elastic beam allowing moderately large vibration is governed by a nonlinear equation of the form:

$$\begin{aligned} \rho A\frac{\partial^2 y}{\partial t^2} + \beta\frac{\partial y}{\partial t} + \frac{\partial^2}{\partial x^2}\left(EI\frac{\partial^2 y}{\partial x^2}\right) + N\frac{\partial^2 y}{\partial x^2} &= f(t, x) \\ N = \frac{EA}{2l}\int_0^l \left(\frac{\partial y}{\partial x}\right)^2 dx \\ x \in (0, l), \qquad t \geq 0 \end{aligned} \tag{7}$$

where E denotes Young's modulus, I the moment of area of the cross section A of the beam, l the length, ρ the mass density, N the membrane force, and f the applied force. For small displacements, neglecting N and β, one obtains the Euler equation for thin beams. The dynamics of Newtonian fluid are governed by the Navier–Stokes equations:

$$\rho\left(\frac{\partial v}{\partial t} + v \cdot \nabla v\right) - \nu\,\Delta v - (3\lambda + \nu)\,\text{grad div } v + \text{grad } p = f \tag{8}$$
$$\frac{\partial \rho}{\partial t} + \text{div}(\rho v) = 0, \qquad t \geq 0, \qquad x \in \Omega \subset \mathbb{R}^n$$

obtained from momentum and mass conservation laws, where ρ is the mass density, v the velocity vector, p the pressure, f the force density, and ν, λ are constant parameters.

Magnetohydrodynamic equations for a homogeneous adiabatic fluid in $\mathbb{R}^3$ is given by:

$$\frac{\partial v}{\partial t} + v \cdot \nabla v + \rho^{-1}\nabla p + (\mu\rho)^{-1}(B \times \text{ rot } B) = f$$
$$\frac{\partial \rho}{\partial t} + \text{ div}(\rho v) = 0, \qquad \frac{\partial s}{\partial t} + v \cdot \nabla s = 0 \tag{9}$$
$$\frac{\partial B}{\partial t} - \text{ rot}(v \times B) = 0, \qquad \text{div} B = 0$$
$$p = g(\rho, s), \qquad \text{rot } B \equiv \nabla \times B$$

where v, ρ, p, and f are as in Eq. (8); s is the entropy; B, the magnetic induction vector; and μ, the permeability.

In recent years semilinear parabolic equations of the form,

$$\frac{\partial \phi}{\partial t} = D\,\Delta\phi + f(t, x; \phi), \qquad t \geq 0, \qquad x \in \Omega \subset \mathbb{R}^n \tag{10}$$

have been extensively used for modeling biological, chemical, and ecological systems, where D represents the migration or diffusion coefficient.

The dynamics of a spacecraft with flexible appendages is governed by a coupled system of ordinary and partial differential equations.

In the following two sections we present abstract models that cover a wide variety of physical systems including those already mentioned.

B. Linear Systems

A general spatial differential operator used to construct system models is given by:

$$A(x, D)\phi \equiv \sum_{|\alpha| \leq 2m} a_\alpha(x)\, D^\alpha \phi$$
$$x \in \Omega \equiv \text{open subset of } \mathbb{R}^n \tag{11}$$

Under suitable smoothness conditions on the coefficient a_α and the boundary $\partial\Omega$ of Ω one can express Eq. (11) in the so-called divergence form:

$$A(x, D)\phi \equiv \sum_{|\alpha|, |\beta| \leq m} (-1)^{|\alpha|} D^\alpha \left(a_{\alpha\beta} D^\beta \phi\right) \tag{12}$$

The operator A is said to be strongly uniformly elliptic on Ω if there exists a $\gamma > 0$ such that

$$(-1)^m \text{ Re}\left\{\sum_{|\alpha|=2m} a_\alpha(x)\xi^\alpha\right\} \geq \gamma\,|\xi|^{2m} \quad \text{for all} \quad x \in \Omega \tag{13}$$

or

$$\text{Re}\left\{\sum_{|\alpha|=|\beta|=m} a_{\alpha\beta}(x)\xi^\alpha\xi^\beta\right\} \geq \gamma\,|\xi|^{2m} \tag{14}$$

Many physical processes in steady state (for example, thermal equilibrium or elastic equilibrium) can be described by elliptic boundary value problems given by:

$$A(x, D)\phi = f \quad \text{on} \quad \Omega$$
$$B(x, D)\phi = g \quad \text{on} \quad \partial\Omega \tag{15}$$

where $B = \{B_j,\ 0 \leq j \leq m-1\}$ is a system of suitable boundary operators. For example, the boundary operator may be given by the Dirichlet operator,

$$B \equiv \left\{\frac{\partial^j}{\partial \nu^j}, 0 \leq j \leq m-1\right\}$$

where $\partial/\partial\nu$ denotes spatial derivatives along the outward normal to the boundary $\partial\Omega$. The boundary operators $\{B_j\}$ cannot be chosen arbitrarily; they must satisfy certain compatibility conditions with respect to the operator A. Only then is the boundary value problem (15) well posed; that is, one can prove the existence of a solution and its continuous dependence on the data f and $g = \{g_j,\ 0 \leq j \leq m-1\}$. The order of each of the operators B_j is denoted by m_j.

Evolution equations of parabolic type arise in the problems of heat transfer, chemical diffusions, and also in the study of Markov processes. The most general model describing such phenomenon is given by:

$$\frac{\partial \phi}{\partial t} + A\phi = f, \qquad t, x \in I \times \Omega = Q$$
$$B\phi = g, \qquad t, x \in I \times \partial\Omega \tag{16}$$
$$\phi(0) = \phi_0, \qquad x \in \Omega$$

where f, g, and ϕ_0 are the given data.

Second-order evolution equations of hyperbolic type describing many vibration and wave propagation problems have the general form:

$$\begin{aligned} \frac{\partial^2\phi}{\partial t^2} + A\phi &= f, \qquad t, x \in I \times \Omega \\ B\phi &= g, \qquad t, x \in I \times \Omega \\ \phi(0) &= \phi_0, \qquad x \in \Omega \\ \dot{\phi}(0) &= \phi_1, \end{aligned} \tag{17}$$

Schrödinger-type evolution equations are obtained if A is replaced by iA in Eq. (16).

In the study of existence of solutions of these problems, Garding's inequality is used for *a priori* estimates. If A is strongly uniformly elliptic, the principal coefficients satisfy Hölder conditions on Ω uniformly with respect to $t \in I$, and the other coefficients are bounded measurable, then one can prove the existence of a $\lambda \in \mathbb{R}$ and $\alpha > 0$ such that

$$a(t, \phi, \phi) + \lambda|\phi|_H^2 > \alpha\|\phi\|_V^2, \qquad t \in I \tag{18}$$

where

$$a(t, \phi, \psi) \equiv \sum_{|\alpha|,|\beta| \le m} \left\langle a_{\alpha,\beta}(t, \cdot) D^\beta\phi, D^\alpha\psi \right\rangle_\Omega \tag{19}$$

and V is any reflexive Banach space continuously and densely embedded in $H = L_2(\Omega)$ where $L_p(\Omega)$ is the equivalence classes of pth-power Lebesgue integrable functions on $\Omega \subset R^n$, with the norm given by:

$$\|f\|_{L_{p(\Omega)}} \equiv \begin{cases} \left(\int \Omega |f(x)|^p \, dx\right)^{1/p} & \text{for} \quad 1 \le p < \infty \\ \text{ess sup}\{f(x), x \in \Omega\} & \text{for} \quad p = \infty \end{cases}$$

For example, V could be $W_0^{m,p}$, $W^{m,p}$, or $W_0^{m,p} \subset V \subset W^{m,p}$ for $p \ge 2$ where $W_0^{m,p}(\Omega)$ is the closure of C^∞ functions with compact support on Ω in the topology of $W^{m,p}$, and

$$W^{m,p}(\Omega) \equiv \left\{ f \in L_p(\Omega) : D^\alpha f \in L_p(\Omega), \quad |\alpha| \le m \right\}$$

furnished with the norm topology:

$$\|f\|_{W^{m,p}} \equiv \sum_{|\alpha| \le m} \|D^\alpha f\|_{L_p(\Omega)}, \qquad p \ge 1$$

$$D^\alpha \equiv D_1^{\alpha_1} D_1^{\alpha_2} \cdots D_n^{\alpha_n}, \qquad D_i^{\alpha_i} \equiv \frac{\partial^{\alpha_i}}{\partial x_i^{\alpha_i}}$$

$$\|\alpha\| = \sum_{i=1}^n \alpha_i, \qquad \alpha_i \equiv \text{nonnegative integers}$$

C. Nonlinear Evolution Equations and Differential Inclusions

Nonlinear systems are more frequently encountered in practical problems than the linear ones. There is no clear-cut classification for these systems. We present here a few basic structures that seem to cover a broad area in the field.

1. Elliptic Systems

The class of elliptic problems which have received considerable attention in the literature is given by:

$$\begin{aligned} A\psi &= 0 \qquad \text{on} \quad \Omega \\ D^\alpha\psi &= 0 \qquad \text{on} \quad \partial\Omega, \qquad |\alpha| \le m - 1 \end{aligned} \tag{20}$$

where

$$A\psi \equiv \sum_{|\alpha| \le m} (-1)^{|\alpha|} D^\alpha(|D^\alpha\psi|^{p-2} D^\alpha\psi) + \sum_{|\beta| \le m-1} (-1)^{|\beta|} \times D^\beta\left(b_\beta(x, \psi, D^1\psi, \ldots, D^m\psi)\right) \tag{21}$$

2. Semilinear Systems

The class of nonlinear evolution equations that can be described in terms of a linear operator and a nonlinear operator containing lower order derivatives has been classified as semilinear. These systems have the form:

$$\begin{aligned} \frac{d\psi}{dt} + A(t)\psi + f(t, \psi) &= 0 \\ \psi(0) &= \psi_0 \end{aligned} \tag{22}$$

and

$$\begin{aligned} \frac{d^2\psi}{dt^2} + A(t)\psi + f(t, \psi) &= 0 \\ \psi(0) &= \psi_0, \\ \dot{\psi}(0) &= \psi_1 \end{aligned} \tag{23}$$

where $A(t)$ may be a differential operator of the form (11) and the nonlinear operator f may be given by:

$$f(t, \psi) \equiv \sum_{|\alpha| \le 2m-1} (-1)^{|\alpha|} D^\alpha \times \left(b_\alpha(t, \cdot; \psi, D^1\psi, \ldots, D^{2m-1}\psi)\right) \tag{24}$$

For example, a second-order semilinear parabolic equation is given by:

$$\frac{\partial\phi}{\partial t} + \sum_{i,J} a_{ij}(t, x)\phi_{x_i x_j} + a(t, x, \phi, \phi_x) = 0 \tag{25}$$

which certainly covers the ecological model, Eq. (10). In short, Eq. (22) is an abstract model for a wide variety of

nonlinear diffusions. It covers the Navier–Stokes equation (8) and many others including the first-order semilinear hyperbolic system:

$$\frac{\partial y}{\partial t} + \left(\sum_{i=1}^{n} A_i(t,x) \frac{\partial y}{\partial x_i} + B(t,x)y \right) + f(t,x,y) = 0 \tag{26}$$

The second-order abstract semilinear equation (23) covers a wide variety of nonlinear vibration and wave propagation problems.

3. Quasilinear Systems

The most general form of systems governed by quasilinear evolution equations that have been treated in the literature is given by:

$$\frac{d\psi}{dt} + A(t,\psi)\psi + f(t,\psi) = 0 \qquad \psi(0) = \psi_0 \tag{27}$$

It covers the quasilinear second-order parabolic equation or systems of the form:

$$\frac{\partial \psi}{\partial t} + \sum_{i,J}^{n} a_{ij}(t,x;\psi,\psi_x)\psi_{x_i x_j} + a(t,x,\psi,\psi_x) = 0 \tag{28}$$

where, for parabolicity, one requires that

$$a_{ij}(t,x,p,q)\xi_i\xi_j \geq \gamma|\xi|^2$$
$$\gamma > 0, \qquad t,x \in Q, \qquad p \in \mathbb{R}, \qquad q \in \mathbb{R}^n$$

It covers the quasilinear hyperbolic systems of the form:

$$A_0(t,x,y)\frac{\partial y}{\partial t} + \sum_{i}^{n} A_i(t,x,y)\frac{\partial y}{\partial x_i} + B(t,x,y) = 0 \tag{29}$$

including the magnetohydrodynamic equation (9).

4. Differential Inclusions

In recent years abstract differential inclusions have been used as models for controlled systems with discontinuities. For example, consider the abstract semilinear equation (22). In case the operator f, as a function of the state ψ, is discontinuous but otherwise well defined as a set-valued function in a suitable topological space, one may consider Eq. (22) as a differential inclusion:

$$-\dot{\psi}(t) \in A(t)\psi(t) + f(t,\psi(t)) \quad \text{a.e.} \qquad \psi(0) = \psi_0 \tag{30}$$

Such equations also arise from variational inequalities.

D. Stochastic Evolution Equations

In certain situations because of lack of precise knowledge of the system parameters arising from inexact observation or due to gaps in our fundamental understanding of the physical world, one may consider stochastic models to obtain most probable answers to many scientific questions. Such models may be described by stochastic evolution equations of the form:

$$d\psi = (A(t)\psi + f(t,\psi))\,dt + \sigma(t,\psi)\,dW \qquad \psi(0) = \psi_0 \tag{31}$$

where A is a differential operator possibly of the form (11), f is a nonlinear operator of the form (24), and σ is a suitable operator-valued function. The variable $W = \{W(t), t \geq 0\}$ represents a Wiener process with values in a suitable topological space and defined on a probability space. The uncertainty may arise from randomness of ψ_0, the process W, and even from the operators A, f, and σ. This is further discussed in Sections II and III.

II. LINEAR EVOLUTION EQUATIONS

In this section, we present some basic results from control theory for systems governed by linear evolution equations.

A. Existence of Solutions for Linear Evolution Equations and Semigroups

A substantial part of control theory for distributed parameter systems has been developed for first- and second-order evolution equations of parabolic and hyperbolic type of the form:

$$\frac{d\phi}{dt} + A(t)\phi = f, \qquad \phi(0) = \phi_0 \tag{32}$$

$$\frac{d^2\phi}{dt^2} + A(t)\phi = f, \qquad \phi(0) = \phi_0, \quad \dot{\phi}(0) = \phi_1 \tag{33}$$

A fundamental question that must be settled before a control problem can be considered is the question of existence of solutions of such equations.

We present existence theorems for problems (32) and (33) and conclude the section with a general result when the operator A is only a generator of a c_0-semigroup (in case of constant A) or merely the generator of an evolution operator in a Banach space (in case A is variable). The concepts of semigroups and evolution operators are introduced at a convenient place while dealing with the questions of existence.

For simplicity we consider time-invariant systems, although the result given below holds for the general case.

Theorem 1. Consider system (33) with the operator A time invariant and self-adjoint and suppose it satisfies the conditions:

(a1) $|\langle A\phi, \psi\rangle| \le c\|\phi\|_V\|\psi\|_V$

$c \ge 0, \qquad \phi, \psi \in V$

(a2) $\langle A\phi, \phi\rangle + \lambda|\phi|_H^2 \ge \alpha\|\phi\|_V^2$

$\lambda \in R, \qquad \alpha > 0, \qquad \phi \in V$

Then, for every $\phi_0 \in V$, $\phi_1 \in H$, and $f \in L_2(I, H)$, system (33) has a unique solution ϕ satisfying:

(c1) $\phi \in L_\infty(I, V) \cap C(\bar{I}, V)$

(c2) $\dot{\phi} \in L_\infty(I, V) \cap C(\bar{I}, H)$

and

(c3) $(\phi_0, \phi_1, f) \to (\phi, \dot{\phi})$

is a continuous mapping from $V \times H \times L_2(I, H)$ to $C(\bar{I}, V) \times C(\bar{I}, H)$.

Proof (Outline). The proof is based on Galerkin's approach which converts the infinite-dimensional system (33) into its finite-dimensional approximation, then by use of *a priori* estimates and compactness arguments one shows that the approximating sequence has a subsequence that converges to the solution of problem (33). □

Note that the second-order evolution equation (33) can be written as a first-order equation $d\psi/dt + \tilde{A}\psi = \tilde{f}$ where

$$\psi = \begin{bmatrix} \phi \\ \dot{\phi} \end{bmatrix}, \qquad \tilde{A} = \begin{bmatrix} 0 & -I \\ A & 0 \end{bmatrix}$$
$$\tilde{f} = \begin{bmatrix} 0 \\ f \end{bmatrix} \tag{34}$$

Defining $X = V \times H$, with the product topology, as the state space, it follows from Theorem 1(c3) that there exists an operator-valued function $S(t)$, $t \ge 0$, with values $S(t) \in \mathcal{L}(X)$ so that

$$\psi(t) = S(t)\psi_0 + \int_0^t S(t-\theta)\tilde{f}(\theta)\, d\theta \tag{35}$$

The family of operators $\{S(t), t \ge 0\}$ forms a c_0-semigroup in X and $\psi \in C(\bar{I}, X)$ where $C(I, X)$ is the space of continuous functions on I with values in the Banach space X with the norm (topology):

$$\|f\| = \sup\{|f(t)|_X, t \in I\}$$

For system (32) one can prove the following result using the same procedure.

Theorem 2. Consider system (32) and suppose that the operator A satisfies assumptions (a1) and (a2) (see Theorem 1). Then, for each $\phi_0 \in H$ and $f \in L_2(I, V^*)$, system (32) has a unique solution:

$$\phi \in L_2(I, V) \cap L_\infty(I, H) \cap C(\bar{I}, H)$$

and further

$$(\phi_0, f) \to \phi$$

is a continuous map from $H \times L_2(I, V^*)$ to $C(\bar{I}, H)$.

As a consequence of this result there exists an evolution operator $U(t, s)$, $0 \le s \le t \le \infty$, with values $U(t, s) \in \mathcal{L}(H)$ such that

$$\phi(t) = U(t, 0)\phi_0 + \int_0^t U(t, \theta) f(\theta)\, d\theta \tag{36}$$

is a (mild) weak solution of problem (32).

We conclude this section with some results on the question of existence of solutions for a class of general time-invariant linear systems on Banach space.

Let X be a Banach space, and $S(t)$, $t \ge 0$, a family of bounded linear operators from X to X satisfying the properties:

(a) $S(0) = I$ (identity operator)

(b) $S(t + \tau) = S(t)S(\tau), \qquad t, \tau \ge 0$ (37)

(c) Strong $\lim_{t \downarrow 0^+} S(t)\xi = \xi \in X$

The operator $S(t)$, $t \ge 0$, satisfying the above properties is called a *strongly continuous semigroup* or, in short, a c_0-semigroup. Let A be a closed, densely defined linear operator with domain $D(A) \subset X$ and range $R(A) \subset X$. Suppose there exist numbers $M > 0$ and $\omega \in R$ such that

$$\|(\lambda I - A)^{-1}\|_X \le M/(\lambda - \omega)$$

for all real $\lambda > \omega$. Then, by a fundamental theorem from the semigroup theory known as the Hille–Yosida theorem, there exists a unique c_0-semigroup $S(t)$, $t \ge 0$, with A as its infinitesimal generator. The semigroup $S(t)$, $t \ge 0$, satisfies the properties:

(a) $\|S(t)\|_{\mathcal{L}(X)} \le Me^{\omega t}, \qquad t \ge 0$

(b) for $\xi \in D(A)$

$S(t)\xi \in D(A)$ for all $t \ge 0$ (38)

(c) for $\xi \in D(A)$

$$\frac{d}{dt}S(t)\xi = AS(t)\xi = S(t)A\xi, \qquad t \ge 0$$

If $\omega = 0$, we have a *bounded semigroup*; for $\omega = 0$ and $M = 1$ we have a *contraction semigroup*, and for $\omega < 0$, we have a *dissipative semigroup*. In general, the abstract Cauchy problem:

$$\frac{dy}{dt} = Ay, \qquad y(0) = \xi \tag{39}$$

has a unique solution $y(t) = S(t)\xi$, $t > 0$, with $y(t) \in D(A)$, provided $\xi \in D(A)$. If $\xi \in D(A)$ and f is any strongly continuously differentiable function with values $f(t) \in X$, then the inhomogeneous problem,

$$\frac{dy}{dt} = Ay + f, \qquad y(0) = \xi \tag{40}$$

has a unique continuously differentiable solution y given by:

$$y(t) = S(t)\xi + \int_0^t S(t-\theta) f(\theta)\, d\theta$$
$$\text{with} \quad y(t) \in D(A) \qquad \text{for all} \quad t \geq 0 \tag{41}$$

A solution satisfying these conditions is called a *classical solution*. For control problems these conditions are rather too strong since, in general, we do not expect the controls, for example $f(t)$, to be even continuous. Thus, there is a need for a broader definition and this is provided by the so-called mild solution.

Any function $y\colon I \to X$ having the integral representation (41) is called a mild solution of problem (40). In this regard we have the following general result.

Theorem 3. Suppose A is the generator of a c_0-semigroup $S(t), t \geq 0$, in X and let $y_0 \in X$ and $f \in L_p(I, X)$, $1 \leq p \leq \infty$, where $L_p(I, X)$ is the space of strongly measurable functions on I taking values in a Banach space X with norm:

$$\|f\|_{L_p(I,X)} = \begin{cases} \left(\int_I |f(t)|_X^p \, dt \right)^{1/p}, & 1 \leq p < \infty \\ \text{ess sup } \{|f(t)|_X, t \in I\}, & p = \infty \end{cases}$$

Then, evolution Eq. (40) has a unique mild solution y given by:

$$y(t) = S(t)y_0 + \int_0^t S(t-\theta) f(\theta)\, d\theta \tag{42}$$

In this case $y(t)$ does not necessarily belong to $D(A)$. Another special but important class of strongly continuous semigroups $S(t)$, $t \geq 0$, that satisfies the property:

$$S(t)X \subset D(A) \qquad \text{for all} \quad t > 0$$

is called the *analytic* (holomorphic, parabolic) *semigroup*. An analytic semigroup has the following properties:

(a) $S(t)X \subset D(A^n)$ for all integers $n \geq 0$ and $t > 0$,

(b) $S(t), d^n S(t)/dt^n = A^n S(t)$ are all bounded operators for $t > 0$.

One reason for calling the analytic semigroups parabolic semigroups is that $S(t)$, $t \geq 0$, turns out to be the fundamental solution of certain parabolic evolution equations. Consider the differential operator,

$$L(x, D) = \sum_{|\alpha| \leq 2m} a_\alpha(x) D^\alpha$$

and suppose it is strongly elliptic of order $2m$ and

$$B_j(x, D) = \sum_{|\alpha| \leq m_j} b_{j,\alpha} D^\alpha, \qquad 0 \leq j \leq m-1$$

is a set of normal boundary operators as defined earlier. Define

$$D(A) = W^{2m,p}(\Omega) \cap W_0^{m,p}(\Omega)$$
$$(A\phi)(x) = -L(x, D)\phi(x), \qquad \phi \in D(A) \tag{43}$$

Then, under certain technical assumptions, A generates an analytic semigroup $S(t)$, $t \geq 0$, in the Banach space $X = L_p(\Omega)$. The initial boundary value problem,

$$\frac{\partial \phi}{\partial t}(t, x) + L(x, D)\phi(t, x) = f(t, x)$$
$$t \in (0, T), \qquad x \in \Omega$$
$$(D^\alpha \phi)(t, x) = 0 \tag{44}$$
$$|\alpha| \leq m-1, \qquad t \in (0, T), \qquad x \in \partial\Omega$$
$$\phi(0, x) = \phi_0(x), \qquad x \in \Omega$$

can be considered to be an abstract evolution equation in $X = L_p(\Omega)$ and be written as:

$$\frac{d\phi}{dt} = A\phi + f$$
$$\phi(0) = \phi_0$$

with mild solution given by:

$$\phi(t) = S(t)\phi_0 + \int_0^t S(t-\theta) f(\theta)\, d\theta \tag{45}$$

where $f \in L_p(I, X)$.

B. Stability

The question of stability is of signnificant interest to system scientists, since every physical system must be stable to function properly.

In this section, we present a Lyapunov-like result for the abstract evolution equation,

$$\frac{dy}{dt} = Ay \tag{46}$$

where A is the generator of a strongly continuous semigroup $S(t)$, $t \geq 0$ on H. Let $\mathscr{L}^+(H)$ denote the class of symmetric, positive, self-adjoint operators in H; that is,

$T \in \mathcal{L}(H)$, $T = T^*$, and $(T\xi, \xi) > 0$ for $\xi\ (\neq 0) \in H$. We can prove the following result.

Theorem 4. A necessary and sufficient condition for the system $dy/dt = Ay$ to be exponentially stable in the Lyapunov sense (i.e., there exist $M \geq 1$, $\beta > 0$ such that $\|y(t)\| \leq Me^{-\beta t}$) is that the operator equation,

$$A^*Y + YA = -\Gamma \tag{47}$$

has a solution $Y \in \mathcal{L}^+(H)$ for each $\Gamma \in \mathcal{L}^+(H)$ with $(\Gamma\xi, \xi) \geq \gamma|\xi|_H^2$ for some $\gamma > 0$.

REMARK. Equation (47) is understood in the sense that the equality,

$$0 = (\Gamma\xi, \eta) + (A\xi, Y\eta) + (Y\xi, A\eta)$$

holds for all $\xi,\ \eta \in D(A)$.

Corollary 5. If the autonomous system (46) is asymptotically stable, then the system,

$$\frac{dy}{dt} = Ay + f, \qquad t \geq 0 \tag{48}$$

is stable in the L_p sense; that is, for every $y_0 \in H$ and input $f \in L_p(0, \infty; H)$, the output $y \in L_p(0, \infty; H)$ for all $1 \leq p \leq \infty$ and in particular, for $1 \leq p < \infty$, $y(t) \to 0$ as $t \to \infty$.

We conclude this section with a remark on the solvability of Eq. (47). Let $\{\lambda_i\}$ be the eigenvalues of the operator A, each repeated as many times as its multiplicity requires, and let $\{\xi_i\}$ denote the corresponding eigenvectors complete in H. Consider Eq. (47) and form

$$\left(A^*Y\xi_i, \xi_j\right) + (YA\,\xi_i, \xi_j) = -(\Gamma\xi_i, \xi_j) \tag{49}$$

for all integers $i,\ j \geq 1$. Clearly, from this equation, it follows that

$$(Y\xi_i, \xi_j) = -(\Gamma\xi_i, \xi_j)/(\lambda_i + \bar{\lambda}_j) \tag{50}$$

Hence, if $\lambda_i + \bar{\lambda}_j \neq 0$ for all i, j, Γ determines Y uniquely, and if $\lambda_i + \bar{\lambda}_j = \text{Re}\ \lambda_i < 0$ for all i, then $Y \in \mathcal{L}^+(H)$. In other words, if system (46) is asymptotically stable then the operator equation (47) always has a positive solution.

C. System Identification

A system analyst may know the structure of the system—for example, the order of the differential operator and its type (parabolic, hyperbolic, etc.)—but the parameters are not all known. In that case, the analyst must identify the unknown parameters from available information. Consider the natural system to be given by $\dot{y}^* = A(q^*)y^*$. Assume that q^* is unknown to the observer but the observer can observe certain data z^* from a Hilbert space K, the output space, which corresponds to the natural history y^*. The observer chooses a feasible set Q, where q^* may possibly lie, and constructs the model system,

$$\begin{aligned} \dot{y}(q) &= A(q)y(q), \qquad y(q)(0) = y_0 \\ z(q) &= Cy(q) \end{aligned} \tag{51}$$

where C is the observation or output operator, an element of $\mathcal{L}(H, K)$. The analyst may choose to identify the parameter approximately by minimizing the functionl,

$$J(q) = \frac{1}{2}\int_0^T |z(q) - z^*|_K^2\, dt \qquad \text{over } Q \tag{52}$$

Similarly, one may consider identification of an operator appearing in system equations. For example, one may consider the system,

$$\begin{aligned} &\frac{d^2y}{dt^2} + Ay + B^*y = f \\ &y(0) = y_0, \qquad \dot{y}(0) = y_1 \\ &z = Cy \end{aligned} \tag{53}$$

where the operator A is known but the operator B^* is unknown. One seeks an element B from a feasible set $P^0 \subset \mathcal{L}(V, V^*)$ so that

$$J(B) = \int_0^T g(t, y(B), \dot{y}(B))\, dt \tag{54}$$

is minimum, where g is a suitable measure of discrepancy between the model output, $z(B) = Cy(B)$, and the observed data z^* corresponds to the natural history y^*.

In general, one may consider the problem of identification of all the operators A, B including the data y_0, y_1, and f. For simplicity, we shall consider only the first two problems and present a couple sample results. First, consider problem (51) and (52).

Theorem 6. Let the feasible set of parameters Q be a compact subset of a metric space and suppose for each $q \in Q$ that $A(q)$ is the generator of a strongly continuous contraction semigroup in H. Let

$$(\gamma I - A(q_n))^{-1} \to (\gamma I - A(q_0))^{-1} \tag{55}$$

in the strong operator topology for each $\gamma > 0$ whenever $q_n \to q_0$ in Q. Then there exists $q^0 \in Q$ at which $J(q)$ attains its minimum.

Proof. The proof follows from the fact that under assumption (55) the semigroup $S_n(t)$, $t \geq 0$, corresponding to q_n strongly converges on compact intervals to the semigroup $S_0(t)$, $t \geq 0$, corresponding to q_0. Therefore, J is continuous on Q and, Q being compact, it attains its minimum on Q. □

The significance of the above result is that the identification problem is well posed.

For the second-order evolution equation (53) we have the following result.

Theorem 7. Consider system (53). Let P^0 be a compact (in the sense of strong operator topology) subset of the ball,

$$P_b \equiv \left\{B \in \mathcal{L}(V, V^*): \|B\|_{\mathcal{L}(V,V^*)} \leq b\right\}$$

Then, for each g defined on $I \times V \times H$ which is measurable in the first variable and lower semi continuous in the rest, the functional $J(B)$ of Eq. (54) attains its minimum on P^0.

The best operator B^0 minimizing the functional $J(B)$ can be determined by use of the following necessary conditions of optimality.

Theorem 8. Consider system (53) along with the functional,

$$J(B) \equiv \frac{1}{2}\int_0^T |Cy(B) - z^*(t)|_K^2 \, dt$$

with the observed data $z^* \in L_2(I, K)$, the observer $C \in \mathcal{L}(H, K)$, $f \in L_2(I, H)$, $y_0 \in V$, $y_1 \in H$ and P^0 as in Theorem 7. Then, for B^0 to be optimal, it is necessary that there exists a pair $\{y, x\}$ satisfying the equations,

$$\begin{aligned} &\ddot{y} + Ay + B^0 y = f \\ &\ddot{x} + A^*x + (B^0)^*x = C^*\Lambda_K(Cy(B^0) - z^*) \\ &y(0) = y_0, \qquad x(T) = 0 \\ &\dot{y}(0) = y_1, \qquad \dot{x}(T) = 0 \end{aligned} \tag{56}$$

and the inequality

$$\int_0^T \langle B^0 y(B^0), x\rangle_{V^*,V} \, dt \geq \int_0^T \langle By(B^0), x\rangle_{V^*,V} \, dt \tag{57}$$

for all $B \in P^0$, where Λ_K is the canonical isomorphism of K onto K^* such that $\|\Lambda_K e\|_{K^*} = \|e\|_K$.

By solving Eqs. (56) and (57) simultaneously one can determine B^0. In fact, a gradient-type algorithm can be developed on the basis of Eqs. (56) and (57).

D. Controllability

Consider the controlled system,

$$\dot{\phi} = A\phi + Bu, \qquad t \geq 0 \tag{58}$$

with ϕ denoting the state and u the control. The operator A is the generator of a strongly continuous semigroup $S(t)$, $t \geq 0$, in a Banach space X and B is the control operator with values in $\mathcal{L}(E, X)$ where E is another Banach space. Let $\mathcal{U}$ denote the class of admissible controls, possibly a proper subset of $L_p^{\text{loc}}(E) \equiv$ locally pth power summable E-valued functions on $\mathbb{R}_0 = [0, \infty)$. For a given initial state $\phi_0 \in \mathcal{U}$,

$$\phi(t) = S(t)\phi_0 + \int_0^t S(t-\theta)Bu(\theta)\, d\theta, \qquad t \geq 0$$

denotes the mild (weak) solution of problem (58).

Given $\phi_0 \in X$ and a desired target $\phi_1 \in X$, is it possible to find a control from $\mathcal{U}$ that transfers the system from state ϕ_0 to the desired state ϕ_1 in finite time? This is the basic question of controllability. In other words, for a given $\phi_0 \in X$, one defines the attainable set:

$$\mathcal{A}(t) \equiv \left\{x \in X: x = S(t)\phi_0 + \int_0^t S(t-\theta)Bu(\theta)\, d\theta, \; u \in \mathcal{U}\right\}$$

and inquires if there exists a finite time $\tau \geq 0$, such that $\phi_1 \in \mathcal{A}(\tau)$ or equivalently,

$$\phi_1 - S(\tau)\phi_0 \in R(\tau) \equiv \mathcal{A}(\tau) - S(\tau)\phi_0$$

The set $\mathcal{R}(\tau)$, given by:

$$\mathcal{R}(\tau) \equiv \left\{\xi \in X: \xi = L_\tau u \equiv \int_0^\tau S(\tau-\theta)Bu(\theta)\, d\theta, \; u \in \mathcal{U}\right\}$$

is called the *reachable set*. If $S(t)B$ is a compact operator for each $t \geq 0$, then $\mathcal{R}(\tau)$ is compact, hence the given target may not be attainable. A similar situation arises if $BE \subset D(A)$ and $\phi_0 \in D(A)$ and $\phi_1 \notin D(A)$. As a result, an appropriate definition of controllability for ∞-dimensional systems may be formulated as follows.

Definition. System (58) is said to be controllable (exactly controllable) in the time interval $[0, \tau]$ if $R(\tau)$ is dense in $X[\mathcal{R}(\tau) = X]$ and it is said to be controllable (exactly controllable) in finite time if $\cup_{\tau>0} R(\tau)$ is dense in $X[\cup_{\tau>0} R(\tau) = X]$. Note that for finite-dimensional systems, $X = R^n$, $E = R^m$, and controllability and exact controllability are all and the same. It is only for ∞-dimensional systems that these concepts are different.

We present here a classical result assuming that both X and E are self-adjoint Hilbert spaces with $\mathcal{U} = L_2^{\text{loc}}(E)$.

Theorem 9. For system (58) the following statements are equivalent:

(a) System (58) is controllable in time τ

(b) $\left(L_\tau L_\tau^*\right) \in \mathcal{L}^+(X)$

(c) $\text{Ker } L_\tau^* \equiv \left\{\xi \in X: L_\tau^*\xi = 0\right\} = \{0\}$

where L_τ^* is the adjoint of the operator L_τ and $L_\tau u \equiv \int_0^\tau S(\tau-\theta)\,Bu(\theta)\,d\theta$.

Note that by our definition, here controllability means approximate controllability; that is, one can reach an arbitrary neighborhood of the target but never exactly at the target itself. Another interesting difference between finite- and infinite-dimensional systems is that in case $X = R^n$, $E = R^m$, condition (b) implies that $(L_\tau L_\tau^*)^{-1}$ exists and the control achieving the desired transfer from ϕ_0 to ϕ_1 is given by:

$$u = \left(L_\tau L_\tau^*\right)^{-1}(\phi_1 - S(\tau)\phi_0)$$

For ∞-dimensional systems, the operator $(L_\tau L_\tau^*)$ does not in general have a bounded inverse even though the operator is positive.

Another distinguishing feature is that in the finite-dimensional case the system is controllable if and only if the rank condition,

$$\operatorname{rank}(B, AB, \ldots, A^{n-1}B) = n$$

holds. In the ∞-dimensional case there is no such condition; however, if $BE \subset \bigcap_{n=1}^\infty D(A^n)$ then the system is controllable if

$$\text{closure}\left\{\bigcup_{n=0}^{\infty} \text{range}(A^n B)\right\} = X \tag{59}$$

This condition is also necessary and sufficient if $S(t)$, $t \geq 0$, is an analytic semigroup and $BE \subset \bigcup_{t>0} S(t)X$.

In recent years, much more general results that admit very general time-varying operators $\{A(t),\ B(t),\ t \geq 0\}$, including hard constraints on controls, have been proved. We conclude this section with one such result. The system,

$$\begin{aligned} \dot{y} &= A(t)y + B(t)u, \qquad t \geq 0 \\ y(0) &= y_0 \end{aligned} \tag{60}$$

is said to be globally null controllable if it can be steered to the origin from any initial state $y_0 \in X$.

Theorem 10. Let X be a reflexive Banach space and Y a Banach space densely embedded in X with the injection $Y \subset X$ continuous. For each $t \geq 0$, $A(t)$ is the generator of a c_0-semigroup satisfying the stability condition and $A \in L_1^{\text{loc}}(0, \infty; \mathcal{L}(Y, X))$ where $\mathcal{L}(X, Z)$ is the space of bounded linear operators from a Banach space X to a Banach space Z; $\mathcal{L}(X) \equiv \mathcal{L}(X, X)$, $B \in L_q^{\text{loc}}(0, \infty; \mathcal{L}(E, X))$, and $\mathcal{U} = \{u \in L_p^{\text{loc}}(E);\ u(t) \in \Gamma \text{ a.e.}\}$ where Γ is a closed bounded convex subset of E with $o \in \Gamma$ and $p^{-1} + q^{-1} = 1$. Then a necessary and sufficient condition for global null controllability of system (60) is that

$$\int_0^\infty H_\Gamma(B^*(t)\psi(t))\,dt = +\infty \tag{61}$$

for all nontrivial weak solutions of the adjoint system

$$\dot{\psi} + A^*(t)\psi = 0, \qquad t \geq 0$$

where $H_\Gamma(\xi) = \sup\{(\xi, e)_{E^*,E},\ e \in \Gamma\}$.

E. Existence of Optimal Controls

The question of existence of optimal controls is considered to be a fundamental problem in control theory. In this section, we present a simple existence result for the hyperbolic system,

$$\begin{aligned} \ddot{\phi} + A\phi &= f + Bu, \qquad t \in I \equiv (0, T) \\ \phi(0) &= \phi_0, \qquad \dot{\phi}(0) = \phi_1 \end{aligned} \tag{62}$$

Similar results hold for parabolic systems. Suppose the operator A and the data ϕ_0, ϕ_1, f satisfy the assumptions of Theorem 1. Let E be a real Hilbert space and $\mathcal{U}_0 \subset L_2(I, E)$ the class of admissible controls and $B \in \mathcal{L}(E, H)$. Let $S_0 \subset V$ and $S_1 \subset H$ denote the set of admissible initial states. By Theorem 1, for each choice of $\phi_0 \in S_0$, $\phi_1 \in S_1$, and $u \in \mathcal{U}_0$ there corresponds a unique solution ϕ called the *response*. The quality of the response is measured through a functional called the *cost functional* and may be given by an expression of the form,

$$\begin{aligned} J(\phi_0, \phi_1, u) \equiv \alpha \int_0^T g_1(t, \phi(t), \dot{\phi}(t))\,dt \\ + \beta g_2(\phi(T), \dot{\phi}(T)) + \lambda g_3(u) \end{aligned} \tag{63}$$

$\alpha + \beta > 0$; $\alpha, \beta, \gamma \geq 0$, where g_1, g_2, and g_3 are suitable functions to be defined shortly. One may interpret g_1 to be a measure of discrepancy between a desired response and the one arising from the given policy $\{\phi_0, \phi_1, u\}$. The function g_2 is a measure of distance between a desired target and the one actually realized. The function g_3 is a measure of the cost of control applied to system (62). A more concrete expression for J will be given in the following section. Let $P \equiv S_0 \times S_1 \times \mathcal{U}_0$ denote the set of admissible policies or controls. The question is, does there exist a policy $p^0 \in P$ such that $J(p^0) \leq J(p)$ for all $p \in P$? An element $p^0 \in P$ satisfying this property is called an *optimal policy*. A set of sufficient conditions for the existence of an optimal policy is given in the following result.

Theorem 11. Consider system (62) with the cost functional (63) and let S_0, S_1, and $\mathcal{U}_0$ be closed bounded convex subsets of V, H, and $L_2(I, E)$, respectively. Suppose for each $(\xi, \eta) \in V \times H$, $t \to g_1(t, \xi, \eta)$ is measurable on I and, for each $t \in I$, the functions $(\xi, \eta) \to g_1(t, \xi, \eta)$ and $(\xi, \eta) \to g_2(\xi, \eta)$ are weakly lower semicontinuous on $V \times H$ and the function g_3 is weakly lower

semicontinuous on $L_2(I, E)$. Then, there exists an optimal policy,

$$p^0 = \left(\phi_0^0, \phi_1^0, u^0\right) \in P$$

Another problem of considerable interest is the question of existence of time-optimal controls. Consider system (33) in the form (34) with

$$f = Bu, \qquad \tilde{f} = \begin{pmatrix} 0 \\ Bu \end{pmatrix}$$

and solution given by:

$$\psi(t) = S(t)\psi_0 + \int_0^t S(t-\theta)\tilde{f}(\theta)\, d\theta \tag{35′}$$

where $S(t)$, $t \geq 0$, is the c_0-semigroup in $X \equiv V \times H$ with the generator $-\tilde{A}$ as given Eq. (34). Here, one is given the initial and the desired final states ψ_0, $\psi_1 \in X$ and the set of admissible controls $\mathcal{U}_0$. Given that the system is controllable from state ψ_0 to ψ_1 in finite time, the question is, does there exist a control that does the transfer in minimum time? A control satisfying this property is called a *time-optimal control*. We now present a result of this kind.

Theorem 12. If $\mathcal{U}_0$ is a closed bounded convex subset of $L_2(I, E)$ and if systems (34) and (35′) are exactly controllable from the state ψ_0 to $\psi_1 \in X$, then there exists a time-optimal control.

Proof. Let $\psi(u)$ denote the response of the system corresponding to control $u \in \mathcal{U}_0$; that is,

$$\psi(u)(t) = S(t)\psi_0 + \int_0^t S(t-\theta)\begin{pmatrix} 0 \\ Bu(\theta) \end{pmatrix} d\theta$$

Let $\mathcal{U}_{00} \subset \mathcal{U}_0$ denote the set of all controls that transfer the system from state ψ_0 to state ψ_1 in finite time. Define

$$\mathcal{I} = \{t \geq 0 : \psi(u)(0) = \psi_0,\ \psi(u)(t) = \psi_1, u \in \mathcal{U}_{00}\}$$

and $\tau^* = \inf\{t \geq 0 : t \in \mathcal{I}\}$. We show that there exists a control $u^* \in \mathcal{U}_{00}$ having the transition time τ^*. Let $\{\tau_n\} \in \mathcal{I}$ such that τ_n is nonincreasing and $\tau_n \to \tau^*$. Since $\tau_n \in \mathcal{I}$ there exists a sequence $u_n \in \mathcal{U}_{00} \subset \mathcal{U}_0$ such that $\psi(u_n)(0) = \psi_0$ and $\psi(u_n)(\tau_n) = \psi_1$. Denote by f_n the element $\binom{0}{B}u_n$. By virtue of our assumption, $\mathcal{U}_0$ is weakly compact, B is bounded, and there exists a subsequence of the sequence $\{f_n\}$ relabeled as $\{f_n\}$ and

$$f^* = \begin{pmatrix} 0 \\ Bu^* \end{pmatrix} \in L_2(I, X)$$

with $u^* \in \mathcal{U}_0$ such that $f_n \to f^*$ weakly in $L_2(I, X)$. We must show that $u^* \in \mathcal{U}_{00}$. Clearly, by definition of $\{\tau_n\}$ we have

$$\psi_1 = \psi(u_n)(\tau_n) \equiv S(\tau_n)\psi_0 + \int_0^{\tau_n} S(\tau_n - \theta) f_n(\theta)\, d\theta \qquad \text{for all} \quad n$$

Let $x^* \in X^* \equiv V^* \times H$, where X^* is the dual of the Banach space X which is the space of continuous linear functionals on X; for example,

$$X = L_p, \qquad 1 \leq p < \infty$$
$$X^* = L_q \qquad \text{with} \quad p^{-1} + q^{-1} = 1$$

Then,

$$x^*(\psi_1) = x^*(S(\tau_n)\phi_0) + x^*\left(\int_0^{\tau_n} S(\tau_n - \theta) f_n(\theta)\, d\theta\right) \tag{64}$$

By virtue of the c_0-property of the semigroup $S(t)$, $t \geq 0$,

$$\lim_{n\to\infty} x^*(S(\tau_n)\psi_0) = x^*\left(S(\tau^*)\psi_0\right) \tag{65}$$

Splitting the integral in Eq. (64) into two parts, we have

$$\begin{aligned} &x^*\left(\int_0^{\tau_n} S(\tau_n - \theta) f_n(\theta)\, d\theta\right) \\ &\quad = x^*\left(\int_0^{\tau^*} S\left(\tau_n - \tau^*\right) S(\tau^* - \theta) f_n(\theta)\, d\theta\right) \\ &\qquad + x^*\left(\int_{\tau^*}^{\tau_n} S(\tau_n - \theta) f_n(\theta)\, d\theta\right) \\ &\quad = \left\langle \int_0^{\tau^*} S(\tau^* - \theta) f_n(\theta)\, d\theta,\ S^*\left(\tau_n - \tau^*\right) x^* \right\rangle_{X, X^*} \\ &\qquad + x^*\left(\int_{\tau^*}^{\tau_n} S(\tau_n - \theta) f_n(\theta)\, d\theta\right) \end{aligned}$$

where S^* is the dual of the operator S. $\mathcal{L}_u(X, Y)$ is the space of unbounded linear operators from X into Y. Let $\{x^*, y^*\} \in X^* \times Y^*$ be the set of all pairs for which

$$\langle y^*, Sx\rangle_{Y^*, Y} = \langle x^*, x\rangle_{X^*, X}$$

for all $x \in D(S) \subset X$ where

$$x^*(x) = \langle x^*, x\rangle_{X^*, X} = \langle x, x^*\rangle_{X, X^*}$$

is the duality pairing between the elements $x \in X$ and $x^* \in X^*$ or the value of x^* at x. If $D(S)$ is dense in X (i.e., closure of $D(S) = X$), then the above relation determines uniquely the dual S^* of S and its domain $D(S^*) \subset Y^*$. If $D(S) = X$, then $S \in \mathcal{L}(X, Y)$ and $S^* \in \mathcal{L}(Y^*, X^*)$. Clearly,

$$\begin{aligned} &\int_0^{\tau^*} S(\tau^* - \theta) f_n(\theta)\, d\theta \\ &\qquad \xrightarrow{w} \int_0^{\tau^*} S(\tau^* - \theta) f^*(\theta)\, d\theta \qquad \text{in } X \end{aligned} \tag{66}$$

and since V and hence V^* are all reflexive Banach spaces S^* is a c_0-semigroup in X^* and consequently

$$S^*\left(\tau_n - \tau^*\right) x^* \xrightarrow{s} x^* \qquad \text{in } X^* \tag{67}$$

Further, by the c_0-property of $S(t)$, $t \geq 0$ there exists a finite $M > 0$ such that

$$\left| x^* \left(\int_{\tau^*}^{\tau_n} S(\tau_n - \theta) f_n(\theta)\, d\theta \right) \right|$$

$$\leq M \|x^*\|_{X^*} \left(\int_{\tau^*}^{\tau_n} \|f_n(\theta)\|_X^2\, d\theta \right)^{1/2} \left(\tau_n - r^*\right)^{1/2} \quad (68)$$

Since $\mathcal{U}_0$ is bounded, it follows from this that

$$\lim_{n\to\infty} x^* \left(\int_{\tau^*}^{\tau_n} S(\tau_n - \theta) f_n(\theta)\, d\theta \right) = 0$$

Using Eqs. (65) to (67) in (64) we obtain:

$$x^*(\psi_1) = x^*\big(S(\tau^*)\psi_0\big) + x^* \left(\int_0^{\tau^*} S(\tau^* - \theta) f^*(\theta)\, d\theta \right)$$

for all $x^* \in X^*$. Hence,

$$\psi_1 = S(\tau^*)\psi_0 + \int_0^{\tau *} S(\tau^* - \theta) f^*(\theta)\, d\theta$$

and $u^* \in \mathcal{U}_{00}$. This completes the proof. □

The method of proof of the existence of time-optimal controls presented above applies to much more general systems.

F. Necessary Conditions of Optimality

After the questions of controllability and existence of optimal controls are settled affirmatively, one is faced with the problem of determining the optimal controls. For this purpose, one develops certain necessary conditions of optimality and constructs a suitable algorithm for computing the optimal (extremal) controls. We present here necessary conditions of optimality for system (62) with a quadratic cost functional of the form,

$$\begin{aligned} J(u) \equiv \alpha \int_0^T (C\phi(t) - z_1(t), C\phi(t) - z_1(t))_{H_1}\, dt \\ + \beta \int_0^T (D\dot{\phi}(t) - z_2(t), D\dot{\phi}(t) - z_2(t))_{H_2}\, dt \\ + \gamma \int_0^T (N(t)u, u)_E\, dt \end{aligned} \quad (69)$$

where $\alpha, \beta, \gamma > 0$. The output spaces, where observations are made, are given by two suitable Hilbert spaces H_1 and H_2 with output operators $C \in \mathcal{L}(H, H_1)$ and $D \in \mathcal{L}(H, H_2)$. The desired trajectories are given by $z_1 \in L_2(I, H_1)$ and $z_2 \in L_2(I, H_2)$. The last integral in Eq. (69) gives a measure of the cost of control with $N(t) \geq \delta I$ for all $t \in I$, with $\delta > 0$. We assume that $N(t) = N^*(t)$, $t \geq 0$. Our problem is to find the necessary and sufficient conditions an optimal control must satisfy. By Theorem 11, we know that, for the cost functional (69) subject to the dynamic constraint (62), an optimal control exists. Since in this case J is strictly convex, there is, in fact, a unique optimal control. For characterization of optimal controls, the concept of Gateaux differentials plays a central role. A real-valued functional f defined on a Banach space X is said to be Gateaux differentiable at the point $x \in X$ in the direction $h \in X$ if

$$\lim_{\varepsilon\to 0} \left\{ \frac{f(x + \varepsilon h) - f(x)}{\varepsilon} \right\} = f'(x, h) \quad (70)$$

exists. In general, $h \to f'(x, h)$ is a homogeneous functional, and, in case it is linear in h, we write:

$$f'(x, h) = (f'(x), h)_{X^*, X} \quad (71)$$

with the Gateaux derivative $f'(x) \in X^*$. Since the functional J, defined on the Hilbert space $L_2(I, E)$, is Gateaux differentiable and strictly convex, and the set of admissible controls $\mathcal{U}_0$ is a closed convex subset of $L_2(I, E)$, a control $u^0 \in \mathcal{U}_0$ is optimal if and only if

$$(J'(u^0), u - u^0) \geq 0 \qquad \text{for all} \quad u \in \mathcal{U}_0 \quad (72)$$

Using this inequality, we can develop the necessary conditions of optimality.

Theorem 13. Consider system (62) with the cost functional (69) and $\mathcal{U}_0$, a closed bounded convex subset of $L_2(I, E)$. For $u^0 \in \mathcal{U}_0$ to be optimal, it is necessary that there exists a pair

$$\{\phi^0, \psi^0\} \in C(\bar{I}, V) \times C(\bar{I}, V)$$

with

$$\{\dot{\phi}^0, \dot{\psi}^0\} \in C(\bar{I}, H) \times C(\bar{I}, H)$$

satisfying the equations:

$$\begin{aligned} \ddot{\phi}^0 + A\phi^0 &= f + Bu^0 \\ \phi^0(0) &= \phi_0, \\ \dot{\phi}^0(0) &= \phi_1 \end{aligned} \quad (73a)$$

$$\begin{aligned} &\ddot{\psi}^0 + A^*\psi^0 + \int_t^T g_1\, d\theta + g_2 = 0 \\ &\psi^0(T) = 0, \qquad \dot{\psi}^0(T) = 0 \\ &g_1 = 2\alpha C^* \Lambda_1 \big(C\phi^0 - z_1\big) \\ &g_2 = 2\beta D^* \Lambda_2 \big(D\dot{\phi}^0 - z_2\big) \end{aligned} \quad (73b)$$

and the inequality

$$\int_0^T \left(u - u^0, 2\gamma N u^0 + \Lambda_E^{-1} B^* \dot{\psi}^0\right)_E dt \geq 0 \qquad (74)$$

for all $u \in \mathcal{U}_0$.

Proof. By taking the Gateaux differential of J at u^0 in the direction w we have

$$\begin{aligned}(J'(u^0), w) \equiv \int_0^T dt \{ & 2\alpha \left(C\hat{\phi}(u^0, w), C\phi(u^0) - z_1\right)_{H_1} \\ & + 2\beta \left(D\dot{\hat{\phi}}(u^0, w), D\dot{\phi}(u^0) - z_2\right)_{H_2} \\ & + 2_\gamma (w, Nu^0)_E \} \end{aligned} \qquad (75)$$

where $\hat{\phi}(u^0, w)$ denotes the Gateaux differential of ϕ at u^0 in the direction w, which is given by the solution of

$$\begin{aligned} \ddot{\hat{\phi}}(u^0, w) + A\hat{\phi}(u^0, w) &= Bw \\ \hat{\phi}(u^0, w)(0) = 0, \qquad \dot{\hat{\phi}}(u^0, w)(0) &= 0 \end{aligned} \qquad (76)$$

Introducing the duality maps,

$$\Lambda_1 : H_1 \to H_1^*, \qquad \Lambda_2 : H_2 \to H_2^*$$

in expression (75) and defining

$$\begin{aligned} g_1(t, \phi(u^0)) &\equiv 2\alpha C^* \Lambda_1 \left(C\phi(u^0) - z_1(t)\right) \\ g_2(t, \phi(u^0)) &\equiv 2\beta D^* \Lambda_2 \left(D\dot{\phi}(u^0) - z_2(t)\right) \end{aligned}$$

we obtain:

$$\begin{aligned}(J'(u^0), w) = \int_0^T dt \{ & \left(\hat{\phi}(u^0, w), g_1\right)_H + \left(\dot{\hat{\phi}}(u^0, w), g_2\right)_H \\ & + (w, 2\gamma N u^0)_E \} \end{aligned} \qquad (77)$$

for all $w \in \mathcal{U}_0$.

Defining $\hat{\phi}_1 = \hat{\phi}$, $\hat{\phi}_2 = \dot{\hat{\phi}}$ one can rewrite Eq. (76) as a first-order evolution equation:

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix}\hat{\phi}_1 \\ \hat{\phi}_2\end{pmatrix} &= \begin{pmatrix} 0 & I \\ -A & 0 \end{pmatrix}\begin{pmatrix}\hat{\phi}_1 \\ \hat{\phi}_2\end{pmatrix} + \begin{pmatrix} 0 \\ Bw \end{pmatrix} \\ \begin{pmatrix}\hat{\phi}_1(0) \\ \hat{\phi}_2(0)\end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned} \qquad (78)$$

Then, by introducing the adjoint evolution equation,

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix}p_1 \\ p_2\end{pmatrix} &= -\begin{pmatrix} 0 & -A^* \\ I & 0 \end{pmatrix}\begin{pmatrix}p_1 \\ p_2\end{pmatrix} + \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \\ \begin{pmatrix}p_1(T) \\ p_2(T)\end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned} \qquad (79)$$

one can easily verify from Egs. (78) and (79) that

$$\int_0^T \{(\hat{\phi}_1, g_1)_H + (\hat{\phi}_2, g_2)_H\}\, dt = -\int_0^T (Bw, p_2)_H\, dt \qquad (80)$$

Using Eq. (80) in Eq. (77) and the duality map Λ_E: $E \to E^*$, we obtain, for $w = u - u^0$,

$$\begin{aligned}(J'(u^0), u - u^0) = \int_0^T \left(2\gamma N u^0 - \Lambda_E^{-1} B^* p_2, u - u^0\right)_E dt \\ u \in \mathcal{U}_0 \end{aligned} \qquad (81)$$

Defining $\psi^0(t) = \int_t^T p_2(\theta)\, d\theta$, one obtains the adjoint equation (73b), and the necessary inequality (74) follows from Eqs. (72) and (81).

REMARK 1. In case $\beta = 0$, the adjoint equation is given by a differential equation rather than the integro-differential equations (73b). That is, p_2 satisfies the equation:

$$\begin{aligned} \ddot{p}_2 + A^* p_2 + g_1 &= 0 \\ p_2(T) = 0, \qquad \dot{p}_2(T) &= 0 \end{aligned}$$

and in Eq. (74) one may replace $\dot{\psi}^0$ by $-p_2$.

REMARK 2. In case of terminal observation, the cost functional (69) may be given by:

$$\begin{aligned} J(u) \equiv \alpha \|C\phi(T) - z_1\|_{H_1}^2 + \beta \|D\dot{\phi}(T) - z_2\|_{H_2}^2 \\ + \gamma \int_0^T (N(t)u, u)_E\, dt \end{aligned} \qquad (82)$$

where $z_1 \in H_1$ and $z_2 \in H_2$.

In this case, the necessary conditions of optimality are given by:

$$\begin{aligned} \int_0^T \left(u - u^0, 2\gamma N u^0 - \Lambda_E^{-1} B^* p_2\right) dt \geq 0 \\ u \in \mathcal{U}_0 \end{aligned} \qquad (83)$$

where p_2 satisfies the differential equation,

$$\begin{aligned} \frac{d^2 p_2}{dt^2} + A^* p_2 &= 0 \\ p_2(T) = g_2 &\equiv 2\beta D^* \Lambda_2 \left(D\dot{\phi}^0(T) - z_2\right) \\ \dot{p}_2(T) = -g_1 &\equiv -2\alpha C^* \Lambda_1 \left(C\phi^0(T) - z_1\right) \end{aligned} \qquad (84)$$

In recent years several interesting necessary conditions of optimality for time- optimal control problems have been reported. We present here one such result. Suppose the system is governed by the evolution equation,

$$\frac{dy}{dt} = Ay + u$$

in a Banach space X, and let

$$\mathcal{U} \equiv \left\{u \in L_p^{\text{loc}}(R_0, X) : u(t) \in B_1\right\}$$

with $p > 1$, $B_1 =$ unit ball in X, denote the class of admissible controls.

Theorem 14 (Maximum Principle.) Suppose A is the generator of a c_0-semigroup $S(t)$, $t \geq 0$, in X and there exists a $t > 0$ such that $S(t)X = X$. Let y_0, $y_1 \in X$, and suppose u^0 is the time-optimal control, with transition time τ, that steers the system from the initial state y_0 to the final state y_1. Then there exists an $x^* \in X^*$ (=dual of X) such that

$$\begin{aligned}\langle S^*(\tau - t)x^*, u^0(t)\rangle_{X^*,X} &= \sup\left\{\langle S^*(\tau - t)x^*, e\rangle, e \in B_1\right\} \\ &= |S^*(\tau - t)x^*|_{X^*} \end{aligned} \tag{85}$$

Suppose X is a reflexive Banach space and there exists a continuous map v: $X^*\backslash\{0\} \to X$ such that, for $\xi^* \in X^*$,

$$|v(\xi^*)|_X = 1$$

and

$$\langle \xi^*, v(\xi^*)\rangle_{X^*,X} = |\xi^*|_{X^*}$$

then the optimal control is bang-bang and is given by $u^0(t) = v(S^*(\tau - t)x^*)$, and it is unique if X is strictly convex.

Maximum principle for more general control problems are also available.

G. Computational Methods

In order to compute the optimal controls one is required to solve simultaneously the state and adjoint equations (73a) and (73b), along with inequality (74). In case the admissible control set $\mathcal{U}_0 \equiv L_2(I, E)$, the optimal control has the form:

$$u^0 = -(1/2\gamma)N^{-1}\Lambda_E^{-1}B^*\psi^0$$

Substituting this expression in Eqs. (73a) and (73b), one obtains a system of coupled evolution equations, one with initial conditions and the other with final conditions specified. This is a two-point boundary-value problem in an ∞-dimensional space, which is a difficult numerical problem. However, in general one can develop a gradient-type algorithm to compute an approximating sequence of controls converging to the optimal. The required gradient is obtained from the necessary condition (74). The solutions for the state and adjoint equations, (73a) and (73b), can be obtained by use of any of the standard techniques for solving partial differential equations, for example, the finite difference, finite element, or Galerkin method.

In order to use the result of Theorem 13 to compute the optimal controls, one chooses an arbitrary control $u^1 \in \mathcal{U}_0$ and solves Eq. (73a) to obtain ϕ^1, which is then used in Eq. (73b) to obtain ψ^1. Then, on the basis of Eq. (74) one takes

$$J'(u^1) = 2\gamma Nu^1 + \Delta_E^{-1}B^*\psi^1$$

as the gradient at u^1 and constructs a new control $u^2 = u^1 - \varepsilon_1 J'(u^1)$, with $\varepsilon_1 > 0$ sufficiently small so that $J(u^2) \leq J(u^1)$. This way one obtains a sequence of approximating controls,

$$u^{n+1} = u^n - \varepsilon_n J'(u^n)$$

with $J(u^{n+1}) \leq J(u^n)$, $n = 1, 2, \ldots$. In practical applications, a finite number of iterations produces a fairly good approximation to the optimal control.

H. Stochastic Evolution Equations

In this section we present a very brief account of stochastic linear systems. Let $(\Sigma, \mathcal{A}, P)$ denote a complete probability space and $\mathcal{F}_t$, $t \geq 0$, a nondecreasing family of right-continuous, completed subsigma algebras of the σ-algebra $\mathcal{A}$; that is, $\mathcal{F}_s \subset \mathcal{F}_t$ for $0 \leq s \leq t$. Let H be a real separable Hilbert space and $\{W(t), t \geq 0\}$ an H-valued Wiener process characterized by the properties:

(a) $P\{W(0) = 0\} = 1$,

(b) $W(t)$, $t \geq 0$, has independent increments over disjoint intervals, and

(c) $E\left\{e^{i(W(t),h)}\right\} = \exp[-t/2(Qh, h)]$,

where $E\{\cdot\} \equiv \int_\Sigma \{\cdot\}\, dP$ and $Q \in \mathcal{L}^+(H)$ is the space of positive self-adjoint bounded operators in H. Symbolically, a stochastic linear differential equation is written in the form

$$dy = Ay\, dt + \sigma(t)\, dW(t), \qquad t \geq 0, \qquad y(0) = y_0$$

where A is the generator of a c_0-semigroup $S(t)$, $t \geq 0$, in H and $\sigma \in \mathcal{L}(H)$ and y_0 is an H-valued random variable independent of the Wiener process. The solution y is given by:

$$y(t) = S(t)y_0 + \int_0^t S(t - \theta)\sigma(\theta)\, dW(\theta), \qquad t \geq \theta$$

Under the given assumptions one can easily show that $E|y(t)|_H^2 < \infty$ for finite t, whenever $E|y_0|_H^2 < \infty$, and further $y \in C(I, H)$ a.s. (a.s. $\equiv$ with probability one). In fact, $\{y(t), t \geq 0\}$ is an $\mathcal{F}_t$–Markov random process and one can easily verify that

$$E\{y(t) \mid \mathcal{F}_\tau\} = S(t - \tau)y(\tau)$$

for $0 \leq \tau \leq t$. The covariance operator $C(t)$, $t \geq 0$, for the process $y(t)$, $t \geq 0$, defined by:

$$(C(t)h, g) = E\{(y(t) - Ey(t), h)(y(t) - Ey(t), g)\}$$

is then given by:

$$(C(t)h, g) = \int_0^t (S(t-\theta)\sigma(\theta)Q\sigma^*(\theta) \times S^*(t-\theta)h, g)\, d\theta + (S(t)C_0S^*(t)h, g)$$

Denoting the positive square root of the operator Q by $\sqrt{Q}$ we have

$$(C(t)h, h) = \int_0^t |\sqrt{Q}\sigma^*(\theta)S^*(t-\theta)h|_H^2\, d\theta + \left(C_0S^*(t)h, S^*(t)h\right)$$

This shows that $C(t) \in \mathscr{L}^+(H)$ if and only if the condition,

$$S^*(t)h = 0, \qquad \text{or}$$
$$\sigma^*(\theta)S^*(t-\theta)h \equiv 0, \qquad 0 \le \theta \le t$$

implies $h = 0$. This is precisely the condition for (approximate) controllability as seen in Theorem 9. Hence, the process $y(t)$, $t \ge 0$, is a nonsingular H-valued Gaussian process if and only if the system is controllable. Similar results hold for more general linear evolution equations on Banach spaces with operators A and σ both time varying.

Linear stochastic evolution equations of the given form arise naturally in the study of nonlinear filtering of ordinary Ito stochastic differential equations in R^n. Such equations are usually written in the weak form,

$$d\pi_t(f) = \pi_t(Af)\, dt + \pi_t(\sigma f)\, dW_t$$
$$t \ge 0, \qquad f \in D(A)$$

where A is a second-order partial differential operator and $\{W(t), t \ge 0\}$ is an $\mathbb{R}^d$-valued Wiener process ($d \le n$). One looks for solutions π_t, $t \ge 0$, that belong to the Banach space of bounded Borel measures satisfying $\pi_t(f) \ge 0$ for $f \ge 0$. Questions of existence of solutions for stochastic systems of the form,

$$dy = (A(t)y + f(t))\, dt + \sigma(t)\, dW$$
$$y(0) = y_0, \qquad t \in (0, T)$$
$$dz = -((A^*(t) - B(t))z + g)\, dt + \sigma(t)\, dW$$
$$z(T) = 0, \qquad t \in (0, T)$$

have been studied in the context of control theory. A fundamental problem arises in the study of the second equation and it has been resolved by an approach similar to the Lax–Milgram theorem in Hilbert space. Questions of existence and stability of solutions of nonhomogeneous boundary value problems of the form,

$$\frac{\partial y}{\partial t} + A(t)y = f(t), \qquad \text{on} \quad I \times \Omega$$
$$By = g(t), \qquad \text{on} \quad I \times \partial\Omega$$
$$u(0) = y_0, \qquad \text{on} \quad \Omega$$

have been studied where f and g have been considered as generalized random processes and y_0 as a generalized random variable. Stability of similar systems with g replaced by a generalized white noise process has been considered recently. Among other things, it has been shown that $y(t)$, $t \ge 0$, is a Feller process on H, a Hilbert space, and there exists a Feller semigroup T_t, $t \ge 0$, on $C_b(H)$, a space of bounded continuous functions on H, whose dual U_t determines the flow $\mu_t = U_t\mu_0$, $t \ge 0$, of the measure induced by $y(t)$ on H. μ_t, $t \ge 0$, satisfies the differential equation:

$$\frac{d}{dt}\mu_t(f) = \mu_t(Gf), \qquad t \ge 0$$

for all $f \in D(G) \subset C_b(H)$ where G is the infinitesimal generator of the semigroup $T_t, t \ge 0$. Optimal control problems for a class of very general linear stochastic systems of the form,

$$dy = (A(t)y + B(t)u)\, dt + \sigma(t)\, dW(t)$$
$$J(u) = E\left\{\int^T \left[|Cy - z_d|^2 + (Nu, u)\right] dt\right\} \equiv \min$$

have been considered in the literature, giving results on the existence of optimal controls and necessary conditions of optimality, including feedback controls. In this work, A, B, σ, C, and N were considered as operator-valued random processes.

III. NONLINEAR EVOLUTION EQUATIONS AND DIFFERENTIAL INCLUSIONS

The two major classes of nonlinear systems that occupy most of the literature are the semilinear and quasilinear systems. However, control theory for such systems has not been fully developed; in fact, the field is wide open. In this section, we shall sample a few results.

A. Existence of Solutions, Nonlinear Semigroup

We consider the questions of existence of solutions for the two major classes of systems, semilinear and quasilinear. In their abstract form we can write them as

$$\frac{d\phi}{dt} + A(t)\phi = f(t, \phi) \qquad \text{(semilinear)} \quad (86)$$

$$\frac{d\phi}{dt} + A(t, \phi)\phi = f(t, \phi) \qquad \text{(quasilinear)} \quad (87)$$

and consider them as evolution equations on suitable state space which is generally a topological space having the structure of a Banach space, or a suitable manifold therein.

For example, let us consider the semilinear parabolic equation with mixed initial and boundary conditions:

$$\begin{aligned} \frac{\partial \phi}{\partial t} + L\phi &= g(t, x; \phi, D\phi, \ldots, D^{2m-1}\phi) \\ &\qquad (t, x) \in I \times \Omega \equiv Q \\ \phi(0, x) &= \phi_0(x), \qquad x \in \Omega \\ D_\nu^k \phi &= 0, \qquad 0 \le k \le m-1 \\ &\qquad (t, x) \in I \times \partial\Omega \end{aligned} \tag{88}$$

where

$$(L\phi)(t, x) \equiv \sum_{|\alpha| \le 2m} a_\alpha(t, x) D^\alpha \phi \tag{89}$$

and $D_\nu^k \phi = \partial^k \phi / \partial \nu^k$ denotes the kth derivative in the direction of the normal ν to $\partial\Omega$. We assume that L is strongly elliptic with principal coefficients a_α, $|\alpha| = 2m$, in $C(\bar{Q})$ and the lower order coefficients a_α, $|\alpha| \le 2m-1$, $L_\infty(Q)$, and further they are all Hölder continuous in t uniformly on $\bar{\Omega}$. Let $1 < p < \infty$ and define the operator-valued function $A(t)$, $t \in I$, by:

$$\begin{aligned} D(A(t)) \equiv \{\psi \in X = L_p(\Omega) : (L\psi)(t, \cdot) \in X \\ \text{and } D_\nu^k \psi \equiv 0 \text{ on } \partial\Omega, 0 \le k \le m-1\} \end{aligned} \tag{90}$$

The domain of $A(t)$ is constant and is given by:

$$D \equiv W^{2m,p} \cap W_0^{m,p} \tag{91}$$

Then, one can show that for each $t \in I$, $-A(t)$ is the generator of an analytic semigroup and there exists an evolution operator $U(t, \tau) \in \mathcal{L}(X)$, $0 \le \tau \le t \le T$, that solves the abstract Cauchy problem:

$$\begin{aligned} \frac{dy}{dt} + A(t)y &= 0 \\ y(0) &= y_0 \end{aligned} \tag{92}$$

for each $y_0 \in X$; that is, $y(t) = U(t, 0)y_0$, $t \in I$, with $y \in C(\bar{I}, X)$ and $\dot{y} \in C((0, T], X)$. In general, if $f \in L_p(I, X)$, then y, given by:

$$y(t) = U(t, 0)y_0 + \int_0^t U(t, \tau) f(\tau)\, d\tau \tag{93}$$

is a mild solution of the Cauchy problem,

$$\begin{aligned} \frac{dy}{dt} + A(t)y &= f \\ y(0) &= y_0 \end{aligned} \tag{94}$$

This would be the generalized (weak) solution of the parabolic initial boundary value problem (88) if g were replaced by $f \in L_p(I, X)$. In order to solve problem (88), one must introduce an operator f such that

$$f(t, u) \equiv g(t, \cdot; u, D^1 u, \ldots, D^{2m-1} u) \in X \qquad \text{a.e.}$$

for u in a suitable subspace Y with $D(A) \subset Y \subset X$. Problem (88) can then be considered as an abstract Cauchy problem,

$$\begin{aligned} \frac{d\phi}{dt} + A(t)\phi &= f(t, \phi) \\ \phi(0) &= \phi_0 \end{aligned} \tag{95}$$

In view of Eq. (93), a mild solution of Eq. (95) is given by a solution of the integral equation,

$$\phi(t) = U(t, 0)\phi_0 + \int_0^t U(t, \theta) f(\theta, \phi(\theta))\, d\theta \tag{96}$$

if one exists. Defining the operator G by:

$$(G\phi)(t) \equiv U(t, 0)\phi_0 + \int_0^t U(t, \theta) f(\theta, \phi(\theta))\, d\theta$$

one then looks for a fixed point for G, that is, an element ϕ such that $\phi = G\phi$. Using *a priori* estimates, the most difficult part of the program, one can establish the existence of a solution by use of a suitable fixed-point theorem—for example, Banach, Schauder, or Leray–Schauder fixed-point theorems. We state the following result without proof.

Theorem 15. Consider the semilinear parabolic problem (88) in the abstract form (95) and suppose A generates the evolution operator $U(t, \tau)$, $0 \le \tau \le t \le T$, and f satisfies the properties:

$$\begin{aligned} \text{(F1)} \qquad & \|f(t, u)\|_X \le c\{1 + \|A^\beta(t)u\|_X\} \\ & \quad t \in I \\ & \text{for constants} \quad c > 0, \quad 0 \le \beta < 1 \\ & \text{and } u \in D(A^\beta), \end{aligned} \tag{97}$$

$$\begin{aligned} \text{(F2)} \qquad & \|f(t, u) - f(t, w)\|_X \\ & \le C\{\|A^\beta(t)v - A^\beta(t)w\|_X^\rho\} \\ & \quad t \in I \\ & \text{for some} \quad 0 < \rho \le 1, \\ & u, w \in D(A^\beta) \end{aligned} \tag{98}$$

Then, Eq. (95) has a mild solution $\phi \in C(I, X)$, hence the semilinear parabolic equation (88) has a generalized solution. The solution is unique if $\rho = 1$.

REMARK 1. Condition (F1) is satisfied if the function g satisfies the growth condition:

$$|g(t, x, u, Du, \ldots, D^{2m-1}u)| \le k\left\{1 + \sum_{j=0}^{2m-1} |D^j u|_j^r\right\}$$

for $0 \le r_j \le (2m + n/q)/(j + n/q)$, $1 < q < \infty$, and a number β satisfying $(2m-1)/2m < \beta < 1$. In the case

$\rho = 1$, condition (F2) is satisfied if the function g is Lipschitz in the last $2m$ variables uniformly with respect to $(t, x) \in \overline{I \times \Omega}$.

If the coefficients $\{a_\alpha, |\alpha| \le 2m\}$ in Eq. (89) are also dependent on ϕ so that $\{a_\alpha = a_\alpha(t, x; \phi, D^1\phi, \ldots, D^{2m-1}\phi)$ then system (88) becomes a quasilinear system and parabolic if

$$(-1)^m \operatorname{Re}\left\{ \sum_{|\alpha|=2m} a_\alpha(t, x; \eta)\xi^\alpha \right\} \ge c|\xi|^{2m} \qquad c > 0 \quad (99)$$

for $(t, x) \in \bar{Q}$ and $\eta \in R^N$, where $N = \sum_{j=0}^{2m-1} N_j$ with N_j denoting the number of terms representing derivatives of order exactly j appearing in the arguments of $\{a_\alpha\}$. System (88) then takes the form:

$$\frac{d\phi}{dt} + A(t, \phi)\phi = f(t, \phi) \qquad \phi(0) = \phi_0 \quad (100)$$

This problem is again solved by use of *a priori* estimates and a fixed-point theorem under the following assumptions:

(A1) The operator $A_0 = A(0, \phi_0)$ is a closed operator with domain D dense in X and

$$\left\|(\lambda I - A_0)^{-1}\right\| \le k/(1 + |\lambda|) \qquad \text{for all } \lambda, \quad \operatorname{Re}\lambda \le 0.$$

(A2) A_0^{-1} is a completely continuous operator that is; it is continuous in X and maps bounded sets into compact subsets of X.

(A3) There exist numbers ε, ρ satisfying $0 < \varepsilon \le 1$, $0 < \rho \le 1$, such that for all $t, \tau \in I$,

$$\|(A(t, u) - A(\tau, w))A^{-1}(\tau, w)\| \le k_R\left\{|t - \tau|^\varepsilon + \left\|A_0^\beta u - A_0^\beta w\right\|^\rho\right\}$$

for all u, w such that $\|A_0^\beta u\|, \|A_0^\beta w\| < R$ with k_R possibly depending on R.

(F1) For all $t, \tau \in I$,

$$\|f(t, v) - f(t, w)| \le k_R\left\{\left\|A_0^\beta v - A_0^\beta w\right\|^\rho\right\}$$

for all $v, w \in X$ such that $\|A_0^\beta v\|, \|A_0^\beta w\| \le R$.

Theorem 16. Under assumptions (A1) to (A3) and (F1) there exists a $t^* \in (0, T)$ such that Eq. (100) has at least one mild solution $\phi \in C([0, t^*], X)$ for each $\phi_0 \in D(A_0^\beta)$ with $\|A_0^\beta \psi_0\| \le R$. Further, if f also satisfies the Hölder condition in t, the solution is C^1 in $t \in (0, t^*]$. If $\rho = 1$, the solution is unique.

Proof. We discuss the outline of a proof. The differential equation (100) is converted into an integral equation and then one shows that the integral equation has a solution. Let $v \in C([0, t^*], X)$ and define:

$$A^v(t) \equiv A\left(t, A_0^{-\beta} v(t)\right), \qquad f^v(t) \equiv f\left(t, A_0^{-\beta} v(t)\right)$$

and consider the linear system,

$$\frac{dy}{dt} + A^v(t)y = f^v(t), \quad t \in [0, t^*] \qquad y(0) = \phi_0 \quad (101)$$

By virtue of assumptions (A1) to (A3), $-A^v(t)$, $t \in [0, t^*]$, is the generator of an evolution operator $U^v(t, \tau)$, $0 \le t \le t^*$. Hence, the system has a mild solution given by:

$$y^v(t) = U^v(t, 0)\phi_0 + \int_0^t U^v(t, \theta) f^v(\theta)\, d\theta \qquad 0 \le t \le t^* \quad (102)$$

Defining an operator G by setting

$$(Gv)(t) \equiv A_0^\beta U^v(t, 0)\phi_0 + \int_0^t A_0^\beta U^v(t, \theta) f^v(\theta)\, d\theta \quad (103)$$

one then looks for a fixed point of the operator G, that is, an element $v^* \in C([0, t^*], X)$ such that $v^* = Gv^*$. In fact, one shows, under the given assumptions, that for sufficiently small $t^* \in I$, there exists a closed convex set $K \subset C([0, t^*], X)$ such that $GK \subset K$ and GK is relatively compact in $C([0, t^*], X)$ and hence, by the Schauder fixed-point theorem (which is precisely as stated) has a solution $v^* \in K$. The solution (mild) of the original problem (100) is then given by $\phi^* = A_0^{-\beta} v^*$. This is a genuine (strong) solution if f is also Holder continuous in t. If $\rho = 1$, G has the contraction property and the solution is unique. □

According to our assumptions, for each $t \in [0, t^*]$ and $y \in D(A_0^\beta)$, the operator $A(t, y)$ is the generator of an analytic semigroup. This means that Theorem 16 can handle only parabolic problems and excludes many physical problems arising from hydrodynamics and wave propagation phenomenon including the semilinear and quasilinear symmetric hyperbolic systems discussed in Section I. This limitation is overcome by allowing $A(t, y)$, for each t, y in a suitable domain, to be the generator of a c_0-semigroup rather than an analytic semigroup. The fundamental assumptions required are:

(H1) X is a reflexive Banach space with Y being another Banach space which is continuously and densely

embedded in X and there is an isometric isomorphism S of Y onto X.

(H2) For $t \in [0, T]$, $y \in W \equiv$ an open ball in Y, $A(t, y)$ is the generator of a c_0-semigroup in X.

(H3) For $t, y \in [0, T] \times W$,

$$(SA(t, y) - A(t, y)S)S^{-1} = B(t, y) \in \mathcal{L}(Y, X)$$

and $\|B(t, y)\|_{\mathcal{L}(Y,X)}$ is uniformly bounded on $I \times Y$.

Theorem 17. Under assumptions (H1) to (H3) and certain Lipschitz and boundedness conditions for A and f on $[0, T] \times W$, the quasilinear system (100) has, for each $\phi_0 \in W$, a unique solution,

$$\phi \in C([0, t^*), W) \cap C^1([0, t^*), X) \qquad \text{for some} \quad 0 < t^* \leq T$$

The proof of this result is also given by use of a fixed-point theorem but without invoking the operator A_0. Here, for any $v \in C([0, t^*), W)$, one defines $A^v(t) = A(t, v(t))$, $f^v(t) = f(t, v(t))$ and $U^v(t, \tau)$, $0 \leq \tau \leq t \leq T$, the evolution operator corresponding to $-A^v$ and constructs the operator G by setting

$$(Gv)(t) = U^v(t, 0)\phi_0 + \int_0^t U^v(t, \tau) f^v(\tau)\, d\tau$$

where the expression on the right-hand side is the mild solution of the linear equation (101). Any v^* satisfying $v^* = Gv^*$ is a mild solution of the original problem (100).

From the preceding results it is clear that the solutions are defined only over a subinterval $(0, t^*) \subset (0, T)$ and it may actually blow up at time t^*. Mathematically this is explained through the existence of singularities, which physically correspond to the occurrence of, for example, turbulence or shocks in hydrodynamic problems. However, global solutions are defined for systems governed by differential equations with monotone operators. We present a few general results of this nature.

Let X be a Banach space with dual X^* and suppose A is an operator from $D(A) \subset X$ to X^*. The operator A is said to be monotone if

$$(Ax - Ay, x - y)_{X^*,X} \geq 0 \qquad \text{for all} \quad x, y \in D(A) \tag{104}$$

It is said to be demicontinuous on X if

$$Ax_n \overset{w}{\to} Ax_0 \qquad \text{in } X^* \qquad \text{whenever} \quad x_n \overset{s}{\to} x_0 \qquad \text{in } X \tag{105}$$

And, it is said to be hemicontinuous on X if

$$A(x + \theta y) \overset{w}{\to} Ax \qquad \text{in } X^* \qquad \text{whenever} \qquad \theta \to 0 \tag{106}$$

Let H be a real Hilbert space and V a subset of H having the structure of a reflexive Banach space with V dense in H. Let V^* denote the (topological) dual of V and suppose H is identified with its dual H^*. Then we have $V \subset H \subset V^*$. Using the theory of monotone operators we can prove the existence of solutions for nonlinear evolution equations of the form,

$$\frac{d\phi}{dt} + B(t)\phi = f, \qquad t \in I = (0, T) \qquad \phi(0) = \phi_0 \tag{107}$$

where $B(t)$, $t \in I$, is a family of nonlinear monotone operators from V to V^*.

Theorem 18. Consider system (107) and suppose the operator B satisfy the conditions:

(B1) B: $L_p(I, V) \to L_q(I, V^*)$ is hemicontinuous:

$$\|B\phi\|_{L_q(I,V^*)} \leq K_1\big(1 + \|\phi\|^{p-1}_{L_p(I,V)}\big) \tag{108}$$

where $K_1 > 0$, and $1 < p, q < \infty$, $1/p + 1/q = 1$.

(B2) For all $\phi, \psi \in L_p(I, V)$,

$$(B\phi - B\psi, \phi - \psi)_{L_q(I,V^*),L_p(I,V)} \geq 0 \tag{109}$$

That is, B is a monotone operator form $L_p(I, V)$ to $L_q(I, V^*)$.

(B3) There exists a nonnegative function C: $\mathbb{R} \to \bar{R}$ with $C(\xi) \to +\infty$ as $\xi \to \infty$ such that for $\psi \in L_p(I, V)$,

$$(B\psi, \psi)_{L_q(I,V^*),L_p(I,V)} \geq C(\|\psi\|)\|\psi\| \tag{110}$$

Then for each $\phi_0 \in H$ and $f \in L_q(I, V^*)$ system (107) has a unique solution $\phi \in L_p(I, V) \cap C(\bar{I}, H)$ and $\dot{\phi} \in L_q(I, V^*)$. Further, ϕ is an absolutely continuous V^*-valued function on $\bar{I}$.

It follows from the above result that, for $\phi_0 \in H$, $\phi \in C(\bar{I}, H)$ and hence $\phi(t) \in H$ for $t \geq 0$. For $f \equiv 0$, the mapping $\phi_0 \to \phi(t)$ defines a nonlinear evolution operator $U(t, \tau)$, $0 \leq \tau \leq t \leq T$, in H. In case B is time invariant, we have a nonlinear semigroup $S(t)$, $t \geq 0$, satisfying the properties:

(a) $S(0) = I$ and, as $t \to 0$,

(b) $S(t)\xi \to {}^s\xi$ in H and, due to uniqueness,

(c) $S(t + \tau)\xi = S(t)S(\tau)\xi, \xi \in H$.

Further, it follows from the equation $\dot{\phi} + B\phi = 0$ that $(\dot{\phi}(t), \phi(t)) + (B\phi(t), \phi(t)) = 0$; hence,

$$|\phi(t)|^2_H = |\phi_0|^2_H - 2\int_0^t (B\phi(\theta), \phi(\theta)\, d\theta, \qquad t \geq 0$$

Thus, by virtue of (B3), $|\phi(t)|_H \leq |\phi_0|_H$; that is, $|S(t)\phi_0|_H \leq |\phi_0|_H$. Hence the semigroup $\{S(t), t \geq 0\}$ is

a family of nonlinear contractions in H, and its generator is $-B$.

A classical example is given by the nonlinear initial boundary-value problem,

$$\begin{aligned} \frac{\partial \phi}{\partial t} - \sum_i \frac{\partial}{\partial x_i}\left(\left|\frac{\partial \phi}{\partial x_i}\right|^{p-2}\frac{\partial \phi}{\partial x_i}\right) &= f \qquad \text{in} \quad I \times \Omega = Q \\ \phi(0, x) &= \phi_0(x) \\ \phi(t, x) &= 0 \qquad \text{on} \quad I \times \partial\Omega \end{aligned} \tag{111}$$

For this example, $V = W_0^{1,p}(\Omega)$ with $V^* = W^{-1,q}$ and $H = L_2(\Omega)$ and

$$Bu = -\sum \frac{\partial}{\partial x_i}\left(\left|\frac{\partial v}{\partial x_i}\right|^{p-2}\frac{\partial v}{\partial x_i}\right), \qquad p \geq 2$$

$$f \in L_q(I, W^{-1,q}), \qquad \phi_0 \in L_2(\Omega)$$

We can rewrite problem (111) in its abstract form,

$$\begin{aligned} \frac{d\phi}{dt} + B\phi &= f, \qquad t \in I \\ \phi(0) &= \phi_0 \end{aligned}$$

noting that V absorbs the boundary condition.

We conclude this section with one of the most general results involving monotone operators. Basically we include a linear term in model (107) which is more singular than the operator B. For convenience of presentation in the sequel we shall write this model as:

$$\frac{d\phi}{dt} = A(t)\phi + f(t, \phi), \qquad \phi(0) = \phi_0 \tag{112}$$

where f represents $-B + f$ of the previous model (107) and $A(t)$, $t \in I$, is a family of linear operators more singular than f in the sense that it may contain partial differentials of higher order than that in f and hence may be an unbounded operator.

For existence of solutions we use the following assumptions for A and f:

(A1) $\{A(t),\ t \in I\}$ is a family of densely defined linear operators in H with domains $D(A(t)) \subset V$ and range $R(A(t)) \subset V^*$ for $t \in I$.

(A2) $\langle A(t)e, e\rangle_{V^*,V} \leq 0$ for all $e \in D(A(t))$.

(F1) The function $t \to \langle f(t, e), g\rangle$ is measurable on I for $e, g \in V$, and $f: I \times V \to V^*$ is demicontinuous in the sense that for each $e \in V$

$$\langle f(t_n, \xi_n), e\rangle_{V^*,V} \to \langle f(t, \xi), e\rangle$$

whenever $t_n \to t$ and $\xi_n \to \xi$ in V.

(F2) $\langle f(t, \xi) - f(t, \eta), \xi - \eta\rangle_{V^*,V} \leq 0$ for all $\xi, \eta \in V$.

(F3) There exists an $h \in L_q(I, R_+)$, $R_+ = [0, \infty)$, and $\alpha \geq 0$ such that

$$|f(t, \xi)|_{V^*} \leq h(t) + \alpha(\|\xi\|_V)^{p/q} \qquad \text{a.e.}$$

for each $\xi \in V$.

(F4) There exists an $h_1 \in L_1(I,\ R)$ and $\beta > 0$ such that

$$\langle f(t, \xi), \xi\rangle_{V^*,V} \leq h_1(t) - \beta(\|\xi\|_V)^p \qquad \text{a.e.}$$

for each $\xi \in V$.

Theorem 19. Consider system (112) and suppose assumptions (A1) to (A2) and (F1) to (F4) hold. Then, for each $\phi_0 \in H$, Eq. (112) has a unique (weak) solution $\phi \in L_p(I,\ V) \cap C(\bar{I},\ H)$ and further the solution ϕ is an absolutely continuous V^*-valued function.

The general result given above also applies to partial differential equations of the form:

$$\begin{aligned} \frac{\partial \phi}{\partial t} &+ \sum_{|\alpha| \leq m+1} (-1)^{|\alpha|} D^\alpha \left(a_{\alpha\beta}(t, x) D^\beta \phi\right) \\ &+ \sum_{|\alpha| \leq m} (-1)^{|\alpha|} D^\alpha F_\alpha(t, x; \phi, D^1\phi, \ldots, D^m\phi) \\ &= 0 \qquad \text{on} \quad I \times \Omega \\ \phi(0, x) &= \phi_0(x) \qquad \text{on} \quad \Omega \\ (D^\alpha \phi)(t, x) &= 0, \qquad 0 \leq |\alpha| \leq m \\ &\qquad \text{on} \quad I \times \partial\Omega \end{aligned} \tag{113}$$

where the operator $A(t)$ in Eq. (112) comes from the linear part and the boundary conditions, and the nonlinear operator f comes from the nonlinear part in Eq. (113). The space V can be chosen as $W_0^{m,p}$, $p \geq 2$, with $V^* \equiv W^{-m,q}$ where $1/p + 1/q = 1$.

REMARK 2. Again, if both A and f are time invariant, it follows from Theorem 19 that there exists a nonlinear semigroup $S(t)$, $t \geq 0$, such that $\phi(t) = S(t)\phi_0$, $\phi_0 \in H$.

In certain situations the function $\{F_\alpha, |\alpha| \leq m\}$ may be discontinuous in the variables $\{\phi, D^1\phi, \ldots, D^m\phi\}$ and as a result the operator f, arising from the corresponding Dirichlet form, may be considered to be a multivalued function. In other words $f(t, \xi)$, for $t, \xi \in I \times V$, is a nonempty subset of V^*. In that case, the differential equation becomes a differential inclusion $\dot{\phi} \in A(t)\phi + f(t, \phi)$. Differential inclusions may also arise from variational evolution inequalities. Define the operator S by:

$$S_t g = U(t, 0)\phi_0 + \int_0^t U(t, \theta) g(\theta)\, d\theta$$

$$t \in I, \qquad g \in L_1(I, V^*) \tag{114}$$

where U is the transition operator corresponding to the generator A and ϕ_0 is the initial state. Then, one questions the existence of a $g \in L_1(I, V^*)$ such that $g(t) \in f(t, S_t g)$ a.e. If such a g exists, then one has proved the existence of a solution of the initial value problem:

$$\begin{aligned} \dot{\phi}(t) &\in A(t)\phi(t) + f(t, \phi(t)) \\ \phi(0) &= \phi_0 \end{aligned} \tag{115}$$

These questions have been considered in control problems.

B. Stability, Identification, and Controllability

We present here some simple results on stability and some comments on the remaining topics.

We consider the semilinear system,

$$\frac{d\phi}{dt} = A\phi + f(\phi) \tag{116}$$

in a Hilbert space $(H, \|\cdot\|)$ and assume that f is weakly nonlinear in the sense that (a) $f(0) = 0$, and (b) $\|f(\xi)\| = o(\|\xi\|)$, where

$$\lim_{\|\xi\| \to 0} \left\{ \frac{o(\|\xi\|)}{\|\xi\|} \right\} = 0 \tag{117}$$

Theorem 20. If the linear system $\dot{\phi} = A\phi$ is asymptotically stable in the Lyapunov sense and f is a weakly nonlinear continuous map from H to H, then the nonlinear system (116) is locally asymptotically stable near the zero state.

The proof is based on Theorem 4.

For finite-dimensional systems, Lyapunov stability theory is most popular in that stability or instability of a system is characterized by a scalar-valued function known as the Lyapunov function. For ∞-dimensional systems a straight forward extension is possible only if strong solutions exist.

Consider the evolution equation (116) in a Hilbert space H, and suppose that A is the generator of a strongly continuous semigroup in H and f is a continuous map in H bounded on bounded sets. We assume that Eq. (116) has strong solutions in the sense that $\dot{\phi}(t) = A\phi(t) + f(\phi(t))$ holds a.e., and $\phi(t) \in D(A)$ whenever $\phi_0 \in D(A)$. Let T_t, $t \geq 0$, denote the corresponding nonlinear semigroup in H so that $\phi(t) = T_t(\phi_0)$, $t \geq 0$. Without loss of generality we may consider $f(0) = 0$ (if necessary after proper translation in H) and study the question of stability of the zero state. Let Ω be a nonempty open connected set in H containing the origin and define $\Omega_D \equiv \Omega \cap D(A)$, and $B_a(D) \equiv \{\xi \in H : |\xi|_H < a\} \cap D(A)$ for each $a > 0$. The system is said to be stable in the region Ω_D if, for each ball $B_R(D) \subset \Omega_D$, there exists a ball $B_r(D) \subset B_R(D)$ such that $T_t(\phi_0) \in B_R(D)$ for all $t \geq 0$ whenever $\phi_0 \in B_r(D)$. The zero state is said to be asymptotically stable if $\lim_{t \to \infty} T_t(\phi_0) = 0$ whenever $\phi_0 \in \Omega_D$.

A function $V : \Omega_D \to [0, \infty]$ is said to be positive definite if it satisfies the properties:

(a) $V(x) > 0$ for $x \in \Omega_D \backslash \{0\}$, $V(0) = 0$.

(b) V is continuous on Ω_D and bounded on bounded sets.

(c) V is Gateaux differentiable on Ω_D in the direction of H, in the sense that, for each $x \in \Omega_D$ and $h \in H$,

$$\lim_{\varepsilon \to 0} \left\{ \frac{V(x + \varepsilon h) - V(x)}{\varepsilon} \right\} \equiv V'(x, h)$$

exists, and for each $h \in H$, $x \to V'(x, h)$ is continuous.

The following result is the ∞-dimensional analog of the classical Lyapunov stability theory.

Theorem 21. Suppose the system $\dot{\phi} = A\phi + f(\phi)$ has strong solutions for each $\phi_0 \in D(A)$ and there exists a positive definite function V on Ω_D such that along any trajectory $\phi(t)$, $t \geq 0$, starting from $\phi_0 \in \Omega_D$,

$$V'(\phi(t), A\phi(t) + f(\phi(t))) \leq 0 \qquad (<0) \tag{118}$$

for all $t \geq 0$. Then the system is stable (asymptotically stable) in the region Ω_D.

If the system admits only mild solutions, Theorem 21 must be modified by using positive definite functions which have Gateaux derivatives in the directions $\{h\}$ in spaces larger than H.

We conclude this section with a result for systems governed by monotone nonlinear operators as in Eqs. (107) and (112).

Theorem 22. Consider system (112) with the operators A and f satisfying the assumptions of Theorem 19 for all $t \geq 0$, and suppose $h_1 \in L_1(0, \infty; R)$ and the injection $V \subset H$ is continuous. Then, the system is globally asymptotically stable with respect to the origin in H.

The questions of identification of parameters appearing in any of the system equations treated above can be dealt with in a similar way as in the linear case. In fact, an identification problem may be considered as a special case of a control problem with controls appearing in the system coefficients. Such classes of problems have been covered well in the literature. However, controllability questions for the general systems are more difficult and almost nothing is known.

C. Existence of Optimal Controls

Existence of optimal controls for strongly nonlinear parabolic systems and more general nonlinear evolution

equations of the form (86) have been treated in the literature. The technical details are rather involved and long. We shall limit ourselves to a brief summary of some results.

Consider system (107) with controls denoted by u:

$$\frac{d\phi}{dt} + B(t)\phi = f(t, u) \tag{119}$$

where the operator B is nonlinear and may be given by the expression:

$$B(t)\phi = \sum_{|\alpha|\le m} (-1)^{|\alpha|} D^{\alpha} F_{\alpha}(t, x; \phi, D^1\phi, \ldots, D^m\phi), \quad t, x \in I \times \Omega \tag{120}$$

Under quite general assumptions on the function F_α, the operator B has properties (B1) to (B3) of Theorem 18. For the space V one may choose $W_0^{m,p}$, $p \ge 2$, or any closed subspace of $W^{m,p}$ so that $W_0^{m,p} \subset V \subset W^{m,p}$. Here V is a reflexive Banach space. For admissible controls we choose any reflexive Banach space E of functions defined on Ω and Γ, a closed bounded convex subset of E, and consider $\mathcal{U}$ to be the class of admissible controls which are strongly measurable functions defined on $I = (0, T)$, with values in Γ. Let $f\colon I \times \Gamma \to V^*$, so that for each t, $f(t, \cdot)$ is weakly continuous (or more generally demicontinuous) on Γ; for each $v \in \Gamma$, $f(\cdot, v)$ is measurable on I (or continuous), and for each $u \in \mathcal{U}$, $f(u) \in L_q(I, V^*)$ where $f(u)(t) \equiv f(t, u(t))$.

The system,

$$\begin{aligned} \frac{d\phi}{dt} + B(t)\phi &= f(t, u(t)), \qquad t \in I \\ \phi(0) &= \phi_0, \qquad\quad u \in \mathcal{U} \end{aligned} \tag{121}$$

is written in its weak form:

$$\begin{aligned} &(L\phi, \psi) + b(\phi, \psi) = (f(u), \psi) \\ &\quad \text{for all} \quad \psi \in L_p(I, V) \cap C(\bar{I}, H) \\ &\qquad \phi(0) = \phi_0, \qquad u \in \mathcal{U} \end{aligned} \tag{122}$$

where L denotes the extension of d/dt as an operator from the space $L_p(I, V)$ to the space $L_q(I, V^*)$ and b is the Dirichlet form given by:

$$b(\phi, \psi) \equiv \int_I \sum_{|\alpha|\le m} \int_\Omega F_\alpha(t, x, \phi, D^1\phi, \ldots, D^m\phi) \cdot D^\alpha \psi \, dx \, dt \tag{123}$$

We consider the following control problems:

(P1) *Terminal control.* Let $J(u) = Z(\phi(T))$, where Z is a real-valued function on H and the pair $\{u, \phi\}$ is subject to the dynamic constraint (122). The problem is to find a control $u \in \mathcal{U}$ that minimizes the functional J.

(P2) *Time-optimal control.* Let M, a subset of H, be the target set. The requirement is to find a control $u \in \mathcal{U}$ that transfers the system from the state $\phi_0 \in H$ to the target set M in minimum time.

The existence of optimal controls depends on the properties of admissible trajectories, attainable sets, and the cost functionals. Let $\mathcal{X}$ denote the set of admissible trajectories, that is, the set of all $\{\phi\} \in L_p(I, V) \cap C(\bar{I}, H)$ such that ϕ is a solution of Eq. (122) corresponding to some control $u \in \mathcal{U}$. Similarly, the attainable set may be defined as:

$$\mathcal{A}(t) \equiv \{\xi \in H : \xi = \phi(t) \quad \text{for some} \quad \phi \in \mathcal{X}\}$$

Under a number of technical assumptions on the operators B and f and the control set Γ one can prove the following result.

Theorem 23. (a) The set of admissible trajectories $\mathcal{X}$ is a weakly closed and weakly sequentially compact subset of $L_p(I, V)$. (b) For each $t \in [0, T]$, the attainable set $\mathcal{A}(t)$, is a weakly compact subset of H.

Using the preceding result, one can prove the existence of optimal controls for problems (P1) and (P2).

Theorem 24. Let Z be a weakly lower semicontinuous functional defined on H and bounded from below. Then there exists an optimal control solving problem (P1).

Theorem 25 (P2)**.** Suppose the given target set M is a weakly closed subset of H and the system is controllable in the sense that there exists an admissible u, and $\tau \in \bar{I}$, such that $\phi(u)(\tau) \in M$. Then there exists an optimal control that steers the systems from state ϕ_0 to the target set M in minimum time.

Optimal control problems for the more general system (112) recently have been studied in several papers giving existence results for measurable controls and measure-valued controls. Systems governed by differential inclusions of the form (115) and their associated control problems also have been studied recently. The technical details are too long for presentation here. Interested readers may consult the Bibliography.

D. Necessary Conditions of Optimality

For completeness we shall present a result on the necessary conditions of optimality. Consider system (122) along with the cost functional given by:

$$J(u) = Z(\phi(T) + \int_0^T f^0(t, \phi(t), u(t)\, dt \tag{124}$$

The problem is to find a control $u^0 \in \mathcal{U}$ that minimizes the functional J subject to constraint (122). Let $\{F_\alpha(t, x, \xi)\}$,

$t,\ x \in I \times \Omega, \xi = \{\xi_\alpha, |\alpha| \le m\} \in R^N$, denote the functions defining the operator B (see Eqs. (119) and (120)) and $\{F_\alpha^\beta, |\beta| \le m\}$ their directional derivatives with respect to $\xi \in R^N$. We assume that for fixed $t, x \in I \times \Omega$, the functions $\xi \to F_\alpha^\beta(t, x, \xi)$ are continuous on R^N for all α, β satisfying $|\alpha|, |\beta| \le m$, and, for fixed $\xi \in R^N, (t, x) \to F_\alpha^\beta(t, x, \xi)$ are measurable on $I \times \Omega$. For any fixed $\phi \in L_p(I, V)$, the bilinear form,

$$b_\phi(\psi, \nu) \equiv \sum_{|\alpha|,|\beta| \le m} \int_I \langle D^\alpha \psi, F_\beta^\alpha(t, x; \phi, D^1\phi, \ldots, D^m\phi) D^\beta \nu \rangle_\Omega \, dt \tag{125}$$

is well defined on $L_p(I,\ V) \times L_p(I,\ V)$. Let

$$\begin{aligned} f_1^0 &: I \times V \times E \to V^* \\ f_2^0 &: I \times V \times E \to E^* \end{aligned} \tag{126}$$

denote the linear Gateaux differentials of f^0 with respect to the state and control variables, respectively, and let

$$F: I \times E \to \mathcal{L}(E, V^*) \tag{127}$$

denote the linear Gateaux differential of f with respect to the control variable.

Under a number of technical assumptions on the functions $\{F_\alpha, |\alpha| \le m\}, f, f^0$, and Z and the control constraint set $\Gamma \subset E$, one can prove the folowing necessary conditions of optimality.

Theorem 26. Consider system (122) with the cost functional (124). For the pair $\{u^0, \phi^0\} \in \mathcal{U} \times \mathcal{X}$ to be optimal it is necessary that there exists a $\psi^0 \in L_p(I, V) \cap C(\bar{I}, H)$ so that the triple $\{u^0, \psi^0, \phi^0\}$ satisfy the conditions:

(a) $(L\phi^0, \nu) + b(\phi^0, \nu) = (f(u^0), \nu), \qquad \phi^0(0) = \phi_0$

for all

$$\nu \in \mathcal{F}_1 \equiv \{\nu \in L_p(I, V) \cap C(\bar{I}, H): \nu(T) = 0\}$$

(b) $$-(L\psi^0, \nu) + b_{\phi^0}(\psi^0, \nu) + \left(f_1^0, \nu\right) = 0, \qquad \psi^0(T) = -Z'(\phi^0(T))$$

for all

$$\nu \in \mathcal{F}_0 \equiv \{\nu \in L_p(I, V) \cap C(\bar{I}, H): \nu(0) = 0\}$$

(c) $$\int_I \langle f_2^0(t, \phi^0(t), u^0(t)) - F^*(t, u^0(t))\psi^0(t),\ w(t) - u^0(t) \rangle_{E^*,E} \, dt \ge 0 \tag{128}$$

for all $w \in \mathcal{U}$, where F^* denotes the dual of the operator $F(t,\ u^0(t)) \in \mathcal{L}(E,\ V^*)$ and Z' the Gateaux derivative of Z on H.

For time-optimal controls, similar necessary conditions exist. In this case, the optimal control is also characterized by inequality (c), with the exceptions that $f_2^0 \equiv 0$ and the upper limit of the integral is the optimal time t^0 instead of T.

A number of interesting observations can be made from the above result. For example, if $f(t,\ u) = T(t)u$ and $f^0(t, \phi,\ u) = \tilde{f}^0(t,\ \phi) + \langle Nu, u \rangle_{E^*,E}$ and the control set $\Gamma = E$, then it follows from inequality (c) that

$$(N + N^*)u^0(t) = T^*(t)\psi^0(t) \tag{129}$$

Hence, if E is reflexive and $N = N^*$ and N is invertible, then $u^0(t) = \frac{1}{2}N^{-1}T^*(t)\psi^0(t)$. This is precisely the form of optimal controls for linear systems with quadratic cost functionals.

E. Nonlinear Stochastic Evolution Equations

We present here a brief account of nonlinear stochastic systems. The simplest nonlinear stochastic evolution equation may be given by:

$$dy = Ay\,dt + f(y)\,dW(t), \qquad t \ge 0, \qquad y(0) = y_0 \tag{130}$$

where W is the Wiener process with covariance operator $Q \in \mathcal{L}^+(H)$ as in the linear case. A is the generator of a c_0-semigroup $S(t)$, $t \ge 0$, on H, and $f: H \to \mathcal{L}(H)$ satisfying:

$$\begin{aligned} &\text{(a)} \qquad \|f(x)\|_{\mathcal{L}(H)}^2 \le K^2\left(1 + |x|_H^2\right) \\ &\text{(b)} \qquad \|f(x) - f(y)\|_{\mathcal{L}(H)}^2 \le K^2|x - y|_H^2 \end{aligned} \tag{131}$$

Define the nonlinear operator G by:

$$(Gx)(t) \equiv S(t)y_0 + \int_0^t S(t - \theta)f(x(\theta))\,dW(\theta) \tag{132}$$

for $x \in X \equiv C(I, L_2(\Omega, H))$, where X is a Banach space with respect to the topology given by:

$$\|x\|_X \equiv \sqrt{\sup\left\{E|x(t)|_H^2, t \in I\right\}}$$

Under the above assumptions one can prove the existence of an integer n, such that the nth iteration of G, denoted by G^n, is a contraction in X. Hence, by the Banach fixed-point theorem, there exists $y \in X$ such that $y = Gy$. In other words, the integral equation in X,

$$y(t) = S(t)y_0 + \int_0^t S(t - \theta)f(y(\theta))\,dW(\theta) \qquad t \in [0, T]$$

has a unique solution $y \in X$ whenever $y_0 \in L_2(\Sigma,\ H)$, hence y is a mild solution of system (130).

Existence theory for semilinear stochastic evolution equations of the form,

$$dy = (A(t)y + f(t, y))\,dt + \sigma(t)\,dW \qquad (133)$$

has been developed under much weaker hypotheses on the operators A and f using the Leray–Schauder degree theory. There are also results in which A has been considered to be a strongly measurable function from $(\Sigma, \mathcal{A}, P)$ to $\mathcal{L}_c(H)$, the space of closed densely defined linear (not necessarily bounded) operators in H. In this case, A is assumed to generate a strongly measurable (random) c_0-semigroup in H. Existence theory for nonlinear stochastic boundary value problems of the form,

$$\begin{aligned} \frac{\partial \phi}{\partial t} + A(t)\phi &= f + F(\phi) \qquad \text{on} \quad I \times \Omega \\ B\phi &= g + G(\phi) \qquad \text{on} \quad I \times \partial\Omega \\ \phi(0) &= \phi_0 \qquad \text{on} \quad \Omega \end{aligned} \qquad (134)$$

has been considered with f, g being generalized random processes, ϕ_0 a generalized random variable, and F, G nonlinear accretive operators. Stability problems for systems of the form (134) with $f = 0$; $g = N$, a generalized white noise; $G = 0$; and F being a monotone operator have been studied. It has been shown that the system is asymptotically stable with respect to a ball around the origin with radius determined by the trace of the covariance operator of the associated Wiener process.

It appears from the literature that for nonlinear systems the theory of optimal control, filtering, identification, and controllability is far from satisfactory. This is a difficult but fascinating field and certainly a challenging subject of the future.

IV. RECENT ADVANCES IN INFINITE-DIMENSIONAL SYSTEMS AND CONTROL

In this section we discuss some recent advances in the theory and applications of distributed parameter systems since the time of first publication of this encyclopedia. Details of these new developments can be found in the references. There has been substantial theoretical development of distributed parameter systems, as indicated in References 11 to 36. These include general boundary control problems, control of deterministic and stochastic evolution inequalities and differential inclusions, uncertain systems, and the so-called B-evolutions. Due to space limitations, we cannot include the details. Here we shall give only a brief outline of the major new concepts introduced in recent years. On the theoretical front, there are three new major developments:

1. The first one is in the area of control theory for systems governed by m-times integrated semigroups or distribution semi groups.
2. The second one is in the area of fundamental concepts, in particular, the notion of solution, a new notion of solutions, called *measure solutions*, has been introduced very recently and has been used in the theory of control of distributed parameter systems.
3. The third front extends the concept of impulsive systems to infinite dimensional Banach spaces; we shall discuss briefly these new developments and their implication in mathematical sciences.
4. The fourth front represents recent applications of the theory of distributed parameter systems to the physical sciences.

A. *m*-Times Integrated Semigroups

The classical semigroup theory, as seen in Section II, is based on the assumptions that A is closed and $D(A)$ is dense in X and that the Hille–Yosida inequality,

$$\|R(\lambda, A)\| \equiv \|(\lambda I - A)^{-1}\| \le M/(\lambda - \omega), \qquad \lambda \in \rho(A) \supset (\omega, \infty) \qquad (135)$$

holds for some $M \ge 0$ and $\omega \in R$. These are the necessary and sufficient conditions for the existence of a C_0-semigroup $S(t)$, $t \ge 0$, and hence the existence of a solution of the Cauchy problem,

$$\dot{x} = Ax, \qquad x(0) = \xi \qquad (136)$$

in X. The solution is given by $x(t) = S(t)\xi$, $t \ge 0$. No doubt this covers a very large class of partial differential operators with given boundary conditions and hence a large class of distributed parameter systems. However, there are classes of operators A which do not satisfy the Hille–Yosida theorem, yet such systems have solutions in some generalized sense. According to the Hille–Yosida theorem, $R(\lambda, A)$ is the Laplace transform of some operator-valued function $S(t)$, $t \ge 0$. It is now known that the Cauchy problem stated above has a solution in some generalized sense even if only $R_m(\lambda, A) \equiv R(\lambda, A)/\lambda^m$, $\lambda \in \rho(A)$, is the Laplace transform of an operator-valued function $T(t)$, $t \ge 0$. In this case, $T(t)$, $t \ge 0$, is said to be the m-times integrated semigroup and A is said to be its infinitesimal generator. The classical solution for the Cauchy problem as stated above is now given by:

$$x(t) = (d^m/dt^m)T(t)\xi, t \ge 0, \qquad \text{for} \quad \xi \in D(A^{m+1})$$

In general, if $\xi \in X$ there is no classical solution but we may admit generalized derivatives and hence generalized solutions. For example, if $D(A^{m+1})$ is dense in X, one can choose a sequence $\{\xi_k\}$ converging strongly to ξ and

consider the entity x as the generalized solution if it satisfies the identity:

$$\int_0^\tau \langle x(t), \phi(t)\rangle_{X,X^*}\, dt = (-1)^m \lim_{k\to\infty} \int_0^\tau \left\langle T(t)\xi_k, D^m\phi(t)\right\rangle_{X,X^*} dt \quad (137)$$

for all ϕ in a class of test functions. A suitable class of test functions for this problem is the Sobolev class $W_0^{m,1}(I, X^*)$ which consists of X^*-valued functions whose derivatives up to order $m-1$ vanish on the boundary of the set $I \equiv (0, \tau)$ and belong to $L_1(I, X^*)$. By duality, the solution $x \in W^{-m,\infty}(I, X)$. For further details on m-times integrated semigroups, generalized solutions, stochastic systems, and optimal controls of systems involving operators that generate such semigroups, the reader may consult References 13, 21, and 24.

B. Measure Solution

Consider the system,

$$\dot{x} = f(x), \qquad x(0) = \xi \in E \quad (138)$$

It is well known that if $E = R^n$ is a finite-dimensional space and f is merely continuous, the system has at least one local solution in the sense that there exists a maximal interval of time $(0, \tau_m)$ and an absolutely continuous function $x^*(t)$, $t \in (0, \tau_m)$, that satisfies the differential equation along with the initial condition $x^*(0) = \xi$, with the possibility of blow-up at time τ_m. That is,

$$\lim_{t\to\tau_m} \|x^*(t)\| = \infty \quad (139)$$

An elementary example is $\dot{y} = y^2$, $y(0) = \gamma$. If $\gamma > 0$, the blow-up time is $\tau_m = (1/\gamma)$.

In contrast, if E is an infinite-dimensional Banach space, mere continuity of $f : E \to E$ is no more sufficient to guarantee even a local solution. Even continuity and boundedness of this map do not guarantee existence. Compactness is necessary. However, under mere continuity and local boundedness assmptions, one can prove the existence of generalized (to be defined shortly) solutions. This is what prompted the concept and development of measure-valued solutions. Consider the semilinear system,

$$\dot{x} = Ax + f(x), t \geq 0, \qquad x(0) = x_0 \quad (140)$$

and suppose A is the infinitesimal generator of a C_0-semigroup in E. Let $BC(E)$ denote the Banach space of bounded continuous functions on E with the standard topology induced by the sup norm $\|\phi\| \equiv \sup\{|\phi\ (\xi)|, \xi \in E\}$. The dual of this space is the space of regular, bounded, finitely additive measures denoted by $\sum_{rba}(E)$ and defined on the algebra of sets generated by closed subsets of E. This is a Banach space with respect to the topology induced by the total variation norm. Let $\Pi_{rba}(E) \subset \sum_{rba}(E)$ denote the space of regular, finitely additive probability measures. For a $\nu \in \sum_{rba}(E)$ and $\phi \in BC(E)$, the pairing

$$(\nu, \phi) \equiv \nu(\phi) \equiv \int_E \phi(\xi)\nu(d\xi)$$

is well defined. Letting $D\phi$ denote the Frechet derivative of $\phi \in BC(E)$, we introduce the class,

$$\mathcal{F} \equiv \{\phi \in BC(E) : D\phi \text{ continuous with } \phi,\ D\phi \text{ having bounded supports}\}$$

for test functions. Define the operator $\mathcal{A}$ with domain:

$$D(\mathcal{A}) \equiv \{\phi \in F : \mathcal{A}\phi \in BC(E^+)\}$$

where for $\phi \in D(\mathcal{A})$,

$$\mathcal{A}\phi \equiv \langle A^* D\phi(\xi), \xi\rangle_{E^*,E} + \langle D\phi(\xi),\ f(\xi)\rangle_{E^*,E} \quad (141)$$

and E^+ is a suitable compactification of E that makes E^+ a compact Hausdorff space containing E as a dense subspace. The new notion of a solution for the Cauchy problem (140) can be stated as follows.

Definition: A measure function μ_t, $t \geq 0$, with values in $\Pi_{rba}(E)$, is said to be a generalized solution of the semilinear evolution equation if, for each $\phi \in D(\mathcal{A})$, the following identity holds:

$$\mu_t(\phi) = \phi(x_0) + \int_0^t \mu_s(\mathcal{A}\phi)\, ds, \qquad t \geq 0 \quad (142)$$

The concepts of measure solutions and stochastic evolution equations have been extended. For example, consider the infinite-dimensional stochastic systems on a Hilbert space H, governed by Eq. (130) or, more generally, the equation:

$$dx = Ax\, dt + f(x)\, dt + \sigma(x)\, dW, \qquad x(0) = x_0 \quad (143)$$

Again, if f and ω are merely continuous and bounded on bounded sets, all familiar notions of solutions (strong, mild, weak, martingale) fail. However, measure solutions are well defined. In this case, expression (142) must be modified as follows:

$$\mu_t(\phi) = \phi(x_0) + \int_0^t \mu_s(\mathcal{A}\phi)\, ds + \int_0^t \langle \mu_s(\mathcal{B}\phi), dW(s)\rangle, \qquad t \geq 0 \quad (144)$$

where now the operator $\mathcal{A}$ is given by the second-order partial differential operator,

$$\mathcal{A}\phi \equiv (1/2)Tr(D^2\phi\sigma\sigma^*) + \langle A^*D\phi(\xi), \xi\rangle_H + \langle D\phi(\xi), f(\xi)\rangle_H \tag{145}$$

and the operator $\mathcal{B}$ is given by:

$$\mathcal{B}\phi \equiv \sigma^* D\phi$$

The last term is a stochastic integral with respect to a cylindrical Brownian motion (for details, see Reference 31). The operators $\mathcal{A}$ and $\mathcal{B}$ are well defined on a class of test functions given by:

$$\mathcal{F} \equiv \{\phi \in BC(E) : \phi, D\phi, D^2\phi \text{ continuous having bounded supports and } Tr(D^2\phi\sigma\sigma^*(\xi)) < \infty, \xi \in H\} \tag{146}$$

A detailed account of measure solutions and their optimal control for semilinear and quasilinear evolution equations can be found in References 20, 22, 28, and 30–32.

C. Impulsive Systems

In many physical problems, a system may be subjected to a combination of continuous as well as impulsive forces at discrete points of time. For example, in building construction, piling is done by dropping massive weights on the top of vertically placed steel bars. This is an impulsive input. In the management of stores, inventory is controlled by an agreement from the supplier to supply depleted goods. An example from physics is a system of particles in motion which experience collisions from time to time, thereby causing instantaneous changes in directions of motion. In the treatment of patients, medications are administered at discrete points of time, which could be considered as impulsive inputs to a distributed physiological system.

The theory of finite dimensional impulsive systems is well known.[12] Only in recent years has the study of infinite dimensional impulsive systems been initiated.[27,33–36] Here, we present a brief outline of these recent advances. Let $I \equiv [0, T]$ be a closed bounded interval of the real line and define the set $D \equiv \{t_1, t_2, \cdot, \cdot, t_n\} \in (0, T)$. A semilinear impulsive system can be described by the following system of equations:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + f(x(t)), \\ t &\in I\backslash D, \qquad x(0) = x_0, \\ \Delta x(t_i) &= F_i(x(t_i)), \\ 0 = t_0 < t_1 < t_2, \cdots < t_n < t_{n+1} &= T \end{aligned} \tag{147}$$

where, generally, A is the infinitesimal generator of a C_0-semigroup in a Banach space E, the function f is a continuous nonlinear map from E to E, and $F_i : E \to E$, $i = 1, 2, \cdot, \cdot, n$, are continuous maps. The difference operator $\Delta_x(t_i) \equiv x(t_i + 0) - x(t_i - 0) \equiv x(t_i + 0) - x(t_i)$ denotes the jump operator. This represents the jump in the state x at time t_i with F_i determining the jump size at time t_i. Similarly, a controlled impulsive system is governed by the following system of equation:

$$\begin{aligned} dx(t) &= [Ax(t) + f(x(t))]\,dt + g(x(t))\,dv(t), \\ t &\in I\backslash D, x(0) = x_0, \\ \Delta x(t_i) &= F_i(x(t_i)), \\ 0 = t_0 < t_1 < t_2, \cdots < t_n < t_{n+1} &= T \end{aligned} \tag{148}$$

where $g : E \to \mathcal{L}(F, E)$ and the control $v \in BV(I, F)$. For maximum generality, one can choose the Banach space $BV(I, F)$ of functions of bounded variation on I with values in another Banach space F as the space of admissible controls. This class allows continuous as well as jump controls. For each $r > 0$, let $B_r \equiv \{\xi \in E : \|\xi\| \le r\}$ denote the ball of radius r around the origin. Let $PWC_\ell(I, E)$ denote the Banach space of piecewise continuous functions on I taking values from E, with each member being left continuous, having right-hand limits. For solutions and their regularity properties we have the following.

Theorem 27. Suppose the following assumptions hold:

(A1) A is the infinitesimal generator of a C_0-semigroup in E.

(A2) The maps $g, F_i, i = 1, 2, \cdot, \cdot, n$ are continuous and bounded on bounded subsets of E with values in $\mathcal{L}(F, E)$ and E, respectively,

(A3) The map f is locally Lipschitz having at most linear growth; that is, there exist constants $K > 0, K_r > 0$, such that:

$$\|f(x) - f(y)\|_E \le K_r\|x - y\|_E, \qquad x, y \in B_r$$
$$\text{and} \quad \|f(x)\| \le K(1 + \|x\|)$$

Then, for every $x_0 \in E$ and $v \in BV(I, F)$, system (148) has a unique mild solution, $x \in PWC_l(I, E)$.

Using this result one can construct necessary conditions of optimality. We present here one such result of the author.[35] Consider $\mathcal{U} \in BV(I, F)$ to be the class of controls comprised of pure jumps at a set of arbitrary but prespecified points of time $J \equiv \{0 = t_0 = s_0 < s_1 < s_2, \cdot, \cdot, < s_{m-1} < s_m, m \ge n\} \subset [0, T)$. Clearly $\mathcal{U}$ is isometrically isomorphic to the product space $\mathcal{F} \equiv \prod_{k=1}^{m+1} F$, furnished with the product topology. We choose a closed convex subset $\mathcal{U}_{ad} \subset \mathcal{U}$ to be the class of admissible controls. The

basic control problem is to find a control policy that imparts a minimum to the following cost functional,

$$J(v) = J(q) \equiv \int_0^T \ell(x(t), v(t))\, dt + \varphi(v) + \Phi(x(T))$$
$$\equiv \int_0^T \ell(x(t), q)\, dt + \varphi(q) + \Phi(x(T))$$

Theorem 28. Suppose assumptions (A1) to (A3) hold, with $\{f, F_i, g\}$ all having Frechet derivatives continuous and bounded on bounded sets, and the functions ℓ, φ, Φ are once continuously Gateaux differentiable on $E \times F, \mathcal{F}, E$, respectively. Then, if the pair $\{v^o(orq^o), x^o\}$ is optimal, there exists a $\psi \in PWC_r(I, E^*)$ so that the triple $\{v^0, x^0, \psi\}$ satisfies the following inequality and evolution equation:

(a) $$\sum_{i=0}^{m} \left\langle \left\{ g^*\big(x^o(s_i)\big)\psi(s_i) + \varphi_{q_i}(q^o) + \int_{s_i}^{T} \ell_u(x^o(t), v^o(t))\, dt \right\}, q_i - q_i^o \right\rangle_{F^*,F} \geq 0$$

for all $q = \{q_0, q_1, \cdot, \cdot, q_m\} \in \mathcal{U}_{ad}$.

(b) $$d\psi = -\big(A^*\psi + f_x^*(x^o(t))\big)\psi - \ell_x(x^o(t), v^o(t))\, dt - g_x^*(x^o(t), \psi(t))\, dv^o, \qquad t \in I \backslash D$$
$$\psi(T) = \Phi_x(x^o(T))$$
$$\Delta_r \psi(t_i) = -F_{i,x}^*\big(x^o(t_i)\big)\psi(t_i), \qquad i = 1, 2, \cdot, \cdot, n$$

where the operator Δ_r is defined by $\Delta_r f(t_i) \equiv f(t_i - 0) - f(t_i)$.

(c) The process x^o satisfies the system equation (148) (in the mild sense) corresponding to the control v^o.

In case a hard constraint is imposed on the final state requiring $x^o(T) \in \mathcal{K}$ where $\mathcal{K}$ is a closed convex subset of E with nonempty interior, the adjoint equation given by (b) requires modification. The terminal equality condition is replaced by the inclusion $\psi(T) \in \partial I_{\mathcal{K}}(x^o(T))$, where $I_{\mathcal{K}}$ is the indicator function of the set $\mathcal{K}$.

Recently a very general model for evolution inclusions has been introduced:[36]

$$\begin{aligned} \dot{x}(t) &\in Ax(t) + F(x(t)), \\ &t \in I \backslash D, \qquad x(0) = x_0, \\ \Delta x(t_i) &\in G_i(x(t_i)), \\ 0 = t_0 < t_1 < t_2, \cdots < t_n < t_{n+1} &\equiv T \end{aligned} \tag{2.3}$$

Here, both F and $\{G_i,\ i = 1, 2, \cdot, \cdot, n\}$ are multivalued maps. This model may arise under many different situations. For example, in case of a control problem where one wishes to control the jump sizes in order to achieve certain objectives, one has the model,

$$\begin{aligned} \dot{x}(t) &\in Ax(t) + F(x(t)), \\ &t \in I \backslash D, \qquad x(0) = x_0, \\ \Delta x(t_i) &= g_i(x(t_i), u_i)), \\ 0 = t_0 < t_1 < t_2, \cdots < t_n < t_{n+1} &\equiv T \end{aligned}$$

where the controls u_i may take values from a compact metric space U. In this case, the multis are given by $G_i(\zeta) = g_i(\zeta, U)$. For more details on impulsive evolution equations and, in general, inclusions, the reader may consult References 35 and 36.

D. Applications

The slow growth of application of distributed systems theory is partly due to its mathematical and computational complexities. In spite of this, in recent years there have been substantial applications of distributed control theory in aerospace engineering, including vibration suppression of aircraft wings, space shuttle orbiters, space stations, flexible artificial satellites, and suspension bridges. Several papers on control of fluid dynamical systems governed by Navier–Stokes equations have appeared over the past decade with particular reference to artificial heart design. During the same period, several papers on the control of quantum mechanical and molecular systems have appeared. Distributed systems theory has also found applications in stochastic control and filtering. With the advancement of computing power, we expect far more applications in the very near future.

SEE ALSO THE FOLLOWING ARTICLES

CONTROLS, LARGE-SCALE SYSTEMS • DIFFERENTIAL EQUATIONS, ORDINARY • DIFFERENTIAL EQUATIONS, PARTIAL • TOPOLOGY, GENERAL • WAVE PHENOMENA

BIBLIOGRAPHY

Ahmed, N. U., and Teo, K. L. (1981). "Optimal Control of Distributed Parameter Systems," North-Holland, Amsterdam.

Ahmed, N. U. (1983). "Properties of relaxed trajectories for a class of nonlinear evolution equations on a Banach space," *SIAM J. Control Optimization* **2**(6), 953–967.

Ahmed, N. U. (1981). "Stochastic control on Hilbert space for linear evolution equations with random operator-valued coefficients," *SIAM J. Control Optimization* **19**(3), 401–403.

Ahmed, N. U. (1985). "Abstract stochastic evolution equations on Banach spaces," *J. Stochastic Anal. Appl.* **3**(4), 397–432.

Ahmed, N. U. (1986). "Existence of optimal controls for a class of systems governed by differential inclusions on a Banach space," *J. Optimization Theory Appl.* **50**(2), 213–237.

Ahmed, N. U. (1988). "Optimization and Identification of Systems Governed by Evolution Equations on Banach Space," Pitman Research Notes in Mathematics Series, Vol. 184, Longman Scientific/Wiley, New York.

Balakrishnan, A. V. (1976). "Applied Functional Analysis," Springer–Verlag, Berlin.

Butkovskiy, A. G. (1969). "Distributed Control Systems," Elsevier, New York.

Curtain, A. F., and Pritchard, A. J. (1978). "Infinite Dimensional Linear Systems Theory," Lecture Notes, Vol. 8, Springer–Verlag, Berlin.

Lions, J. L. (1971). "Optimal Control of Systems Governed by Partial Differential Equations," Springer–Verlag, Berlin.

Barbu, V. (1984). "Optimal Control of Variational Inequalities," Pitman Research Notes in Mathematics Series, Vol. 246, Longman Scientific/Wiley, New York.

Lakshmikantham, V., Bainov, D. D., and Simeonov, P. S. (1989). "Theory of Impulsive Differential Equations," World Scientific, Singapore.

Ahmed, N. U. (1991). "Semigroup Theory with Applications to Systems and Control," Pitman Research Notes in Mathematics Series, Vol. 246, Longman Scientific/Wiley, New York.

Ahmed, N. U. (1992). "Optimal Relaxed Controls for Nonlinear Stochastic Differential Inclusions on Banach Space," Proc. First World Congress of Nonlinear Analysis, Tampa, FL, de Gruyter, Berlin, pp. 1699–1712.

Ahmed, N. U. (1994). "Optimal relaxed controls for nonlinear infinite dimensional stochastic differential inclusions," *Lect. Notes Pure Appl. Math., Optimal Control of Differential Equations* **160,** 1–19.

Ahmed, N. U. (1995). "Optimal control of infinite dimensional systems governed by functional differential inclusions," *Discussiones Mathematicae (Differential Inclusions)* **15,** 75–94.

Ahmed, N. U., and Xiang, X. (1996). "Nonlinear boundary control of semilinear parabolic systems," *SIAM J. Control Optimization* **34**(2), 473–490.

Ahmed, N. U. (1996). "Optimal relaxed controls for infinite-dimensional stochastic systems of Zakai type," *SIAM J. Control Optimization* **34**(5), 1592– 1615.

Ahmed, N. U., and Xiang, X. (1996). "Nonlinear uncertain systems and necessary conditions of optimality," *SIAM J. Control Optimization* **35**(5), 1755–1772.

Ahmed, N. U. (1996). "Existence and uniqueness of measure-valued solutions for Zakai equations," *Publicationes Mathematicae* **49**(3–4), 251–264.

Ahmed, N. U. (1996). "Generalized solutions for linear systems governed by operators beyond Hille–Yosida type," *Publicationes Mathematicae* **48**(1–2), 45–64.

Ahmed, N. U. (1997). "Measure solutions for semilinear evolution equations with polynomial growth and their optimal control," *Discussiones Mathematicae (Differential Inclusions)* **17,** 5–27.

Ahmed, N. U. (1997). "Stochastic B-evolutions on Hilbert spaces," *Nonlinear Anal.* **30**(1), 199–209.

Ahmed, N. U. (1997). "Optimal control for linear systems described by m- times integrated semigroups," *Publicationes Mathematicae* **50**(1–2), 1–13.

Xiang, X., and Ahmed, N. U. (1997). "Necessary conditions of optimality for differential inclusions on Banach space," *Nonlinear Anal.* **30**(8), 5437–5445.

Ahmed, N. U., and Kerbal, S. (1997). "Stochastic systems governed by B-evolutions on Hilbert spaces," *Proc. Roy. Soc. Edinburgh* **127A,** 903–920.

Rogovchenko, Y. V. (1997). "Impulsive evolution systems: main results and new trends," *Dynamics of Continuous, Discrete, and Impulsive Systems* **3**(1), 77–78.

Ahmed, N. U. (1998). "Optimal control of turbulent flow as measure solutions," *IJCFD* **11**, 169–180.

Fattorini, H. O. (1998). "Infinite dimensional optimization and control theory," *In* "Encyclopedia of Mathematics and Its Applications," Cambridge Univ. Press, Cambridge, U.K.

Ahmed, N. U. (1999). "Measure solutions for semilinear systems with unbounded nonlinearities," *Nonlinear Anal.* **35,** 478–503.

Ahmed, N. U. (1999). "Relaxed solutions for stochastic evolution equations on Hilbert space with polynomial nonlinearities," *Publicationes Mathematicae* **54**(1–2), 75–101.

Ahmed, N. U. (1999). "A general result on measure solutions for semilinear evolution equations," *Nonlinear Anal.* **35**.

Liu, J. H. (1999). "Nonlinear impulsive evolution equations: dynamics of continuous, discrete, and impulsive systems," **6,** 77–85.

Ahmed, N. U. (1999). "Measure solutions for impulsive systems in Banach space and their control," *J. Dynamics of Continuous, Discrete, and Impulsive Systems* **6,** 519–535.

Ahmed, N. U. (2000). "Optimal impulse control for impulsive systems in Banach spaces," *Int. J. of Differential Equations and Applications* **1**(1).

Ahmed, N. U. (2000). "Systems governed by impulsive differential inclusions on Hilbert space," *J. Nonlinear Analysis.*



Dynamic Programming

Martin L. Puterman
The University of British Columbia

I. Sequential Decision Problems
II. Finite-Horizon Dynamic Programming
III. Infinite-Horizon Dynamic Programming
IV. Further Topics

GLOSSARY

Action One of several alternatives available to the decision maker when the system is observed in a particular state.
Decision rule Function that determines for the decision maker which action to select in each possible state of the system.
Discount factor Present value of a unit of currency received one period in the future.
Functional equation Basic entity in the dynamic programming approach to solving sequential decision problems. It relates the optimal value for a $(t+1)$-stage decision problem to the optimal value for a t-stage problem. Its solution determines an optimal decision rule at a particular stage.
Horizon Number of stages.
Markov chain Sequence of random variables in which the conditional probability of the future is independent of the past when the present state is known.
Markov decision problem Stochastic sequential decision problem in which the set of actions, the rewards, and the transition probabilities depend only on the current state of the system and the current action selected; the history of the problem has no effect on current decisions.
Myopic policy Policy in which each decision rule ignores the future consequence of the decision and uses the action at each stage that maximizes the immediate reward.
Policy Sequence of decision rules.
Stage Point in time at which a decision is made.
State Description of the system that provides the decision maker with all the information necessary to make future decisions.
Stationary Referring to a problem in which the set of actions, the set of states, the reward function, and the transition function are the same at each decision point or to a policy in which the same decision rule is used at every decision point.

IN ALL AREAS of endeavor, decisions are made either explicitly or implicitly. Rarely are decisions made in isolation. Today's decision has consequences for the future because it could affect the availability of resources or limit the options for subsequent decisions. A sequential decision problem is a mathematical model for the problem faced by a decision maker who is confronted with a sequence of interrelated decisions and wishes to make them in an optimal fashion. Dynamic programming is a collection of mathematical and computational tools for analyzing sequential decision problems. Its main areas of

application are operations research, engineering, statistics, and resource management. Improved computing capabilities will lead to the wide application of this technique in the future.

I. SEQUENTIAL DECISION PROBLEMS

A. Introduction

A system under the control of a decision maker is evolving through time. At each point of time at which a decision can be made, the decision maker, who will be referred to as "he" with no sexist connotations intended, observes the state of the system. On the basis of this information, he chooses an action from a set of alternatives. The consequences of this action are two-fold; the decision maker receives an immediate reward or incurs an immediate cost, and the state that the system will occupy at subsequent decision epochs is influenced either deterministically or probabilistically. The problem faced by the decision maker is to choose a sequence of actions that will optimize the performance of the system over the decision-making horizon. Since the action selected at present affects the future evolution of the system, the decision maker cannot choose his action without taking into account future consequences.

Dynamic programming is a procedure for finding optimal policies for sequential decision problems. It differs from linear, nonlinear, and integer programming in that there is no standard dynamic programming problem formulation. Instead, it is a collection of techniques based on developing mathematical recursions to decompose a multistage problem into a series of single-stage problems that are analytically or computationally more tractable. Its implementation often requires ingenuity on the part of the analyst, and the formulation of dynamic programming problems is considered by some practitioners to be an art. This subject is best understood through examples. This section proceeds with a formal introduction of the basic sequential decision problem and follows with several examples. The reader is encouraged to skip back and forth between these sections to understand the basic ingredients of such a problem. Dynamic programming methodology is discussed in Sections II and III.

B. Problem Formulation

Some formal notation follows. Let T denote the set of time points at which decisions can be made. The set T can be classified in two ways; it is either finite or infinite and either a discrete set or a continuum. The primary focus of this article is when T is discrete. Discrete-time problems are classified as either finite horizon or infinite horizon according to whether the set T is finite or infinite. The problem formulation in these two cases is almost identical; however, the dynamic programming methods of solution differ considerably. For discrete-time problems, T is the set $\{1, 2, \ldots, N\}$ in the finite case and $\{1, 2, \ldots\}$ in the infinite case. The present decision point is denoted by t and the subsequent point by $t+1$. The points of time at which decisions can be made are often called stages. Almost all the results in this article concern discrete-time models; the continuous-time model is briefly mentioned in Section IV.

The set of possible states of the system at time t is denoted by S_t. In finite-horizon problems, this is defined for $t = 1, 2, \ldots, N+1$, although decisions are made only at times $t = 1, 2, \ldots, N$. This is because the decision at time N often has future conseqences that can be summarized by evaluating the state of the system at time $N+1$. This is analagous to providing boundary values for differential equations. If at time t the decision maker observes the system in state $s \in S_t$, he chooses an action a from the set of allowable actions at time t, $A_{s,t}$. As above, S_t and $A_{s,t}$ can be either finite or infinite and discrete or continuous. This distinction has little consequence for the problem formulation.

As a result of choosing action a when the system is in state s at time t, the decision maker receives an immediate reward $r_t(s, a)$. This reward can be positive or negative. In the latter case it can be thought of as a cost. Furthermore, the choice of action affects the system evolution either deterministically or probabilistically. In the deterministic case, the choice of action determines the state of the system at time $t+1$ with certainty. Denote by $w_t(s, a) \in S_{t+1}$ the state the system will occupy if action a is chosen in state s at time t; $w_t(s, a)$ is called the transfer function. When the system evolves probabilistically, the subsequent state is random and choice of action specifies its probability distribution. Let $p_t(j|s, a)$ denote the probability that the system is in state $j \in S_{t+1}$ if action a is chosen in state s at time t; $p_t(j|s, a)$ is called the transition probability function. When S_t is a continuum, $p_t(j|s, a)$ is a probability density. Such models are discussed briefly in Section IV. A sequential decision problem in which the transitions from state to state are governed by a transition probability function and the set of actions and rewards depends only on the current state and stage is called a Markov decision problem.

The deterministic model is a special case of the probabilistic model obtained by choosing $p_t(j|s, a) = 1$ if $j = w_t(s, a)$ and $p_t(j|s, a) = 0$ if $j \neq w_t(s, a)$. Even though there is this equivalence, the transfer function representation is more convenient for deterministic problems.

A decision rule is a function $d_t : S_t \to A_{s,t}$ that specifies the action the decision maker chooses when the system is

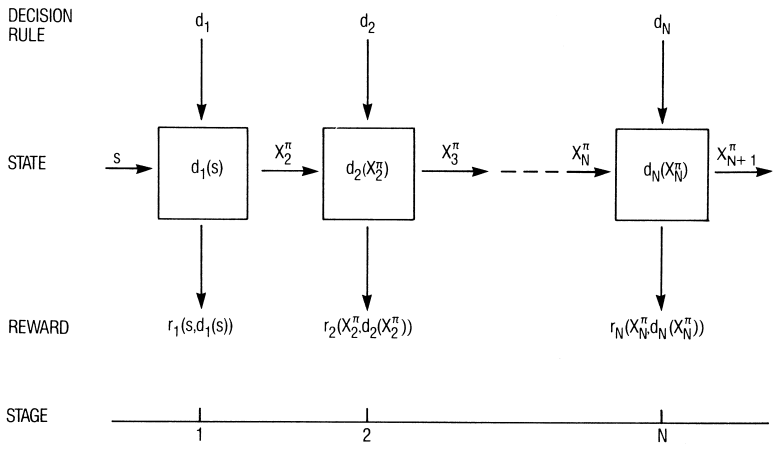


FIGURE 1 Evolution of the sequential decision model under the policy $\pi = (d_1, d_2, \ldots, d_N)$. The state at stage 1 is s.

in state s at time t; that is, $d_t(s)$ specifies an action in $A_{s,t}$ for each s in S_t. A decision rule of this type is called Markovian because it depends only on the current state of the system. The set of allowable decision rules at time t is denoted by D_t and is called the decision set. Usually it is the set of all functions mapping S_t to $A_{s,t}$, but in some applications, it might be a proper subset.

Many generalizations of deterministic Markovian decision rules are possible. Decision rules can depend on the entire history of the system, which is summarized in the sequence of observed states and actions observed up to the present, or they can depend only on the initial and current state of the system. Furthermore, the decision rule might be randomized; that is, in each state it specifies a probability distribution on the set of allowable actions so that by using such a rule the decision maker chooses his action at each decision epoch by a probabilistic mechanism. For the problems considered in this article, using deterministic Markovian decision rules at each stage is optimal so that the generalizations referred to above will not be discussed further.

A policy specifies the sequence of decision rules to be used by the decision maker over the course of the planning horizon. A policy π is a finite or an infinite sequence of decision rules; that is, $\pi = \{d_1, d_2, \ldots, d_N\}$, where $d_t \in D_t$ for $t = 1, 2, \ldots, N$ if the horizon is finite, or $\pi = \{d_1, d_2, \ldots\}$, where $d_t \in D_t$ for $t = 1, 2, \ldots$ if the horizon is infinite. Let Π denote the set of all possible policies; $\Pi = D_1 \times D_2 \times \cdots \times D_N$ in the finite case and $\Pi = D_1 \times D_2 \times \cdots$ in the infinite case.

In deterministic problems, by specifying a policy at the start of the problem, the decision maker completely determines the future evolution of the system. For each policy the sequence of states the system will occupy is known with certainty, and hence the sequence of rewards the decision maker will receive over the planning horizon is known. Let X_t^π denote the state the system occupies at time t if the decision maker uses policy π over the planning horizon. In the first period the decision maker receives a reward of $r_1(s, d_1(s))$, in the second period a reward of $r_2(X_2^\pi, d_2(X_2^\pi))$, and in the tth period $r_t(X_t^\pi, d_t(X_t^\pi))$. Figure 1 depicts the evolution of the process under a policy $\pi = \{d_1, d_2, \ldots, d_N\}$ in both the deterministic and stochastic cases. The quantity in each box indicates the interaction of the incoming state with the prespecified decision rule to produce the indicated action $d_t(X_t^\pi)$. The arrow to the right of a box indicates the resulting state, and the arrow downward the resulting reward to the decision maker. The system is assumed to be in state s before the first decision.

The decision maker evaluates policies by comparing the value of a function of the policy's income stream. Many such evaluation functions are available, but it is most convenient to assume a linear, additive, and risk-neutral utility function over time, which leads to using the total reward over the planning horizon for evaluation. Let $v_N^\pi(s)$ be the total reward over the planning horizon. It is given by the expression

$$v_N^\pi(s) = \sum_{t=1}^{N+1} r_t\left(X_t^\pi, d_t\left(X_t^\pi\right)\right), \tag{1}$$

in which it is implicit that $X_1^\pi = s$. For deterministic problems, evaluation formulas such as Eq. (1) always depend on the initial state of the process, although this is not explicitly stated below.

In probabilistic problems, by specifying a policy at the start of the problem, the decision maker determines the transition probability functions of a nonstationary Markov chain. The sequence of states the system will occupy is not known with certainty, and consequently the sequence of rewards the decision maker will receive over the planning

horizon is not known. Instead, what is known is the joint probability distribution of system states and rewards. In this case, expectations with respect to the joint probability distributions of the Markov chain conditional on the state at the first decision epoch are often used to evaluate policy performance. As in the deterministic case, let X_t^π denote the state the system occupies at time t if the decision maker uses policy π over the planning horizon. For finite-horizon problems, let $v_N^\pi(s)$ equal the total expected reward over the planning horizon. It is given by the expression

$$v_N^\pi(s) = E_{\pi,s}\left\{\sum_{t=1}^{N+1} r_t\left(X_t^\pi, d_t\left(X_t^\pi\right)\right)\right\}, \qquad (2)$$

where $E_{\pi,s}$ denotes the expectation with respect to probability distribution determined by π conditional on the initial state being s.

In both the deterministic and stochastic problems the decision maker's problem is to choose at time 1 a policy π in Π to make $v_N^\pi(s)$ as large as possible and to find the maximal reward,

$$v_N^*(s) = \sup_{\pi\in\Pi} v_N^\pi(s) \qquad \text{for all} \quad s \in S_1. \qquad (3)$$

Frequently the problem is such that the supremum in Eq. (3) is attained—for example, when both $A_{s,t}$ and S_t are finite for all $t \in T$. In such cases the decision maker's objective is to maximize $v_N^\pi(s)$ and find its maximal value.

For infinite-horizon problems, the total reward or the expected total reward the decision maker receives will not necessarily be finite; that is, the summations in Eqs. (1) and (2) usually will not converge. To evaluate policies in infinite-horizon problems, decision makers often use discounting or averaging. Let λ represent the discount factor, usually $0 \le \lambda < 1$. It measures the value at present of one unit of currency received one period from now. Let $v_\lambda^\pi(s)$ equal the total discounted reward in deterministic problems or the expected total discounted reward for probabilistic problems if the system is in state s, before choosing the first action. For deterministic problems it is given by

$$v_\lambda^\pi(s) = \sum_{t=1}^{\infty} \lambda^{t-1} r_t\left(X_t^\pi, d_t\left(X_t^\pi\right)\right), \qquad (4)$$

and for stochastic problems it is given by

$$v_\lambda^\pi(s) = E_{\pi,s}\left\{\sum_{t=1}^{\infty} \lambda^{t-1} r_t\left(X_t^\pi, d_t\left(X_t^\pi\right)\right)\right\}. \qquad (5)$$

In this setting the decision maker's problem is to choose at time 1 a policy π in Π to make $v_\pi^\lambda(s)$ as large as possible and to find the supremal reward,

$$v_\lambda^*(s) = \sup_{\pi\in\Pi} v_\lambda^\pi(s) \qquad \text{for all} \quad s \in S_1. \qquad (6)$$

Alternatively in the infinite-horizon setting, the decision maker might not be willing to assume that a reward received in the future is any less valuable than a reward received at present. For example, if decision epochs are very close together in real time, then all rewards the decision maker receives would have equal value. In this case the decision maker's objective might be to choose a policy that maximizes the average or expected average reward per period. This quantity is frequently called the gain of a policy. For a specified policy it is denoted by $g^\pi(s)$ and in both problems is given by

$$g^\pi = \lim_{N\to\infty} \frac{1}{N} v_N^\pi(s), \qquad (7)$$

where $v_N^\pi(s)$ is defined in Eqs. (1) and (2).

In this setting the decision maker's problem is to choose at time 1 a policy π in Π to make $g^\pi(s)$ as large as possible and to find the supremal average reward,

$$g^*(s) = \sup_{\pi\in\Pi} g^\pi(s) \qquad \text{for all} \quad s \in S_1. \qquad (8)$$

Dynamic programming methods with the average reward criteria are quite complex and are discussed only briefly in Section III. The reader is referred to the works cited in the Bibliography for more details.

Frequently in infinite-horizon problems, the data are stationary. This means that the set of states, the set of allowable actions in each state, the one-period rewards, the transition or transfer functions, and the decision sets are the same at every stage. When this is the case, the time subscript t is deleted and the notation S, A_s, $r(s,a)$, $p(j|s,a)$ or $w(s,a)$, and D is used. Often stationary policies are optimal in this setting. By a stationary policy is meant a policy that uses the identical decision rule in each period; that is, $\pi = (d, d, \ldots)$. Often it is denoted by d when there is no possible source of confusion.

C. Examples

The following examples clarify the notation and formulation described in the preceding sections. The first example illustrates a deterministic, finite-state, finite-action, finite-horizon problem; the second a deterministic, infinite-state, infinite-action, finite-horizon problem; and the third a stochastic, finite-state, finite-action problem with both finite- and infinite-horizon versions. In Sections II and III, these examples will be solved by using dynamic programming methodology.

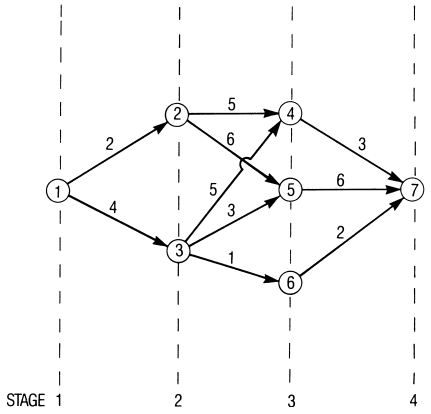


FIGURE 2 Network for the longest-route problem.

1. A Longest-Route Problem

A finite directed graph is depicted in Fig. 2. The circles are called nodes, and the lines connecting them are called arcs. On each arc, an arrow indicates the direction in which movement is possible. The numerical value on the arc is the reward the decision maker receives if he chooses to traverse that arc on his journey from node 1 to node 7. His objective is to find the path from node 1 to node 7 that maximizes the total reward he receives on his journey. Such a problem is called a longest-route problem. A practical application is determining the length of time needed to complete a project. In such problems, the arc length represents the time to complete a task. The entire project is not finished until all tasks are performed. Finding the longest path through the network gives the minimum amount of time for the entire project to be completed since it corresponds to that sequence of tasks that requires the most time. The longest path is called a critical path because, if any task in this sequence is delayed, the entire project will be delayed.

In other applications the values on the arcs represent lengths of road segments, and the decision maker's objective is to find the shortest route from the first node to the terminal node. Such a problem is called a shortest-route problem. All deterministic, finite-state, finite-action, finite-horizon dynamic programming problems are equivalent to shortest- or longest-route problems. Another example of this will appear in Section II.

An important assumption is that the network contains no directed cycle; that is, there is no route starting at a node that returns to that node. If this were the case, the longest route would be infinite and the problem would not be of interest.

The longest-route problem is now formulated as a sequential decision problem. This requires defining the set of states, actions, decision sets, transfer functions, and rewards. They are as follows:

Decision points:

$$T = \{1, 2, 3\}$$

States (numbers correspond to nodes):

$$S_1 = \{1\}; \qquad S_2 = \{2, 3\}; \qquad S_3 = \{4, 5, 6\};$$
$$S_4 = \{7\}$$

Actions (action j selected in node i corresponds to choosing to traverse the arc between nodes i and j; the first subscript on A is the state and the second the stage):

$$A_{1,1} = \{2, 3\}$$
$$A_{2,2} = \{4, 5\}; \qquad A_{3,2} = \{4, 5, 6\}$$
$$A_{4,3} = \{7\}; \qquad A_{5,3} = \{7\}; \qquad A_{6,3} = \{7\}$$

Rewards:

$$r_1(1, 2) = 2; \qquad r_1(1, 3) = 4$$
$$r_2(2, 4) = 5; \qquad r_2(2, 5) = 6;$$
$$r_2(3, 4) = 5, \qquad r_2(3, 5) = 3; \qquad r_2(3, 6) = 1$$
$$r_3(4, 7) = 3; \qquad r_3(5, 7) = 6; \qquad r_3(6, 7) = 2$$

Transfer function:

$$w_t(s, a) = a$$

The remaining ingredients in the sequential decision problem formulation are the decision set, the set of policies, and an evaluation formula. The decision set at stage t is the set of all arcs emanating from nodes at stage t. A policy is a list of arcs in which there is one arc starting at each node (except 7) in the network. The policy set contains all such lists. Each policy contains a route from node 1 to node 7 and some superfluous action selections. The value of the policy is the total of the rewards along this route, and the decision maker's problem is to choose a policy that maximizes this total reward.

The structure of a policy is described in more detail through an example. Consider the policy $\pi = \{(1, 2), (2, 4), (3, 4), (4, 7), (5, 7), (6, 7)\}$. Implicit in this definition is a sequence of decision rules $d_t(s)$ for each state and stage. They are $d_1(1) = 2$, $d_2(2) = 4$, $d_2(3) = 4$, $d_3(4) = 7$, $d_3(5) = 7$, and $d_3(6) = 7$. This policy can be formally denoted by $\pi = \{d_1, d_2, d_3\}$. The policy contains one unique routing through the graph, namely, $1 \to 2 \to 4 \to 7$ and several unnecessary decisions. We use the formal notation $X_1^\pi = 1$, $X_2^\pi = 2$, $X_3^\pi = 4$, and $X_4^\pi = 7$ so that

$$\begin{aligned} v_3^\pi(1) &= r_1(1, d_1(1)) + r_2(2, d_2(2)) + r_3(4, d_3(4)) \\ &= r_1(1, 2) + r_2(2, 4) + r_3(4, 7) \\ &= 2 + 5 + 3 = 10. \end{aligned}$$

In such a small problem, one can easily evaluate all policies by enumeration and determine that the longest route through the network is $1 \to 2 \to 5 \to 7$ with a return of 14. For larger problems this is not efficient; dynamic programming methods will be seen to offer an efficient means of determining the longest route.

The reader might note that the formal sequential decision process notation is quite redundant here. The subscript for stage does not convey any useful information and the specification of a policy requires making decisions in nodes that will never be reached. Solution by dynamic programming methods will require this superfluous information. In other settings this information will be useful.

2. A Resource Allocation Problem

A decision maker has a finite amount K of a resource to allocate between N possible activities. Using activity i at level x_i consumes $c_i(x_i)$ units of the resource and yields a reward or utility of $f_i(x_i)$ to the decision maker. The maximum level of intensity for activity i is M_i. His objective is to determine the intensity for each of the activities that maximizes his total reward. When any level of the activity is possible, this is a nonlinear programming problem. When the activity can operate only at a finite set of levels, this is an integer programming problem. In the special case that the activity can be either utilized or not ($M_i = 1$ and x_i is an integer) this is often called a knapsack problem. This is because it can be used to model the problem of a camper who has to decide which of N potential items to carry in his knapsack. The value of item i is $f_i(1)$ and it weighs $c_i(1)$. The camper wishes to select the most valuable set of items that do not weigh more than the capacity of the knapsack.

The mathematical formulation of the resource allocation problem is as follows:

Maximize $f_1(x_1) + f_2(x_2) + \cdots + f_N(x_N)$
subject to

$$c_1(x_1) + c_2(x_2) + \cdots + c_N(x_N) \le K \tag{9}$$

and

$$0 \le x_t \le M_t, \qquad t = 1, 2, \ldots, N. \tag{10}$$

The following change of variables facilitates the sequential decision problem formulation. Define the new variable $s_i = c_i(x_i)$ and assume that c_i is a monotone increasing function on $[0, M_i]$. this assumption says that the more intense the activity level, the more resource utilized. Define $g_i(s_i) = f_i(c_i^{-1}(s_i))$ and $m_i = c_i^{-1}(M_i)$. This change of variables corresponds to formulating the problem in terms of the quantity of resource being used. In this notation the formulation above becomes

Maximize $g_1(s_1) + g_2(s_2) + \cdots + g_N(s_N)$
subject to

$$s_1 + s_2 + \cdots + s_N \le K \tag{11}$$

and

$$0 \le s_t \le m_t, \qquad t = 1, 2, \ldots, N. \tag{12}$$

It is not immediately obvious that this problem is a sequential decision problem. The formulation is based on treating the problem as if the decision to allocate resources to the activities were done sequentially through time with allocation to activity 1 first and activity N last. Decisions are coupled, since successive allocations must take the quantity of the resource allocated previously into account. That is, if $K - s$ units of resource have been allocated to the first t activities, then s units are available for activities $t+1, t+2, \ldots, N$.

The following sequential decision problem formulation is based on the second formulation above:

Decision points (correspond to activity number):

$$T = \{1, 2, \ldots, N\}$$

States (amount of resource available for allocation in remaining stages): For $0 \le t \le N$,

$$S_t = \begin{cases} \{s : 0 \le s \le m_t\} \\ \qquad \text{if resource levels are continuous} \\ \{0, 1, 2, \ldots, m_t\} \\ \qquad \text{if resource levels are discrete} \end{cases}$$

For $t = N + 1$,

$$S_t = \begin{cases} \{s : 0 \le s \le K\} \\ \qquad \text{if resource levels are continuous} \\ \{0, 1, 2, \ldots, K\} \\ \qquad \text{if resource levels are discrete} \end{cases}$$

Actions (s is amount of resource available for stages t, $t+1, \ldots, N$):

$$A_{s,t} = \begin{cases} \{u : 0 \le u \le \min(s, m_t)\} \\ \qquad \text{if resource levels are continuous} \\ \{0, 1, 2, \ldots, \min(s, m_t)\} \\ \qquad \text{if resource levels are discrete} \end{cases}$$

Rewards:

$$r_t(s, a) = g_t(a)$$

Transfer function:

$$w_t(s, a) = s - a$$

The decision set at stage t is the set of all functions from S_t to $A_{s,t}$, and a policy is a sequence of such functions, one for each $t \in T$. A decision rule specifies the amount of resource to allocate to activity t if s units are available for allocation to activities $t, t+1, \ldots, N$, and a policy

specifies which decision rule to use at each stage of the sequential allocation. As in the longest-route problem, a policy specifies decisions in many eventualities that will not occur using that policy. This may seem wasteful at first but is fundamental to the dynamic programming methodology. The quantity X_t^π is the amount of remaining resource available for allocation to activities $t, t+1, \ldots, N$ using policy π. Clearly, $X_1^\pi = K$, $X_2^\pi = K - d_1(K)$, and so forth. The decision maker compares policies through the quantity $v_N^\pi(K)$, which is given by

$$v_N^\pi(K) = g_1(d_1(K)) + g_2\big(d_2\big(X_2^\pi\big)\big) + \cdots + g_N\big(d_N\big(X_N^\pi\big)\big).$$

When the set of activities is discrete, the resource allocation problem can be formulated as a longest-route problem, as can any discrete state and action sequential decision problem. This is depicted in Fig. 3 for the following specific example:

$$\begin{aligned} &\text{Maximize} && 3s_1^2 + s_2^3 + 4s_3 \\ &\text{subject to} \\ & && s_1 + s_2 + s_3 \le 4, \end{aligned}$$

s_1, s_2, and s_3 are integers, and

$$0 \le s_1 \le 2; \qquad 0 \le s_2 \le 2; \qquad 0 \le s_3 \le 2.$$

In the longest-route formulation, the node labels are the amount of resource available for allocation at subsequent stages. A fifth stage is added so that there is a unique destination and all decisions at stage 4 correspond to moving from a node at stage 4 to node 0 with no reward. This is because an unallocated resource has no value to the decision maker in this formulation. The number on each arc is the reward, and the amount of resource allocated is the difference between the node label at stage t and that at stage $t+1$. For instance, if at stage 2, there are 3 units of resource available for allocation over successive stages and the decision maker decides to allocate 2 units, he will receive a reward of $2^3 = 8$ units and move to node 1.

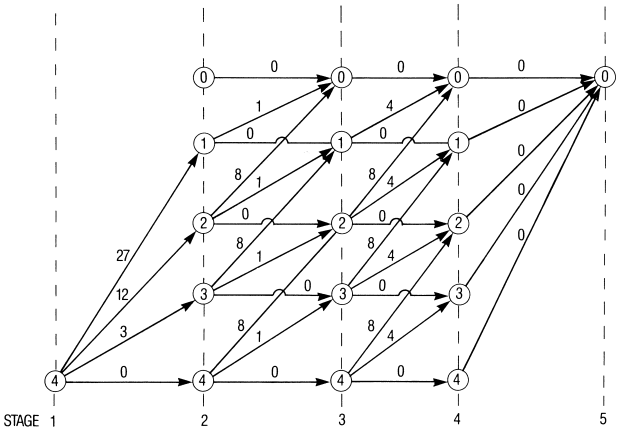


FIGURE 3 Network for the resource allocation problem.

When the resource levels form a continuum, the network representation is no longer valid. The problem is reduced to a sequence of constrained one-dimensional nonlinear optimization problems through dynamic programming. Such an example will be solved in Section II by using dynamic programming methods.

3. A Stochastic Inventory Control Problem

Each month, the manager of a warehouse must determine how much of a product to keep in stock to satisfy customer demand for the product. The objective is to maximize expected total profit (sales revenue less inventory holding and ordering costs), which may or may not be discounted. The demand throughout the month is random, with a known probability distribution. Several simplifying assumptions make a concise formulation possible:

a. The decision to order additional stock is made at the beginning of each month and delivery occurs instantaneously.

b. Demand for the product arrives throughout the month but is filled on the last day of the month.

c. If demand exceeds the stock on hand, the customer goes elsewhere to purchase the product.

d. The revenues and costs and the demand distribution are identical each month.

e. The product can be sold only in whole units.

f. The warehouse capacity is M units.

Let s_t denote the inventory on hand at the beginning of month t, a_t the additional product ordered in month t, and D_t the random demand in month t. The demand has a known probability distribution given by $p_d = P\{D_t = d\}$, $d = 0, 1, 2, \ldots$. The cost of ordering u units in any month is $O(u)$ and the cost of storing u units for 1 month is $h(u)$. The ordering cost is given by

$$O(u) = \begin{cases} K + c(u), & \text{if} \quad u > 0 \\ 0, & \text{if} \quad u = 0, \end{cases} \tag{13}$$

where $c(u)$ and $h(u)$ are increasing functions of u. For finite-horizon problems, if u units of inventory are on hand at the end of the planning horizon, its value is $g(u)$. Finally, if u units of product are demanded in a month and the inventory is sufficient to satisfy demand, the manager receives $f(u)$. Define $F(u)$ to be the expected revenue in a month if the inventory before receipt of customer orders is u units. It is given in period t by

$$F(u) = \sum_{j=0}^{u-1} f(j)p_j + f(u)P\{D_t \ge u\}. \tag{14}$$

Equation (14) can be interpreted as follows. If the inventory on hand exceeds the quantity demanded, j, the revenue is $f(j)$; p_j is the probability that the demand in a period is j units. If the inventory on hand is u units and the quantity demanded is at least u units, then the revenue is $f(u)$ and $P\{D_t \geq u\}$ is the probability of such a demand. The combined quantity is the probability-weighted, or expected, revenue.

This is a stochastic sequential decision problem (Markov decision problem), and its formulation will include a transition probability function instead of a transfer function. The formulation follows:

Decision points:

$$T = \{1, 2, \ldots, N\};$$
N may be finite or infinite

States (units of inventory on hand at the start of a month):

$$S_t = \{0, 1, 2, \ldots M\}, \qquad t = 1, 2, \ldots, N+1$$

Actions (the amount of additional stock to order in month t):

$$A_{s,t} = \{0, 1, 2, \ldots, M - s\}$$

Expected rewards (expected revenue less ordering and holding costs):

$$r_t(s, a) = F(s + a) - O(a) - h(s + a),$$
$$t = 1, 2, \ldots, N$$

Value of terminal inventory (no actions are possible):

$$r_{N+1}(s, a) = g(s)$$

Transition probabilities (see explanation below):

$$p_t(j|s, a) = \begin{cases} 0, & \text{if} \quad j > s + a \\ p_j, & \text{if} \quad j = s + a - D_t, \\ & \quad s + a \leq M, \quad \text{and} \\ & \quad s + a > D_t \\ q_{s+a}, & \text{if} \quad j = 0, \quad s + a \leq M, \\ & \quad \text{and} \quad s + a \leq D_t \end{cases}$$

where

$$q_{s+a} = P\{D_t \geq s + a\} = \sum_{d=s+a}^{\infty} p_d.$$

A brief explanation of the transition probabilities might be helpful. If the inventory on hand at the beginning of period t is s units and an order is placed for a units, the inventory before external demand is $s + a$ units. If the demand of j units is less than $s + a$ units, then the inventory at the beginning of period $t + 1$ is $s + a - j$ units. This occurs with probability p_j. If the demand exceeds $s + a$ units, then the inventory at the start of period $t + 1$ is 0 units. This occurs with probability q_{s+a}. Finally, the probability that the inventory level ever exceeds $s + a$ units is zero, since this demand is nonnegative.

The decision sets consist of all rules that assign the quantity of inventory to be ordered each month to each possible starting inventory position in a month. A policy is a sequence of such ordering rules. Unlike deterministic problems, in which a decision rule is specified for many states that will never be reached, in stochastic problems such as this, it is necessary for the decision maker to determine the decision rule for *all* states. This is because the evolution of the inventory level over time is random, which makes any inventory level possible at any decision point. Consequently the decision maker must plan for each of these eventualities.

An example of a decision rule is as follows: Order only if the inventory level is below 3 units at the start of the month and order the quantity that raises the stock level to 10 units on receipt of the order. In month t this is given by

$$d_t(s) = \begin{cases} 10 - s, & s < 3 \\ 0, & s \geq 3. \end{cases}$$

The evaluation method for a policy depends on the time horizon under consideration. For finite-horizon problems, the total expected cost conditional on the initial stock level is a convenient summary. Assuming the stock level at time 1 is s, the expected total reward for policy π is

$$v_N^{\pi}(s) = E_{\pi,s}\left\{\left[\sum_{t=1}^{N} F\left(X_t^{\pi} + d_t^{\pi}\left(X_t^{\pi}\right)\right) - O\left(d_t^{\pi}\left(X_t\right)\right) - h\left(X_t^{\pi} + d_t\left(X_t^{\pi}\right)\right)\right] - g\left(X_{N+1}^{\pi}\right)\right\}.$$

If, instead, the decision maker wishes to discount future profit at a monthly discount rate of λ, $0 \leq \lambda < 1$, the term λ^{t-1} is inserted before each term in the above summation and λ^N before the terminal reward g. For an infinite-horizon problem, the expected total discounted profit is given by

$$v_{\lambda}^{\pi} = E_{\pi,s}\left\{\sum_{t=1}^{\infty} \lambda^{t-1}\left[F\left(X_t^{\pi} + d_t\left(X_t^{\pi}\right)\right) - O\left(d_t^{\pi}\left(X_t^{\pi}\right)\right) + h\left(X_t^{\pi} + d_t(X^{\pi})\right)\right]\right\}.$$

The decision maker's problem is to choose a sequence of decision rules to maximize expected total or total discounted profits.

Many modifications of this inventory problem are possible; for example, excess demand in any period could be backlogged and a penalty for carrying unsatisfied demand could be charged, or there could be a time lag between placing the order and its receipt. The formulation herein

can easily be modified to include such changes; the interested reader is encouraged to consult the Bibliography for more details.

A numerical example is now provided in complete detail. It will be solved in subsequent sections by using dynamic programming methods. The data for the problem are as follows: $K = 4$, $c(u) = 2u$, $g(u) = 0$, $h(u) = u$, $M = 3$, $N = 3$, $f(u) = 8(u)$, and

$$p_d = \begin{cases} \frac{1}{4}, & \text{if} \quad d = 0 \\ \frac{1}{2}, & \text{if} \quad d = 1 \\ \frac{1}{4}, & \text{if} \quad d = 2. \end{cases}$$

The inventory is constrained to be 3 or fewer units, and the decision maker wishes to consider the effects over three periods. All the costs and revenues are linear. This means that for each unit ordered the per unit cost is 2, for each unit held in inventory for 1 month the per unit cost is 1, and for each unit sold the per unit revenue is 8. The expected revenue when u units of stock are on hand before receipt of an order is given by

u	$F(u)$
0	0
1	$0 \times \frac{1}{4} + 8 \times \frac{3}{4} = 6$
2	$0 \times \frac{1}{4} + 8 \times \frac{1}{2} + 16 \times \frac{1}{4} = 8$
3	$0 \times \frac{1}{4} + 8 \times \frac{1}{2} + 16 \times \frac{1}{4} = 8$

Combining the expected revenue with the expected shortage and holding costs gives the expected profit in period t if the inventory level is s at the start of the period and an order for a units is placed. If $a = 0$, the ordering and holding cost equals s, and if a is positive, it equals $4 + s + 3a$. It is summarized in the tabulations below, where an X corresponds to an action that is infeasible. Transition probabilities depend only on the total inventory on hand before the receipt of orders. They are the same for any s and a that have the same total $s + a$. So that redundant information is reduced, transition probabilities are presented as functions of $s + a$ only. The information in the following tabulations defines this problem completely:

	$r_t(s,a)$			
s	$a=0$	$a=1$	$a=2$	$a=3$
0	0	−1	−2	−5
1	5	0	−3	X
2	6	−1	X	X
3	5	X	X	X

	$p_t(j\|s,a)$			
$s+a$	$j=0$	$j=1$	$j=2$	$j=3$
0	0	0	0	0
1	$\frac{3}{4}$	$\frac{1}{4}$	0	0
2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0
3	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

II. FINITE-HORIZON DYNAMIC PROGRAMMING

A. Introduction

Dynamic programming is a collection of methods for solving sequential decision problems. The methods are based on decomposing a multistage problem into a sequence of interrelated one-stage problems. Fundamental to this decomposition is the *principle of optimality*, which was developed by Richard Bellman in the 1950s. Its importance is that an optimal solution for a multistage problem can be found by solving a functional equation relating the optimal value for a $(t+1)$-stage problem to the optimal value for a t-stage problem.

Solution methods for problems depend on the time horizon and whether the problem is deterministic or stochastic. Deterministic finite-horizon problems are usually solved by backward induction, although several other methods, including forward induction and reaching, are available. For finite-horizon stochastic problems, backward induction is the only method of solution. In the infinite-horizon case, different approaches are used. These will be discussed in Section III. The backward induction procedure is described in the next two sections. This material might seem difficult at first; the reader is encouraged to refer to the examples at the end of this section for clarification.

B. Functional Equation of Dynamic Programming

Let $v^t(s)$ be the maximal total reward received by the decision maker during stages $t, t+1, \ldots, N+1$, if the system is in state s immediately before the decision at stage t. When system transitions are stochastic, $v^t(s)$ is the maximal expected return. Recall that decisions are made at stages $1, 2, \ldots, N$ and not at time $N+1$; however, a reward might be received at stage $N+1$ as a consequence of the decision in stage N. In most deterministic problems, S_1 consists of one element, whereas in stochastic problems, solutions are usually required for all possible initial states.

The basic entity of dynamic programming is the functional equation, or Bellman equation, which relates $v^t(s)$ to $v^{t+1}(s)$. For deterministic problems it is given by

$$v^t(s) = \max_{a \in A_{s,t}} \{ r_t(s,a) + v^{t+1}(w_t(s,a)) \}, \quad t = 1, \ldots, N \tag{15}$$

and

$$v^{N+1}(s) = 0, \tag{16}$$

where Eq. (15) is valid for all $s \in S_t$ and Eq. (16) is valid for all $s \in S_{N+1}$. Equation (15) is the basis for the backward

induction algorithm for solving sequential decision problems. This equation corresponds to a one-stage problem in which the decision maker observes the system in state s and must select an action from the set $A_{s,t}$. The consequence of this action is that the decision maker receives an immediate reward of $r_t(s, a)$ and the system moves to state $w_t(s, a)$, at which he receives a reward of $v^{t+1}(w_t(s, a))$. Equation (15) says that he chooses the action that maximizes the total of these two rewards. This is exactly the problem faced by the decision maker in a one-stage sequential decision problem when the terminal reward function is v^{t+1}. Equation (16) provides a boundary condition. When the application dictates, this value 0 can be replaced by an arbitrary function that assigns a value to the terminal state of the system. Such might be the case in the inventory control example.

Equation (15) emphasizes the dynamic aspects of the sequential decision problem. The decision maker chooses that action which maximizes his immediate reward *plus* his reward over the remaining decision epochs. This is in contrast to the situation in which the decision maker behaves myopically and chooses the decision rule that maximizes the reward only in the current period and ignores future consequences. Some researchers have given conditions in which such a myopic policy is optimal; however, in almost all problems dynamic aspects must be taken into account.

The expression "max" requires explanation because it is fundamental to the dynamic programming methodology. If $f(x, y)$ is any function of two variables with $x \in X$ and $y \in Y$, then

$$g(x) = \max_{y \in Y}\{f(x, y)\}$$

if for each $x \in X$, $g(x) \geq f(x, y)$ for all $y \in Y$ and there exists a $y^* \in Y$ with the properties that $f(x, y^*) \geq f(x, y)$ for all $y \in Y$ and $g(x) = f(x, y^*)$. Thus, Eq. (15) states that the decision maker chooses $a \in A_{s,t}$ to make the expression in braces as large as possible. The quantity $v^t(s)$ is set equal to this maximal value.

In stochastic problems, the functional equation (15) is modified to account for the probabilistic transition structure. It is given by

$$v^t(s) = \max_{a \in A_{s,t}}\left\{r_t(s, a) + \sum_{j \in S_{t+1}} p_t(j|s, a)v^{t+1}(j)\right\}, \quad t = 1, \ldots, N. \tag{17}$$

The stochastic nature of the problem is accounted for in Eq. (17) by replacing the fixed transition function $w_t(s, a)$ by the random state j, which is determined by the probability transition function corresponding to selecting action a. The second expression in this equation equals the expected reward received over the remaining periods as a consequence of choosing action a in period t.

C. Backward Induction and the Principle of Optimality

Backward induction is a procedure that uses the functional equation in an iterative fashion to find the optimal total value function and an optimal policy for a finite-horizon sequential decision problem. That this method achieves these objectives is demonstrated by the principle of optimality. The principle of optimality is not a universal truth that applies to all sequential decision problems but a mathematical result that requires formal proof in each application. For problems in which the (expected) total reward criterion is used, as considered in this article, it is valid. A brief argument of why it holds in such problems is given below.

To motivate backward induction, the following iterative procedure for finding the total reward of some specified policy $\pi = (d_1, d_2, \ldots, d_N)$ is given. It is called the policy evaluation algorithm. To simplify notation, assume that $p_t(j|s, a) = p(j|s, a)$ and $r_t(s, a) = r(s, a)$ for all s, a, and j.

a. Set $t = N + 1$ and $v^{N+1}(s) = 0$ for all $s \in S_{N+1}$.

b. Substitute $t - 1$ for $t(t - 1 \to t)$ and compute $v^t(s)$ for each $s \in S_t$ in the deterministic case by

$$v^t(s) = r(s, d_t(s)) + v^{t+1}(w_t(s, d_t(s))), \tag{18}$$

or in the stochastic case by

$$v^t(s) = r(s, d_t(s)) + \sum_{j \in S_{t+1}} p(j|s, d_t(s))v^{t+1}(j). \tag{19}$$

c. If $t = 1$, stop; otherwise, return to step b.

This procedure inductively evaluates the policy by first fixing its value at the last stage and then computing its value at the previous stage by adding its immediate reward to the previously computed total value. This process is repeated until the first stage is reached. This computation process yields the quantities $v^1(s)$, $v^2(s), \ldots, v^{N+1}(s)$. The quantity $v^1(s)$ equals the expected total value of policy π, which in earlier notation is given by $v_\pi^N(s)$. The quantities $v^1(s)$ correspond to the value of this policy from stage t onward. This procedure is extended to optimization by iteratively choosing and evaluating a policy consisting of the actions that give the maximal return from each stage to the end of the planning horizon instead of just evaluating a fixed prespecified policy.

The backward induction algorithm proceeds as follows:

a. Set $t = N + 1$ and $v^{N+1}(s) = 0$ for all $s \in S_{N+1}$.
b. Substitute $t - 1$ for $t(t - 1 \rightarrow t)$ and compute $v^t(s)$ for each $s \in S_t$ using Eq. (15) or (17) depending on which is appropriate. Denote by $A^*_{s,t}$ the set of actions a^* for which in the deterministic case,

$$v^t(s) = r(s, a^*) + v^{t+1}\big(w_t(s, a^*)\big), \tag{20}$$

or in the stochastic case,

$$v^t(s) = r(s, a^*) + \sum_{j \in S_{t+1}} p(j|s, a^*) v^{t+1}(j). \tag{21}$$

c. If $t = 1$, stop; otherwise, return to step b.

By comparing this procedure with the policy evaluation procedure above, we can easily see that the backward induction algorithm accomplishes three objectives:

a. It finds sets of actions $A^*_{s,t}$ that contain all actions in $A_{s,t}$ that obtain the maximum in Eq. (15) or (17).
b. It evaluates any policy made up of actions selected from the sets $A^*_{s,t}$.
c. It gives the total return or expected total return $v_t(s)$ that would be obtained if a policy corresponding to selecting actions in $A^*_{s,t}$ were used from stage t onward.

Thus, if the decision maker had specified a policy that selected actions in the sets $A^*_{s,t}$ before applying the policy evaluation algorithm, these two procedures would be identical. It will be argued below that any policy obtained by selecting an action from $A^*_{s,t}$ in each state at every stage is optimal and consequently $v^1(s)$ is the optimal value function for the problem; that is, $v^1(s) = v^*_N(s)$.

In deterministic problems, specifying a policy often provides much superfluous information, since if the state of the system is known before the first decision, a policy determines the system evolution with certainty and only one state is reached at each stage. Since all deterministic problems are equivalent to longest-route problems, the objective in such problems is *only* to find a longest route. The following route selection algorithm does this. The system state is known before decision 1, so S_1 contains a single state.

a. Set $t = 1$ and for $s \in S_t$ define $d_t(s) = a^*$ for some $a^* \in A^*_{s,t}$. Set $u = w_t(s, a^*)$.
b. For $u \in S_{t+1}$, define $d_{t+1}(u) = a^*$ for some $a^* \in A^*_{u,t+1}$. Replace u by $u = w_{t+1}(u, a^*)$.
c. If $t + 1 = N$, stop; otherwise, $t + 1 \rightarrow t$ and return to step b.

In this algorithm, the choice of an action at each node determines which arc will be traversed, and decisions are necessary only at nodes that can be reached from the initial state. the algorithm traces forward through the network along the path determined by decisions that obtain the maximum in Eq. (15). It produces *one* route through the network with longest length. If at any stage the set $A^*_{s,t}$ contains more than one action, then several optimal routings exist, and if all are desired, the procedure must be carried out to trace each path.

A problem closely related to the longest-route problem is that of finding the longest route from *each* node to the final node. When there is only one action in $A^*_{s,t}$ for each s and t, then specifying the decision rule that in each state is equal to the unique maximizing action produces these routings. This is closer in spirit to the concept of a policy than a longest route.

That the above procedure results in an optimal policy and optimal value function is due to the additivity of rewards in successive periods. A formal proof of these results is based on induction, but the following argument gives the main idea. The backward induction algorithm chooses maximizing actions in reverse order. It does not matter what happened before the current stage. The only important information for future decisions is the current state of the system. First, for stage N the best action in each state is selected. Clearly, $v^N(s)$ is the optimal value function for a one-stage problem beginning in stage s at stage N. Next, in each state at stage $N - 1$, an action is found to maximize the immediate reward plus the reward that will be obtained if, after reaching a state at stage N, the decision maker chooses the optimal action at that stage. Clearly, $v^{N-1}(s)$ is the optimal value for a one-stage problem with terminal reward $v^N(s)$. Since $v^N(s)$ is the optimal value for the one-stage problem starting at stage N, no greater total reward can be obtained over these two stages. Hence, $v^{N-1}(s)$ is the optimal reward from stage $N - 1$ onward starting in state s. Now, since the sets $A^*_{s,N}$ and $A^*_{s,N-1}$ have been determined by the backward induction algorithm, choosing any policy that selects actions from these sets at each stage and evaluating it with the policy evaluation algorithm above will also yield $v^{N-1}(s)$. Thus, this policy is optimal over these two stages since its value equals the optimal value. This argument is repeated at stages $N - 2, N - 3, \ldots, 1$ to conclude that a policy that selects an action from $A^*_{s,t}$ at each stage is optimal.

The above argument contains the essence of the principle of optimality, which appeared in its original form on p. 83 of Bellman's classic book, "Dynamic Programming," as follows:

> An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

The functional equations (15) and (17) are mathematical statements of this principle.

It might not be obvious to the reader why the backward induction algorithm is more attractive than an enumeration procedure for solving sequential decision problems. To see this, suppose that there are N stages, M states at each stage, and K actions that can be chosen in each state. Then there are $(K^M)^N$ policies. Solving a deterministic problem by enumeration would require $N(K^M)^N$ additions and $(K^M)^N$ comparisons. By backward induction, solution would require NMK additions and NK comparisons, a potentially astronomical savings in work. Solving stochastic problems requires additional M multiplications at each state at each stage to evaluate expectations. Enumeration requires $MN(K^M)^N$ multiplications, whereas backward induction would require NKM multiplications. Clearly, backward induction is a superior method for solving any problem of practical significance.

D. Examples

In this section, the use of the backward induction algorithm is illustrated in terms of the examples that were presented in Section I. First, the longest-route problem in Fig. 2 is considered.

1. A Longest-Route Problem

Note that $N = 3$ in this example.

a. Set $t = 4$ and $v^4(7) = 0$.
b. Since $t \neq 1$, continue. Set $t = 3$ and

$$\begin{aligned} v^3(4) &= r_3(4, 7) + v^4(7) \\ &= 3 + 0 = 3, \\ A^*_{6,3} &= \{7\}, \\ v^3(5) &= r_3(5, 7) + v^4(7) \\ &= 6 + 0 = 6, \\ A^*_{5,3} &= \{7\}, \\ v^3(6) &= r_3(6, 7) + v^4(7) \\ &= 2 + 0 = 2, \\ A^*_{4,3} &= \{7\}. \end{aligned}$$

c. Since $t \neq 1$, continue. Set $t = 2$ and

$$\begin{aligned} v^2(2) &= \max\{r_2(2, 4) + v^3(4), r_2(2, 5) + v^3(5)\} \\ &= \max\{5 + 3, 6 + 6\} = 12, \\ v^2(3) &= \max\{r_2(3, 4) + v^3(4), r_2(3, 5) + v^3(5), \\ &\quad r_2(3, 6) + v^3(6)\} \\ &= \max\{5 + 3, 3 + 6, 1 + 2\} = 9, \\ A^*_{2,2} &= \{5\}; \qquad A^*_{3,2} = 5. \end{aligned}$$

d. Since $t \neq 1$, continue. Set $t = 1$ and

$$\begin{aligned} v^1(1) &= \max\{r_1(1, 2) + v^2(2), r_1(1, 3) + v^2(3)\} \\ &= \max\{2 + 12, 4 + 9\} = 14, \\ A^*_{1,1} &= \{2\}. \end{aligned}$$

e. Since $t = 1$, stop.

This algorithm yields the information that the longest route from node 1 to node 7 has length 14, the longest route from node 2 to node 7 has length 12, and so on. To find the choice of arcs that corresponds to the longest route, we must apply the route selection algorithm:

a. Set $t = 1$, $d_1(1) = 2$, and $u = 2$.
b. Set $d_2(2) = 5$ and $u = 5$.
c. Since $t + 1 \neq 3$, continue. Set $t = 2$, $d_3(5) = 7$, and $u = 7$.
d. Since $t + 1 = 3$, stop.

This procedure gives the longest path through the network, namely, $1 \to 2 \to 5 \to 7$, which can easily be seen to have length 14. Note that choosing the myopic policy at each node would not have been optimal. At node 1, the myopic policy would have selected action 3; at node 3, action 4; and at node 4, action 7. The path $1 \to 3 \to 4 \to 7$ has length 12. By not taking future consequences into account at the first stage, the decision maker would have found himself in a poor position for subsequent decisions.

The backward induction algorithm has also obtained an optimal policy. It is given by

$$\pi^* = (d_1^*, d_2^*, d_3^*),$$

where

$$\begin{aligned} &d_1^*(1) = 2, \qquad d_2^*(2) = 5, \qquad d_2^*(3) = 5, \\ &d_3^*(4) = 7, \qquad d_3^*(5) = 7, \qquad \text{and} \qquad d_3^*(6) = 7. \end{aligned}$$

This policy provides the longest route from each node to node 7, as promised by the principle of optimality. In the language of graph theory, this corresponds to a maximal spanning tree. The longest route from each node to node 7 is depicted in Fig. 4.

2. A Resource Allocation Problem

The backward induction algorithm is now applied to the continuous version of the resource allocation problem of Section 1. Computation in the discrete problem is almost identical to that in the longest-route problem and will be left to the reader.

The bounds on the s_i's are changed to simplify exposition. The problem that will be solved is given by

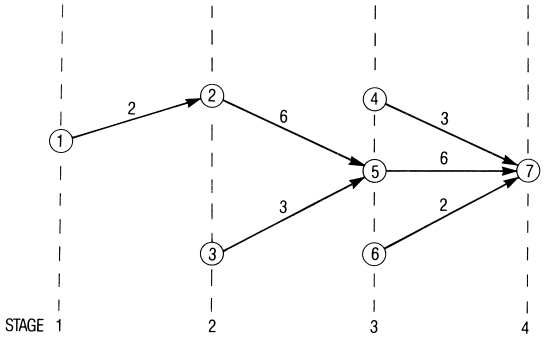


FIGURE 4 Solution to the longest-route problem.

Maximize $3s_1^2 + s_2^3 + 4s_3$

subject to

$$s_1 + s_2 + s_3 \le 4$$

and

$$0 \le s_1 \le 3; \qquad 0 \le s_2; \qquad 0 \le s_3.$$

The backward induction is applied as follows:

a. Set $t = 4$ and $v^4(s) = 0$, $0 \le s \le 4$.
b. Since $t \ne 1$, continue. Set $t = 3$ and

$$v^3(s) = \max_{0 \le a \le s} \{r_3(s, a) + v^4(s - a)\}$$
$$= \max_{0 \le a \le s} \{4a\} = 4s$$

and $A^*_{s,3} = \{s\}$.
c. Since $t \ne 1$, continue. Set $t = 2$ and

$$v^2(s) = \max_{0 \le a \le s} \{r_2(s, a) + v^3(s - a)\}$$
$$= \max_{0 \le a \le s} \{a^3 + 4(s - a)\},$$
$$v^2(s) = \begin{cases} 4s, & 0 \le s \le 2 \\ s^3, & 2 \le s \le 4, \end{cases}$$

and

$$A^*_{s,2} = \begin{cases} \{0\}, & 0 \le s \le 2 \\ \{s\}, & 2 \le s \le 4. \end{cases}$$

d. Since $t \ne 1$, continue. Set $t = 1$. A solution is obtained for $v^1(4)$ only. Obviously it is optimal to allocate all resources in this problem. In most problems one would obtain a $v^1(s)$ for all s. That is quite tedious in this example and unnecessary for solution of the original problem.

$$v^1(4) = \max_{0 \le a \le 3} \{r_1(s, a) + v^2(4 - a)\}$$
$$= \max \left\{ \max_{\substack{0 \le a \le 3 \\ 0 \le 4 - a \le 2}} \{3a^2 + 4(4 - a)\}, \right.$$
$$\left. \max_{\substack{0 \le a \le 3 \\ 2 \le 4 - a}} \{3a^2 + 4(4 - a)^3\} \right\}$$
$$= \max \left\{ \max_{2 \le a \le 3} \{3a^2 + 4(4 - a)\}, \right.$$
$$\left. \max_{0 \le a \le 2} \{3a^2 + (4 - a)^3\} \right\}$$
$$= \max\{31, 30\} = 31$$

and $A^*_{4,1} = \{3\}$.
e. Since $t = 1$, stop.

The objective function value for this constrained resource allocation problem is 31. To find the optimal resource allocation, we must use the second algorithm above:

a. Set $t = 1$, $d_1(4) = 3$, and $u = 1$.
b. Set $d_2(1) = 0$ and $u = 1$.
c. Since $t + 1 \ne 3$, continue. Set $t = 2$, $d_3(1) = 1$, and $u = 0$.
d. Since $t + 1 = 3$, stop.

This procedure gives the optimal allocation, namely, $s_1 = 3$, $s_2 = 0$, and $s_3 = 1$, which corresponds to the optimal value of 31.

3. The Inventory Example

The backward induction algorithm is now applied to the numerical stochastic inventory example of Section I. Since the data are stationary, the time index will be deleted. Also, the following additional notation will be useful. Define $v^t(s, a)$ by

$$v^t(s, a) = r(s, a) + \sum_{j \in S} p(j|s, a) v^{t+1}(j). \tag{22}$$

a. Set $t = 4$ and $v^4(s) = 0$, $s = 0, 1, 2, 3$.
b. Since $t \ne 1$, continue. Set $t = 3$, and for $s = 0, 1, 2, 3$,

$$v^3(s) = \max_{a \in A_s} \left\{ r(s, a) + \sum_{j \in S} p(j|s, a) v^4(j) \right\}$$
$$= \max_{a \in A_s} \{r(s, a)\}. \tag{23}$$

It is obvious from inspecting the values of $r(s, a)$ that in each state the maximizing action is 0; that is, do not order. Thus,

s	$v^3(s)$	$A^*_{s,3}$
0	0	0
1	5	0
2	6	0
3	5	0

c. Since $t \neq 1$, continue. Set $t = 2$ and

$$v^2(s) = \max_{a \in A_s}\{v^2(s, a)\},$$

where, for instance,

$$\begin{aligned} v^2(0, 2) &= r(0, 2) + p(0|0, 2)v^3(0) + p(1|0, 2)v^3(1) \\ &\quad + p(2|0, 2)v^3(2) + p(3|0, 2)v^3(3) \\ &= -2 + \left(\tfrac{1}{4}\right) \times 0 + \left(\tfrac{1}{2}\right) \times 5 + \left(\tfrac{1}{4}\right) \times 6 + \\ &\quad 0 \times 5 = 2. \end{aligned}$$

The quantities $v^2(s, a)$, $v^2(s)$, and $A^*_{s,2}$ are summarized in the following tabulation, where Xs denote infeasible actions:

	$v^2(s, a)$					
s	$a = 0$	$a = 1$	$a = 2$	$a = 3$	$v^2(s)$	$A^*_{s,2}$
0	0	$\frac{1}{4}$	2	$\frac{1}{2}$	2	2
1	$6\frac{1}{4}$	4	$2\frac{1}{2}$	X	$6\frac{1}{4}$	0
2	10	$4\frac{1}{2}$	X	X	10	0
3	$10\frac{1}{2}$	X	X	X	$10\frac{1}{2}$	0

d. Since $t \neq 1$, continue. Set $t = 1$ and

$$v^1(s) = \max_{a \in A_s}\{v^1(s, a)\}.$$

The quantities $v^1(s, a)$, $v^1(s)$, and $A^*_{s,1}$ are summarized in the following tabulation, where Xs denote infeasible actions:

	$v^1(s, a)$					
s	$a = 0$	$a = 1$	$a = 2$	$a = 3$	$v^1(s)$	$A^*_{s,1}$
0	2	$\frac{33}{16}$	$\frac{66}{16}$	$\frac{67}{16}$	$\frac{67}{16}$	3
1	$\frac{129}{16}$	$\frac{98}{16}$	$\frac{99}{16}$	X	$\frac{129}{16}$	0
2	$\frac{194}{16}$	$\frac{131}{16}$	X	X	$\frac{194}{16}$	0
3	$\frac{227}{16}$	X	X	X	$\frac{227}{16}$	0

e. Since $t = 1$, stop.

This procedure has produced the optimal reward function $v^*_3(s)$ and optimal policy $\pi^* = (d^*_1(s), d^*_2(s), d^*_3(s))$, which are as follows:

s	$d^*_1(s)$	$d^*_2(s)$	$d^*_3(s)$	$v^*_3(s)$
0	3	2	0	$\frac{67}{16}$
1	0	0	0	$\frac{129}{16}$
2	0	0	0	$\frac{194}{16}$
3	0	0	0	$\frac{227}{16}$

This policy has a particularly simple form: If at decision point 1 the inventory in stock is 0 units, order 3 units; otherwise, do not order. If at decision point 2 the inventory in stock is 0 units, order 2 units; otherwise, do not order. And at decision point 3 do not order. The quantity $v^*_3(s)$ gives the expected total reward obtained by using this policy when the inventory before the first decision epoch is s units.

A policy of this type is called an (s, S) policy. An (s, S) policy is implemented as follows. If in period t the inventory level is s^t units or below, order the number of units required to bring the inventory level up to S^t units. Under certain convexity and linearity assumptions on the cost functions, Scarf showed in an elegant and important 1960 paper that (s, S) policies are optimal for the stochastic inventory problem. His proof of this result is based on using backward induction to show analytically that, for each t, $v^t(s)$ is K-convex, which ensures that there exists a maximizing policy of (s, S) type. This important result plays a fundamental role in stochastic operations research and has been extended in several ways.

III. INFINITE-HORIZON DYNAMIC PROGRAMMING

A. Introduction

Solution methods for infinite-horizon sequential decision problems are based on solving a stationary version of the functional equation. In this section, the state and action sets, the rewards, and the transition probabilities are assumed to be stationary and only the stochastic version of the problem is considered. Reward streams are summarized using the expected total discounted reward criterion, and the objective is to find a policy with maximal expected total discounted reward as well as this value. The two main solution techniques are value iteration and policy iteration. The former is the extension of backward induction to infinite-horizon problems and is best analyzed from the perspective of obtaining a fixed point for the Bellman equation. Policy iteration corresponds to using a generalization of Newton's method for finding a zero of the functional equation. It is not appropriate for finite-horizon problems. Other solution methods include modified policy iteration, linear programming, Gauss-Seidel

iteration, successive overrelaxation, and extrapolation. Of these, only modified policy iteration and linear programming will be discussed.

B. Basic Results

The basic data of the stationary, infinite-horizon, stochastic sequential decision model are the state space S; the set of allowable actions in state s, A_s; the one-period expected reward if action a is selected in state s, $r(s, a)$; the probability the system is in state j at the next stage if it is in state s and action a is selected, $p(j|s, a)$; and the discount factor λ, $0 \leq \lambda < 1$. Both S and A_s are assumed to be finite. Let M denote the number of elements in S.

For a policy $\pi = (d_1, d_2, \ldots)$, the infinite-horizon expected total discounted reward is denoted by $v_\lambda^\pi(s)$. Let $p_\pi^m(j|s) = P\{X_{m+1}^\pi = j | X_1^\pi = s\}$. For $m = 2$ it can be computed by

$$p_\pi^2(j|s) = \sum_{k \in S} p(j|k, d_2(k)) p(k|s, d_1(s)).$$

This is the matrix product of the transition probability matrices corresponding to the decision rules $d_1(s)$ and $d_2(s)$. In general $p_\pi^m(j|s)$ is given by the matrix product of the matrices corresponding to $d_1, d_2, \ldots, d_m$. Using this notation, we can compute $v_\lambda^\pi(s)$ by

$$\begin{aligned} v_\lambda^\pi(s) = r(s, d_1(s)) &+ \sum_{j \in S} \lambda p(j|s, d_1(s)) r(j, d_2(j)) \\ &+ \sum_{j \in S} \lambda^2 p_\pi^2(j|s) r(j, d_3(j)) + \cdots . \end{aligned} \tag{24}$$

Equation (24) cannot be implemented since an arbitrary nonstationary infinite-horizon policy cannot be completely specified. However, if the rewards are bounded, the above infinite series is convergent, because $\lambda < 1$. When the policy is stationary, Eq. (24) simplifies so that a policy can be evaluated either inductively or by solution of a system of linear equations. Let d denote the stationary policy that uses decision rule $d(s)$ at each stage and $p_d^m(j|s)$ its corresponding m-step transition probabilities. Then,

$$\begin{aligned} v_\lambda^d(s) &= r(s, d(s)) + \sum_{j \in S} \lambda p(j|s, d(s)) r(j, d(j)) \\ &\quad + \sum \lambda^2 p_d^2(j|s) r(j, d(j)) \\ &\quad + \sum_{j \in S} \lambda^3 p_d^3(j|s) r(j, d(j)) + \cdots \qquad (25) \\ &= r(s, d(s)) + \sum_{j \in S} \lambda p(j|s, d(s)) v_\lambda^d(j). \qquad (26) \end{aligned}$$

Equation (26) is derived from Eq. (25) by explicitly writing out the matrix powers, factoring out the term $p(j|s, d(s))$ from all summations, and recognizing that the remaining terms are exactly $v_\lambda^d(s)$ Equation (26) can be rewritten as

$$\sum_{j \in S} [\delta(j, s) - \lambda p(j|s, d(s)] v_\lambda^d(j) = r(s, d(s)),$$

where $\delta(j, s) = 1$ if $j = s$ and 0 if $j \neq s$. This equation can be re-expressed in matrix terms as

$$(I - \lambda P_d) \mathbf{v}_\lambda^\mathbf{d} = \mathbf{r}_\mathbf{d}, \tag{27}$$

where P_d is the matrix with entries $p(j|s, d(s))$, I is the identity matrix, and $\mathbf{r}_\mathbf{d}$ is the vector with elements $r(s, d(s))$. Equation (27) has a unique solution, which can be obtained by Gaussian elimination, successive approximations, or any other numerical method. A consequence of this equation is the following convenient representation for $v_\lambda^d(s)$:

$$\mathbf{v}_\lambda^\mathbf{d} = (I - \lambda P_d)^{-1} \mathbf{r}_\mathbf{d}. \tag{28}$$

The inverse in Eq. (28) exists because P_d is a probability matrix, so that its spectral radius is less than or equal to 1 and $0 \leq \lambda < 1$.

The functional equation of infinite-horizon discounted dynamic programming is the stationary equivalent of Eq. (17). It is given by

$$\begin{aligned} v(s) &= \max_{a \in A_s} \left\{ r(s, a) + \sum_{j \in S} \lambda p(j|s, a) v(j) \right\} \\ &\doteq T v(s). \end{aligned} \tag{29}$$

Equation (29) defines the nonlinear operator T on the space of bounded M vectors or real-valued functions on S. Since T is a contraction mapping (see Section III.C.1), this equation has a unique solution. This solution is the expected discounted reward of an optimal policy. To see this, let $v^*(s)$ denote the solution of this equation. Then for any decision rule $d_t(s)$,

$$v^*(s) \geq r(s, d_t(s)) + \sum_{j \in S} \lambda p(j|s, d_t(s)) v^*(j). \tag{30}$$

By repeatedly substituting the above inequality into the right-hand side of Eq. (30) and noting that $\lambda^n \to 0$ as $n \to \infty$, we can see that what follows from Eq. (24) is that $v^*(s) \geq v_\lambda^\pi(s)$ for any policy π. Also, if $d^*(s)$ is the decision rule that satisfies

$$\begin{aligned} r(s, d^*(s)) &+ \sum_{j \in S} \lambda p(j|s, d^*(s)) v^*(j) \\ &= \max \left\{ r(s, a) + \sum_{j \in S} \lambda p(j|s, a) v^*(j) \right\}, \end{aligned} \tag{31}$$

then the stationary policy that uses $d^*(s)$ each period is optimal. This follows because if d^* satisfies Eq. (31), then

$$v^*(s) = r(s, d^*(s)) + \sum_{j \in S} \lambda p(j|s, d^*(s)) v^*(j),$$

but since this equation has a unique solution,

$$v^*(s) = v_\lambda^{d^*}(s) = v_\lambda^*(s).$$

These results are summarized as follows. There exists an optimal policy to the infinite-horizon discounted stochastic sequential decision problem that is stationary and can be found by using Eq. (31). Its value function is the unique solution of the functional equation of discounted dynamic programming.

This result plays the same role as the principle of optimality in finite-horizon dynamic programming. It says that for us to solve the infinite-horizon dynamic programming problem, it is sufficient to obtain a solution to the functional equation. In the next section, methods for solving the functional equation will be demonstrated.

C. Computational Methods

The four methods to be discussed here are value iteration, policy iteration, modified policy iteration, and linear programming. Only the first three are applied to the example in Section III.D. Value iteration and modified policy iteration are iterative approximation methods for solving the dynamic programming functional equation, whereas policy iteration is exact. To study the convergence of an approximation method, we must have a notion of distance. If $\mathbf{v}$ is a real-valued function on S (an M vector), the norm of $\mathbf{v}$, denoted by $\|\mathbf{v}\|$ is defined as

$$\|\mathbf{v}\| = \max_{s \in S} |v(s)|.$$

The distance between two vectors $\mathbf{v}$ and $\mathbf{u}$ is given by $\|\mathbf{v} - \mathbf{u}\|$. This means that two vectors are ε units apart if the maximum difference between any two components is ε units. This is often called the L^∞ norm.

A policy π is said to be ε-optimal if $\|\mathbf{v}_\lambda^\pi - \mathbf{v}_\lambda^*\| < \varepsilon$. If ε is specified sufficiently small, the two iterative algorithms can be used to find policies whose expected total discounted reward is arbitrarily close to optimum. Of course, the more accurate the approximation, the more iterations of the algorithm that are required.

1. Value Iteration

Value iteration, or successive approximation, is the direct extension of backward induction to infinite-horizon problems. It obtains an ε-optimal policy d^ε as follows:

a. Select $\mathbf{v}^0$, specify $\varepsilon > 0$, and set $n = 0$.
b. Compute $\mathbf{v^{n+1}}$ by

$$v^{n+1}(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda p(j|s,a) v^n(j) \right\}. \quad (32)$$

c. If $\|\mathbf{v^{n+1}} - \mathbf{v^n}\| < \varepsilon(1-\lambda)/2\lambda$, go to step d. Otherwise, increment n by 1 and return to step b.
d. For each $s \in S$, set $d^\varepsilon(s)$ equal to an $a \in A_s$ that obtains the maximum on the right-hand side of Eq. (32) at the last iteration and stop.

This algorithm can best be understood in vector space notation. In Eq. (29), the operator T is defined on the set of bounded real-valued M vectors. Solving the functional equation corresponds to finding a fixed point of T, that is, a $\mathbf{v}$ such that $\mathbf{Tv} = \mathbf{v}$. The value iteration algorithm starts with an arbitrary $\mathbf{v^0}$ (0 is usually a good choice) and iterates according to $\mathbf{v^{n+1}} = T\mathbf{v^n}$. Since T is a contraction mapping, that is,

$$\|T\mathbf{v} - T\mathbf{u}\| \leq \boldsymbol{\lambda}\|\mathbf{v} - \mathbf{u}\|$$

for any $\mathbf{v}$ and $\mathbf{u}$, the iterative method is convergent for any v^0. This is because

$$\|\mathbf{v^{n+1}} - \mathbf{v^n}\| \leq \boldsymbol{\lambda}^n \|\mathbf{v^1} - \mathbf{v^0}\|,$$

and the space of bounded real-valued M vectors is a Banach space (a complete normed linear space) with respect to the norm used here. Since a contraction mapping has a unique fixed point, $\mathbf{v}_\lambda^*$, $\mathbf{v_n}$ converges to it. The rate of convergence is geometric with parameter λ, that is,

$$\|\mathbf{v^n} - \mathbf{v}^*\| \leq \boldsymbol{\lambda}^n \|\mathbf{v^0} - \mathbf{v}^*\|.$$

The algorithm terminates with a value function $\mathbf{v_{n+1}}$ and a decision rule $\mathbf{d}^\varepsilon$ with the following property:

$$v^{n+1}(s) = r(s, d^\varepsilon(s)) + \sum_{j \in S} \lambda p(j|s, d^\varepsilon(s)) v^n(j). \quad (33)$$

The stopping rule in step c ensures that the stationary policy that uses $\mathbf{d}^\varepsilon$ every period is ε-optimal.

The sequence of iterates $\mathbf{v^n}$ have interesting interpretations. Each iterate corresponds to the optimal expected total discounted return in an n-period problem in which the terminal reward equals $\mathbf{v^0}$. Alternatively, they correspond to the expected total discounted returns for the policy in an n-period problem that is obtained by choosing a maximizing action in each state at each iteration.

2. Policy Iteration

Policy iteration, or approximation in the policy space, is an algorithm that uses the special structure of infinite-horizon stationary dynamic programming problems to find all optimal policies. The algorithm is as follows:

a. Select a decision rule $d^0(s)$ for all $s \in S$ and set $n = 0$.

b. Solve the system of equations

$$\sum_{j\in S}[\delta(j,s)-\lambda p(j|s,d^n(s))]v^n(j)=r(s,d^n(s)) \quad (34)$$

for $v^n(s)$.

c. For each s, and each $a\in A_s$, compute

$$r(s,a)+\sum_{j\in S}\lambda p(j|s,a)v^n(j). \quad (35)$$

For each s, put a in $A^*_{n,s}$ if a obtains the maximum value in Eq. (35).

d. If, for all s, $d^n(s)$ is contained in $A^*_{n,s}$, stop. Otherwise, proceed.

e. Set $d^{n+1}(s)$ equal to any a in $A^*_{n,s}$ for each s in S, increment n by 1, and return to step b.

The algorithm consists of two main parts: step b, which is called policy evaluation, and step c, which is called policy improvement. The algorithm terminates when the set of maximizing actions found in the improvement stage repeats, that is, if the same decision obtains the maximum in step b on two successive passes through the iteration loop.

This algorithm terminates in a *finite* number of iterations with an optimal stationary policy and its expected total discounted reward. This is because the improvement procedure guarantees that $v^{n+1}(s)$ is strictly greater than $v^n(s)$, for some $s\in S$, until the termination criterion is satisfied, at which point $v^n(s)$ is the solution of the dynamic programming functional equation. Since each $\mathbf{v^n}$ is the expected total discounted reward of the stationary policy $\mathbf{d^n}$, and there are only finitely many stationary policies, the procedure must terminate in a finite number of iterations.

If only an ε-optimal policy is desired, a stopping rule similar to that in step c of the value iteration procedure can be used.

3. Modified Policy Iteration

The evaluation step of the policy iteration algorithm is usually implemented by solving the linear system

$$(I-\lambda P_{d^n})\mathbf{v^{d^n}}=\mathbf{r_{d^n}} \quad (36)$$

by using Gaussian elimination, which requires $\frac{1}{3}M^3$ multiplications and divisions. When the number of states is large, exact solution of Eq. (34) can be computationally prohibitive. An alternative is to use successive approximations to obtain an approximate solution. This is the basis of the modified policy iteration, or value-oriented successive approximation, method. The modified policy iteration algorithm of order m is as follows:

a. Select $\mathbf{v^0}$, specify $\varepsilon>0$, and set $n=0$.

b. For each s and each $a\in A_s$, compute

$$r(s,a)+\sum_{j\in S}\lambda p(j|s,a)v^n(j). \quad (37)$$

For each s, put a in $A^*_{n,s}$ if a obtains the maximum value in Eq. (37).

c. For each s in S, set $d^n(s)$ equal to any a in $A^*_{n,s}$.

(i) Set $k=0$ and define $u^0(s)$ by

$$u^0(s)=\max_{a\in A_s}\left\{r(s,a)+\sum_{j\in S}\lambda p(j|s,a)v^n(j)\right\}. \quad (38)$$

(ii) If $\|\mathbf{u^0}-\mathbf{v^n}\|<\varepsilon(1-\lambda)/2\lambda$, go to step d. Otherwise, go to step (iii).

(iii) Compute u^{k+1} by

$$u^{k+1}(s)=r(s,d^n(s))+\sum_{j\in S}\lambda p(j|s,d^n(s))u^k(j). \quad (39)$$

(iv) If $k=m$, go to step (v). Otherwise, increment k by 1 and return to step (i).

(v) Set $\mathbf{v^{n+1}}=\mathbf{u^{m+1}}$, increment k by 1, and go to step b.

d. For each $s\in S$, set $d^\varepsilon(s)=d^n(s)$ and stop.

This algorithm combines features of both policy iteration and value iteration. Like value iteration, it is an iterative algorithm that terminates with an ε-optimal policy; however, value iteration avoids step c above. The stopping criterion used in step (ii) is identical to that of value iteration, and the computation of $\mathbf{u^0}$ in step (i) requires no extra work because it has already been determined in step b. Like policy iteration, the algorithm contains an improvement step, step b, and an evaluation step, step c. However, the evaluation is not done exactly. Instead, it is carried out iteratively in step c, which is repeated m times. Note that m can be selected in advance or adaptively during the algorithm. For instance, m can be chosen so that $\|\mathbf{u^{m+1}}-\mathbf{u^m}\|$ is less than some prespecified tolerance that can vary with n. Recent studies have shown that low orders of m work well, while adaptive choice is better.

The modified policy iteration algorithm serves as a bridge between value iteration and policy iteration. When $m=0$, it is equivalent to value iteration, and when m is infinite, it is equivalent to policy iteration. It will converge in fewer iterations than value iteration and more iterations than policy iteration; however, the computational effort per iteration exceeds that for value iteration and is less than that for policy iteration. When the number of states, M, is large, it has been shown to be the most computationally efficient method for solving Markov decision problems.

4. Linear Programming

The stationary infinite-horizon discounted stochastic sequential decision problem can be formulated and solved by linear programming. The primal problem is given by

Minimize

$$\sum_{j\in S} \alpha_j v(j)$$

subject to, for $a \in A_s$ and $s \in S$,

$$v(s) \geq r(s,a) + \sum_{j\in S} \lambda p(j|s,a)v(j),$$

and $v(s)$ is unconstrained.

The constants α_j are positive and arbitrary. The dual problem is given by

Maximize

$$\sum_{s\in S}\sum_{a\in A_s} x(s,a)r(s,a)$$

subject to, for $J \in S$,

$$\sum_{a\in A_s} x(j,a) - \sum_{s\in S}\sum_{a\in A_s} \lambda p(j|s,a)x(s,a) = \alpha_j,$$

and $x(j,a) \geq 0$ for $a \in A_j$ and $j \in S$.

Using a general-purpose linear programming code for solving dynamic programming problems is not computationally attractive. The dynamic programming methods are more efficient. The interest in the linear programming formulation is primarily theoretical but allows inclusion of side constraints. Some interesting observations are as follows:

a. The dual problem is always feasible and bounded. Any optimal basis has the property that for each $s \in S$, $x(s,a) > 0$ for only one $a \in A_s$. An optimal stationary policy is given by $d^*(s) = a$ if $x(s,a) > 0$.

b. The same basis is optimal for all α_j's.

c. When the dual problem is solved by the simplex algorithm with block pivoting, it is equivalent to policy iteration.

d. When policy iteration is implemented by only changing the action that gives the maximum improvement over all states, it is equivalent to solving the dual problem by the simplex method.

D. Numerical Examples

In this section, an infinite-horizon version of the stochastic inventory example presented earlier is solved by using value iteration, policy iteration, and modified policy iteration. The data are as analyzed in the finite-horizon case; however, the discount rate λ is chosen to be .9. The objective is to determine the stationary policy that maximizes the expected total infinite-horizon discounted reward.

1. Value Iteration

To initiate the algorithm, we will take $\mathbf{v^0}$ to be the zero vector; ε is chosen to be .1. The algorithm will terminate with a stationary policy that is guaranteed to have an expected total discounted reward within .1 of optimal. Calculations proceed as in the finite-horizon backward induction algorithm until the stopping criterion of

$$\|\mathbf{v^{n+1}} - \mathbf{v^n}\| \leq \frac{\varepsilon(1-\lambda)}{2\lambda} = \frac{.1\times.1}{2\times.9} = .0056$$

is satisfied. The value functions $\mathbf{v^n}$ and the maximizing actions obtained in step b at each iteration are provided in the tabulation on the following page. The above algorithm terminates after 58 iterations, at which point $\|\mathbf{v^{58}} - \mathbf{v^{57}}\| = .0054$. The .1-optimal stationary policy is $\mathbf{d}^\varepsilon = (3,0,0,0)$, which means that if the stock level is 0, order 3 units; otherwise, do not order. Observe that the optimal policy was first identified at iteration 3, but the algorithm did not terminate until iteration 58. In larger problems such additional computational effort is extremely wasteful. Improved stopping rules and more efficient algorithms are described in Section III.E.

2. Policy Iteration

To initiate policy iteration, choose the myopic policy, namely, that which maximizes the immediate one-period reward $r(s,a)$. The algorithm proceeds as follows:

a. Set $d^0 = (0,0,0,0)$ and $n = 0$.

b. Solve the evaluation equations:

$$\begin{aligned}
(1-.9\times1)v^0(0) &= 0,\\
(-.9\times.75)v^0(0) + (1-.9\times25)v^0(1) &= 5,\\
(-.9\times.25)v^0(0) + (-.9\times.5)v^0(1) &\\
+ (1-.9\times25)v^0(2) &= 6,
\end{aligned}$$

and

$$\begin{aligned}
(-.9\times.25)v^0(1) + (-.9\times.50)v^0(2) &\\
+ (1-.9\times.25)v^0(3) &= 5.
\end{aligned}$$

These equations are obtained by substituting the transition probabilities and rewards corresponding to policy $\mathbf{d^0}$ into Eq. (34). The solution of these equations is $\mathbf{v^0} = (0, 6.4516, 11.4880, 14.9951)$.

c. For each s, the quantities

$$r(s,a) + \sum_{j=0}^{3} \lambda p(j|s,a)v^0(j)$$

are computed for $a = 0,\ldots,3-s$, and the actions that achieve the maximum are placed into $A^*_{0,s}$. In this example

	$v^n(s)$				$d^n(s)$			
n	$s=0$	$s=1$	$s=2$	$s=3$	$s=0$	$s=1$	$s=2$	$s=3$
0	0	0	0	0	0	0	0	0
1	0	5.0	6.0	5.0	2	0	0	0
2	1.6	6.125	9.6	9.95	2	0	0	0
3	3.2762	7.4581	11.2762	12.9368	3	0	0	0
4	4.6632	8.8895	12.6305	14.6632	3	0	0	0
5	5.9831	10.1478	13.8914	15.9831	3	0	0	0
10	10.7071	14.8966	18.6194	20.7071	3	0	0	0
15	13.5019	17.6913	21.4142	23.0542	3	0	0	0
30	16.6099	20.7994	24.5222	26.6099	3	0	0	0
50	17.4197	21.6092	25.3321	27.4197	3	0	0	0
56	17.4722	21.6617	25.3845	27.4722	3	0	0	0
57	17.4782	21.6676	25.3905	27.4782	3	0	0	0
58	17.4736	21.6730	25.3959	27.4836	3	0	0	0

there is a unique maximizing action in each state, and it is given by

$$A^*_{0,0} = \{3\}; \qquad A^*_{0,1} = \{2\};$$
$$A^*_{0,2} = \{0\}; \qquad A^*_{0,3} = \{0\}.$$

d. Since $d^0(0) = 0$, it is not contained in $A^*_{0,0}$, so continue.

e. Set $\mathbf{d^1} = (3, 2, 0, 0)$ and $n = 1$, and return to the evaluation step.

The detailed step-by-step calculations for the remainder of the algorithm are omitted. The value functions and corresponding maximizing actions are presented below. Since there is a unique maximizing action in the improvement step at each iteration, $A^*_{n,s}$ is equivalent to $d^n(s)$ and only the latter is displayed.

The algorithm terminates in three iterations with the optimal policy $\mathbf{d}^* = (3, 0, 0, 0)$. Observe that an evaluation was unnecessary at iteration 3 since $\mathbf{d^2} = \mathbf{d^3}$ terminated the algorithm before the evaluation step. Unlike value iteration, the algorithm has produced an optimal policy as well as its expected total discounted reward $\mathbf{v^3}$, which is the optimal expected total discounted reward. This computation shows that the .1-optimal policy found by using value iteration is in fact optimal. This information could not be obtained by using value iteration unless the action elimination method described in Section III.E were used.

3. Modified Policy Iteration

The following illustrates the application of modified policy iteration of order 5. The first pass through the algorithm is described in detail; calculations for the remainder are presented in tabular form below.

a. Set $\mathbf{v^0} = (0, 0, 0, 0)$, $n = 0$, and $\varepsilon = .1$.

b. Observe that

$$r(s, a) + \sum_{j=0}^{3} \lambda p(j|s, a) v^0(j) = r(s, a),$$

so that for each s the maximum value occurs for $a = 0$. Thus, $A^*_{n,s} = \{0\}$ for $s = 0, 1, 2, 3$ and $\mathbf{d^n} = (0, 0, 0, 0)$.

(i) Set $k = 0$ and $\mathbf{u^0} = (0, 5, 6, 5)$.

(ii) Since $\|\mathbf{u^0} = \mathbf{v^0}\| = 6 > .0056$, continue.

(iii) Compute $\mathbf{u^1}$ by

$$u^1(s) = r(s, d^0(s)) + \sum_{j=0}^{3} \lambda p(j|s, d^0(s)) u^0(j)$$
$$= r(s, 0) + \sum_{j=0}^{3} \lambda p(j|s, 0) u^0(j) \qquad (40)$$
$$= 0 + .9 \times 1 \times 0 = 0, \qquad \text{for} \quad s = 0,$$
$$= 5 + .9 \times \tfrac{3}{4} \times 0 + .9 \times \tfrac{1}{4} \times 5 = 6.125, \quad \text{for} \quad s = 1,$$
$$= 6 + .9 \times \tfrac{1}{4} \times 0 + .9 \times \tfrac{1}{2} \times 5 + .9 \times \tfrac{1}{4} \times 6 = 9.60, \quad \text{for} \quad s = 2,$$
$$= 6 + 9 \times \tfrac{1}{4} \times 5 + .9 \times \tfrac{1}{2} \times 6 + .9 \times \tfrac{1}{4} \times 5 = 10.95, \quad \text{for} \quad s = 3,$$

so that $\mathbf{u^1} = (0, 6.125, 9.60, 10.95)$.

(iv) Since $k = 1 < 5$, continue.

	$v^n(s)$				$d^n(s)$			
n	$s=0$	$s=1$	$s=2$	$s=3$	$s=0$	$s=1$	$s=2$	$s=3$
0	0	6.4516	11.4880	14.9951	0	0	0	0
1	10.7955	12.7955	18.3056	20.7955	3	2	0	0
2	17.5312	21.7215	25.4442	27.5318	3	0	0	0
3	X	X	X	X	3	0	0	0

	$v^n(s)$				$d^n(s)$			
n	$s=0$	$s=1$	$s=2$	$s=3$	$s=0$	$s=1$	$s=2$	$s=3$
0	0	0	0	0	0	0	0	0
1	0	6.4507	11.4765	14.9200	3	2	0	0
2	7.1215	9.1215	14.6323	17.1215	3	0	0	0
3	11.5709	15.7593	19.4844	21.5709	3	0	0	0
4	14.3639	18.5534	22.2763	24.3639	3	0	0	0
5	15.8483	20.0377	23.7606	25.8483	3	0	0	0
10	17.4604	21.6499	25.3727	27.4604	3	0	0	0
11	17.4938	21.6833	25.4062	27.4938	3	0	0	0

The loop is repeated four more times to evaluate $\mathbf{u^2}$, $\mathbf{u^3}$, $\mathbf{u^4}$, and $\mathbf{u^5}$. Then $\mathbf{v^1}$ is set equal to $\mathbf{u^5}$ and the maximization in step b is carried out. The resulting iterates are shown at the bottom of the page. In step (ii), following iteration 11, the computed value of $\mathbf{u^0}$ is (17.4976, 21.6871, 25.4100, 27.4976), so that $\|\mathbf{u^0} - \mathbf{v^{11}}\| = .0038$, which guarantees that the policy (3, 0, 0, 0) is ε-optimal with $\varepsilon = .1$.

4. Comparison of the Algorithms

Value iteration required 58 iterations to obtain a .1-optimal solution. At each iteration a maximization was required so that each action had to be evaluated to determine Eq. (32) at each iteration. Modified policy iteration of order 5 required 11 iterations to determine a .1-optimal policy so that the maximization step was carried out only 11 times. However, at each iteration the inner loop of the algorithm in step c was invoked five times. Thus, modified policy iteration required far fewer maximizations than did value iteration. This would lead to considerable computational savings when the action sets or the problem had many states.

Policy iteration found an optimal policy in three iterations; however, each iteration required both the evaluation of all actions in each state and the solution of a linear system of equations. In this small example, it is not time consuming to solve the linear system by using Gaussian elimination, but when the number of states is large, this can be prohibitive. In such cases, modified iteration is preferred over both policy iteration and value iteration; however, which order is the best to use is an open question.

E. Bounds on the Optimal Total Expected Discounted Reward

At each stage of the value iteration, policy iteration, and modified policy iteration, the computed value of $\mathbf{v^n}$ can be used to obtain upper and lower bounds on the optimal expected discounted reward. These bounds can be used to terminate any of the iterative algorithms, eliminate suboptimal actions at each iteration of an algorithm, and develop improved algorithms.

Bounds are given for value iteration; however, they have also been obtained for policy iteration and modified policy iteration. First, define the following two quantities:

$$L^n = \min_{s\in S}\{v^{n+1}(s) - v^n(s)\}$$

and

$$U^n = \max_{s\in S}\{v^{n+1}(s) - v^n(s)\}.$$

Then, for each n and $s \in S$,

$$v^n(s) + \frac{1}{1-\lambda}L^n \le v^{d^n}(s) \le v^*(s) \le v_\lambda^n(s) + \frac{1}{1-\lambda}U^n. \tag{41}$$

We can easily see, using the two extreme bounds, that if $U^n - L^n < \varepsilon(1-\lambda)$, then

$$0 \le v_\lambda^*(s) - \left(v^n(s) + \frac{1}{1-\lambda}L^n\right) \le \varepsilon. \tag{42}$$

This provides an alternative stopping criterion for value iteration and can be modified for any of the above algorithms. In particular in value iteration, if the algorithm is terminated when $U^n - L^n < \varepsilon(1-\lambda)$, then the quantity $v_n(s) + (1-\lambda)^{-1}L^n$ is within ε of the optimal value function. If this had been implemented in the value iteration algorithm, in Section III.C., it would have terminated after 9, as opposed to 58, iterations, when $\varepsilon(1-\lambda) = .1 \times (1 - .9) = .01$.

These bounds can be used at each iteration to eliminate actions that cannot be part of an optimal policy. This is important computationally because the maximization in the improvement step can be made more efficient if all of A_s need not be evaluated at each iteration. Elimination is based on the result that action a is suboptimal in state s, if

$$r(s,a) + \sum_{j\in S}\lambda p(j|s,a)v_\lambda^*(j) < v_\lambda^*(s). \tag{43}$$

Of course, $v_\lambda^*(s)$ is not known in Eq. (43), but by substituting an upper bound for it on the left-hand side and a lower bound on the right-hand side, one can use the result to eliminate suboptimal actions in state s. The bounds in Eq. (41) can be used.

Action elimination procedures are especially important in approximation algorithms such as value iteration and modified policy iteration that produce only ε-optimal policies. If a unique optimal policy exists, by eliminating suboptimal actions at each iteration, one obtains an optimal policy when only one action remains in each state. This will occur in finitely many iterations.

F. Turnpike and Planning-Horizon Results

Infinite-horizon sequential decision models are usually approximations to long finite-horizon problems with many decision points. A question of practical concern is, *When is the optimal policy for the infinite-horizon model optimal for the finite-horizon problem?* An answer to this question is provided through planning-horizon, or turnpike, theory. The basic result is the following. There exists an N^*, called the planning horizon, such that, for all $n \geq N^*$, the optimal decision when there are n periods remaining is one of the decisions that is optimal when the horizon is infinite. This result means that if there is a unique optimal stationary policy d^* for the infinite-horizon problem, then in an n-period problem with $n \geq N^*$, it is optimal to use the stationary strategy for the initial $n - N^*$ decisions and to find the optimal policy for the remaining N^* periods by using the backward induction methods of Section II. The optimal infinite-horizon strategy is called the turnpike, and it is reached after traveling N^* periods on the nonstationary "side roads." The term *turnpike* originates in mathematical economics, where it refers to the policy path that produces optimal economic growth.

Another interpretation of the above result is that it is optimal to use d^* for the first decision in any finite-horizon problem in which it is known that the horizon exceeds N^*. Thus, it is not necessary to specify the horizon, only to know that there are at least N^* decisions to be made. Bounds on N^* are available, and this concept has been extended to nonstationary infinite-horizon problems and the expected average reward criteria. This is referred to as a rolling horizon approach.

The computational results for the value iteration algorithm in Section III.D give further insight into the planning-horizon result. There, $N^* = 3$, so that in any problem with horizon greater than 3, it is optimal to use the decision rule (3, 0, 0, 0) until there are three decisions left to be made, at which point it is optimal to use the decision rule (2, 0, 0, 0) for two periods and (0, 0, 0, 0) in the last period.

G. The Average Expected Reward Criteria

In many applications in which infinite-horizon formulations are natural, the total discounted reward criterion is not relevant. For instance, in a large telecommunications network, millions of packet and call routing decisions are made every second, so that discounting the consequences of latter decisions makes little sense. An alternative is the expected average reward criterion defined in Eq. (7). Using this criterion means that rewards received at each time period receive equal weight.

Computational methods for the average reward criterion are more complex than those for the discounted reward problem. This is because the form of the average reward function, $g^\pi(s)$, depends on the structure of the Markov chain corresponding to the stationary policy π. If the policy is unichain, that is, the Markov chain induced by the policy has exactly one recurrent class and possibly several transient classes, then $\mathbf{g}^\pi$ is a constant vector. In this case, the functional equation is

$$v(s) = \max_{a \in A_s} \left\{ r(s,a) - g + \sum_{s \in S} p(j|s,a) v(j) \right\}. \quad (44)$$

Solving this equation uniquely determines g and determines $v(s)$ up to an additive constant. The quantity g is the optimal expected average reward. For us to specify $\mathbf{v}$ uniquely, it is sufficient to set $v(s_0) = 0$ for some s_0 in the recurrent class of a policy corresponding to choosing a maximizing action in each state. If this is done, $v(s)$ is called a relative value function and $v(j) - v(k)$ is the difference in expected total reward obtained by using an optimal policy and starting the system in state j as opposed to state k.

As in the discounted case, an optimal policy is found by solving the functional equation. This is best done by policy iteration. The theory of value iteration is quite complex in this setting and is not discussed here. The policy iteration algorithm is given below. It is assumed that all policies are unichain.

a. Select a decision rule $\mathbf{d^0}$ and set $n = 0$.

b. Solve the system of equations

$$\sum_{j \in S} [\delta(j,s) - p(j|s, d^n(s))] v^n(j) - g^n = r(s, d^n(s)), \quad (45)$$

where $v^n(s_0) = 0$ for some s_0 in the recurrent class of d^n.

c. For each s, and each $a \in A_s$, compute

$$r(s,a) + \sum_{j \in S} p(j|s,a) v^n(j). \quad (46)$$

For each s, put a in $A^*_{n,s}$ if a obtains the maximum value in Eq. (46).

d. If for each s, $d^n(s)$ is contained in $A^*_{n,s}$, stop. Otherwise, proceed.

e. Set $d^{n+1}(s)$ equal to any a in $A^*_{n,s}$ for each s in S, increment n by 1, and return to step b.

Note that this algorithm is almost identical to that for the discounted case. The only difference is the linear system of equations solved in step b. If the assumption that all policies are unichain is dropped, solution of the functional equation [Eq. (44)] is no longer sufficient to determine an optimal policy. Instead, a nested pair of optimality equations is required.

IV. FURTHER TOPICS

A. Historical Perspective

The development of dynamic programming is usually credited to Richard Bellman. His numerous papers in the 1950s presented a formal development of this subject and numerous interesting examples. Most of this pioneering work is summarized in his book, "Dynamic Programming." However, many of the themes of dynamic programming and sequential decision processes are scattered throughout earlier works. These include studies that appeared between 1946 and 1953 on water resource management by Masse; sequential analysis in statistics by Wald; games of pursuit by Wald; inventory theory by Arrow, Blackwell, and Girshick; Arrow, Harris, and Marshak; and Dvoretsky, Kiefer, and Wolfowitz; and stochastic games by Shapley.

Although Bellman coined the phrase "Markov decision processes," this aspect of dynamic programming got off the ground with Howard's monograph, "Dynamic Programming and Markov Processes" in 1960. The first formal theoretical treatment of this subject was by Blackwell in 1962. In 1960, deGhellinck demonstrated the equivalence between Markov decision processes and linear programming. Other major contributions are those of Denardo in 1968, in which he showed that value iteration can be analyzed by the theory of contraction mappings; Veinott in 1969, in which he introduced a new family of optimality criteria for dynamic programming problems; and Federgruen and Schweitzer between 1978 and 1980, in which they investigated the properties of the sequences of policies obtained from the value iteration algorithm. Modified policy iteration is usually attributed to Puterman and Shin in 1978; however, similar ideas appeared earlier in works of Kushner and Kleinman and of van Nunen. In 1978, Puterman and Brumelle demonstrated the equivalence of policy iteration to Newton's method. Puterman's book "Markov Decision Processes" provides a comprehensive overview of theory, application, and calculations.

B. Applications

Dynamic programming methods have been applied in many areas. These methods have been used numerically to compute optimal policies, as well as analytically to determine the form of an optimal policy under various assumptions on the rewards and transition probabilities. A brief and by no means complete summary of applications appears in Table I. Only stochastic dynamic programming is considered; however, in many cases, the problems have also been analyzed in the deterministic setting. In these applications, probability distributions of the random quantities are assumed to be known before solution of the problem; adaptive estimation of parameters is not necessary.

A major limitation in the practical application of dynamic programming has been computational. When the set of states at each stage is large—for example, if the state description is vector-valued—then solving a sequential decision problem by dynamic programming requires considerable storage as well as computationl time. Bellman recognized the difficulty early on and referred to it as the "curse of dimensionality." Research in the 1990s addressed this issue by developing approximation methods for large-scale applications. These methods combined concepts from stochastic approximation, simulation, and artificial intelligence and are sometimes referred to as reinforcement learning. A comprehensive treatment of this line of research appears in "Neuro-Dynamic Programming" by Bertsekas and Tsitsiklis.

C. Extensions

In the models considered in this article, it has been assumed that the decision maker knows the state of the system before making a decision, that decisions are made at discrete time points, that the set of states is finite and discrete (with the exception of the example in Section II.D.2), and the model rewards and transition probabilities are known. These models can be modified in several ways: the state of the system may be only partially observed by the decision maker, the sets of decision points and states may be continuous, or the transition probabilities or rewards may not be known. These modifications are discussed briefly below.

1. Partially Observable Models

This model differs from the fully observable model in that the state of the system is not known to the decision maker at the time of decision. Instead, the decision maker receives a signal from the system and on the basis of this signal updates his estimate of the probability distribution of the system state. Updating is done using Bayes' theorem. Decisions are based on this probability distribution, which is a sufficient statistic for the history of the process.

When the set of states is discrete, these models are referred to as partially observable Markov decision processes. Computational methods in this case are quite

TABLE I Stochastic Dynamic Programming Applications

Area	States	Actions	Reward	Stochastic aspect
Capacity expansion	Size of plant	Maintain or add capacity	Costs of expansion and production at current capacity	Demand for product
Cash management	Cash available	Borrow or invest	Transaction costs less interest	External demand for cash
Catalog mailing	Customer purchase record	Type of catalog to send to customer, if any	Purchases in current period less mailing costs	Customer purchase amount
Clinical trials	Number of successes with each treatment	Stop or continue the trial, and if stopped, choose best treatment if any	Costs of treatment and incorrect decisions	Response of a subject to a treatment
Economic growth	State of the economy	Investment or consumption	Utility of consumption	Effect of investment
Fisheries management	Fish stock in each age class	Number of fish to harvest	Value of the catch	Population size
Football	Position of ball	Play to choose	Expected points scored	Outcome of play
Forest management	Size and condition of stand	Harvesting and reforestation activities	Revenues less harvesting costs	Stand growth and price fluctuations
Gambling	Current wealth	Stop or continue playing the game	Cost of playing	Outcome of the game
Hotel and airline reservations	Number of confirmed reservations	Accept, wait-list, or reject new reservations	Profit from satifisfied reservations less overbooking penalties	Demand for reservations and number of arrivals
Inventory control	Stock on hand	Order additional stock	Revenue per item sold less ordering, holding, and penalty costs	Demand for items
Project selection	Status of each project	Project to invest in at present	Return from investing in project	Change of project status
Queuing control	Number in the queue	Accept or reject arriving customers or control service rate	Revenue from serving customer less delay costs	Interarrival times and service times
Reliability	Age or status of equipment	Inspect and repair or replace if necessary	Inspection and repair costs plus failure cost	Failure and deterioration
Scheduling	Activites completed	Next activity to schedule	Cost of activity	Length of time to complete activity
Selling an asset	Current offer	Accept or reject the offer	The offer less the cost of holding the asset for one period	Size of the offer
Water resource management	Level of water in each reservoir in river system	Quantity of water to release	Value of power generated	Rainfall and runoff

complex, and only very small problems have been solved numerically. When the states form a continuum, this problem falls into the venue of control theory. An extremely important result in this area is the Kalman filter, which provides an updating formula for the expected value and covariance matrix of the system state. Another important result is the separation theorem, which gives conditions that allow decomposition of this problem into separate problems of estimation and control.

2. Continuous-State, Discrete-Time Models

The resource allocation problem in Section I is an example of a continuous-state, discrete-time, deterministic model. Its solution using dynamic programming methodology is given in Section II. When transitions are stochastic, only minor modifications to the general sequential decision problem are necessary. Instead of a transition probability function, a transition probability density is used, and summations are replaced by integrations throughout. This modification causes considerable theoretical complexity; the main issues concern measurability and integrability of value functions.

Problems of this type fall into the realm of stochastic control theory. Although dynamic programming is used to solve such problems, the formulation is quite different. Instead of explicitly giving a transition probability function for the state, the theory requires use of a dynamic equation

to relate the state at time $t+1$ to the state at time t. A major result in this area is that when the state dynamics are linear in the state, action, and random component and the cost is quadratic in the state and the action, then a closed-form solution is available for the optimal decision rule and it is linear in the system state. These problems have been studied extensively in the engineering literature.

3. Continous-Time Models

Stochastic continuous-time models are categorized according to whether the state space is continuous or discrete. The discrete-time model has been widely studied in the operations research literature. The stochastic nature of the problem is modeled as either a Markov process, a semi-Markov process, or a general jump process. The decision maker can control the transition rates, transition probabilities, or both. The infinite-horizon versions of the Markov and semi-Markov decision models are analyzed in a similar fashion to the discrete-time Markov decision process; however, general jump processes are considerably more complex. These models have been widely applied to problems in queuing and inventory control.

When the state space is continuous and Markovian assumptions are made, diffusion processes are used to model the transitions. The decision maker can control the drift of the system or can cause instantaneous state transitions or jumps. The discrete-time optimality equation is replaced by a nonlinear second-order partial differential equation and is usually solved numerically. These models are studied in the stochastic control theory literature and have been applied to inventory control, finance, and statistical modeling.

4. Adaptive Control

When transition probabilities and/or rewards are unknown, the decision maker must adaptively estimate them to control the system optimally. The usual approach to analysis of such systems is to assume that the rewards and transition probabilities depend on an unknown parameter, such as the arrival rate to a queuing system, and then use the observed sequence of system states to adaptively estimate this parameter. In a Bayesian analysis of such models, uncertainty about the parameter value is described through a probability distribution which is periodically updated as information becomes available. The classical approach to such models uses maximum likelihood theory to estimate the parameter and derive its statistical properties.

ACKNOWLEDGMENT

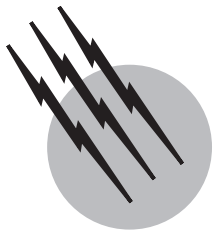
Preparation of this article was supported by Natural Sciences and Engineering Research Council Grant A-5527.

SEE ALSO THE FOLLOWING ARTICLES

LINEAR OPTIMIZATION • NONLINEAR PROGRAMMING • OPERATIONS RESEARCH

BIBLIOGRAPHY

Bellman, R. E. (1957). "Dynamic Programming," Princeton University Press, Princeton, N.J.
Bertsekas, D. P. (1995). "Dynamic Programming and Optimal Control," Vols. 1 and 2, Athena Scientific, Belmont, Mass.
Bertsekas, D. P., and Tsitsiklis, J. M. (1995). "Neuro-Dynamic Programming," Athena Scientific, Belmont, Mass.
Blackwell, D. (1962). *Ann. Math. Stat*. **35**, 719–726.
Denardo, E. V. (1967). *SIAM Rev*. **9**, 169–177.
Denardo, E. V. (1982). "Dynamic Programming, Models and Applications," Prentice-Hall, Englewood Cliffs, N.J.
Fleming, W. H., and Rishel, R. W. (1975). "Deterministic and Stochastic Optimal Control," Springer-Verlag, New York.
Howard, R. A. (1960). "Dynamic Programming and Markov Processes," MIT Press, Cambridge, Mass.
Puterman, M. L. (1994). "Markov Decision Processes," Wiley, New York.
Ross, S. M. (1983). "Introduction to Stochastic Dynamic Programming," Academic Press, New York.
Scarf, H. E. (1960). *In* "Studies in the Mathematical Theory of Inventory and Production," (K. J. Arrow, S. Karlin, and P. Suppes, eds.), Stanford University Press, Stanford, Calif.
Veinott, A. F., Jr. (1969). *Ann. Math. Stat*. **40**, 1635–1660.
Wald, A. (1947). "Sequential Analysis," Wiley, New York.
White, D. J. (1985). *Interfaces* **15**, 73–83.
White, D. J. (1988). *Interfaces* **18**, 55–61.



Fourier Series

James S. Walker
University of Wisconsin–Eau Claire

I. Historical Background
II. Definition of Fourier Series
III. Convergence of Fourier Series
IV. Convergence in Norm
V. Summability of Fourier Series
VI. Generalized Fourier Series
VII. Discrete Fourier Series
VIII. Conclusion

GLOSSARY

Bounded variation A function f has *bounded variation* on a closed interval $[a, b]$ if there exists a positive constant B such that, for all finite sets of points $a = x_0 < x_1 < \cdots < x_N = b$, the inequality $\sum_{i=1}^{N} |\mathrm{f}(x_i) - \mathrm{f}(x_{i-1})| \leq B$ is satisfied. Jordan proved that a function has bounded variation if and only if it can be expressed as the difference of two nondecreasing functions.

Countably infinite set A set is *countably infinite* if it can be put into one-to-one correspondence with the set of natural numbers $(1, 2, \ldots, n, \ldots)$. Examples: The integers and the rational numbers are countably infinite sets.

Continuous function If $\lim_{x \to c} \mathrm{f}(x) = \mathrm{f}(c)$, then the function f is *continuous* at the point c. Such a point is called a *continuity point* for f. A function which is continuous at all points is simply referred to as continuous.

Lebesgue measure zero A set S of real numbers is said to have *Lebesgue measure zero* if, for each $\epsilon > 0$, there exists a collection $\{(a_i, b_i)\}_{i=1}^{\infty}$ of open intervals such that $S \subset \cup_{i=1}^{\infty}(a_i, b_i)$ and $\sum_{i=1}^{\infty}(b_i - a_i) \leq \epsilon$. Examples: All finite sets, and all countably infinite sets, have Lebesgue measure zero.

Odd and even functions A function f is *odd* if $\mathrm{f}(-x) = -\mathrm{f}(x)$ for all x in its domain. A function f is *even* if $\mathrm{f}(-x) = \mathrm{f}(x)$ for all x in its domain.

One-sided limits $\mathrm{f}(x-)$ and $\mathrm{f}(x+)$ denote limits of $\mathrm{f}(t)$ as t tends to x from the left and right, respectively.

Periodic function A function f is *periodic,* with *period* $P > 0$, if the identity $\mathrm{f}(x + P) = \mathrm{f}(x)$ holds for all x. Example: $\mathrm{f}(x) = |\sin x|$ is periodic with period π.

FOURIER SERIES has long provided one of the principal methods of analysis for mathematical physics, engineering, and signal processing. It has spurred generalizations and applications that continue to develop right up to the present. While the original theory of Fourier series applies to periodic functions occurring in wave motion, such as with light and sound, its generalizations often

relate to wider settings, such as the time-frequency analysis underlying the recent theories of wavelet analysis and local trigonometric analysis.

I. HISTORICAL BACKGROUND

There are antecedents to the notion of Fourier series in the work of Euler and D. Bernoulli on vibrating strings, but the theory of Fourier series truly began with the profound work of Fourier on heat conduction at the beginning of the 19th century. Fourier deals with the problem of describing the evolution of the temperature $T(x, t)$ of a thin wire of length π, stretched between $x = 0$ and $x = \pi$, with a constant zero temperature at the ends: $T(0, t) = 0$ and $T(\pi, t) = 0$. He proposed that the initial temperature $T(x, 0) = \mathrm{f}(x)$ could be expanded in a series of sine functions:

$$\mathrm{f}(x) = \sum_{n=1}^{\infty} b_n \sin nx \tag{1}$$

with

$$b_n = \frac{2}{\pi} \int_0^{\pi} \mathrm{f}(x) \sin nx \, dx. \tag{2}$$

A. Fourier Series

Although Fourier did not give a convincing proof of convergence of the infinite series in Eq. (1), he did offer the conjecture that convergence holds for an "arbitrary" function f. Subsequent work by Dirichlet, Riemann, Lebesgue, and others, throughout the next two hundred years, was needed to delineate precisely which functions were expandable in such trigonometric series. Part of this work entailed giving a precise definition of function (Dirichlet), and showing that the integrals in Eq. (2) are properly defined (Riemann and Lebesgue). Throughout this article we shall state results that are always true when Riemann integrals are used (except for Section IV where we need to use results from the theory of Lebesgue integrals).

In addition to positing Eqs. (1) and (2), Fourier argued that the temperature $T(x, t)$ is a solution to the following *heat equation with boundary conditions:*

$$\begin{aligned} \frac{\partial T}{\partial t} &= \frac{\partial^2 T}{\partial x^2}, && 0 < x < \pi, \ t > 0 \\ T(0, t) &= T(\pi, t) = 0, && t \geq 0 \\ T(x, 0) &= f(x), && 0 \leq x \leq \pi. \end{aligned}$$

Making use of Eq. (1), Fourier showed that the solution $T(x, t)$ satisfies

$$T(x, t) = \sum_{n=1}^{\infty} b_n e^{-n^2 t} \sin nx. \tag{3}$$

This was the first example of the use of Fourier series to solve *boundary value problems* in partial differential equations. To obtain Eq. (3), Fourier made use of D. Bernoulli's method of *separation of variables,* which is now a standard technique for solving boundary value problems.

A good, short introduction to the history of Fourier series can be found in *The Mathematical Experience*. Besides his many mathematical contributions, Fourier has left us with one of the truly great philosophical principles: "The deep study of nature is the most fruitful source of knowledge."

II. DEFINITION OF FOURIER SERIES

The Fourier sine series, defined in Eqs. (1) and (2), is a special case of a more general concept: the Fourier series for a *periodic function.* Periodic functions arise in the study of wave motion, when a basic waveform repeats itself periodically. Such periodic waveforms occur in musical tones, in the plane waves of electromagnetic vibrations, and in the vibration of strings. These are just a few examples. Periodic effects also arise in the motion of the planets, in AC electricity, and (to a degree) in animal heartbeats.

A function f is said to have period P if $\mathrm{f}(x + P) = \mathrm{f}(x)$ for all x. For notational simplicity, we shall restrict our discussion to functions of period 2π. There is no loss of generality in doing so, since we can always use a simple change of scale $x = (P/2\pi)t$ to convert a function of period P into one of period 2π.

If the function f has period 2π, then its *Fourier series* is

$$c_0 + \sum_{n=1}^{\infty} \{a_n \cos nx + b_n \sin nx\} \tag{4}$$

with *Fourier coefficients* c_0, a_n, and b_n defined by the integrals

$$c_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{f}(x) \, dx \tag{5}$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} \mathrm{f}(x) \cos nx \, dx, \tag{6}$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} \mathrm{f}(x) \sin nx \, dx. \tag{7}$$

[*Note:* The sine series defined by Eqs. (1) and (2) is a special instance of Fourier series. If f is initially defined over the interval $[0, \pi]$, then it can be extended to $[-\pi, \pi]$ (as an odd function) by letting $\mathrm{f}(-x) = -\mathrm{f}(x)$, and then extended periodically with period $P = 2\pi$. The Fourier series for this odd, periodic function reduces to the sine series in Eqs. (1) and (2), because $c_0 = 0$, each $a_n = 0$, and each b_n in Eq. (7) is equal to the b_n in Eq. (2).]

It is more common nowadays to express Fourier series in an algebraically simpler form involving complex exponentials. Following Euler, we use the fact that the complex exponential $e^{i\theta}$ satisfies $e^{i\theta} = \cos\theta + i\sin\theta$. Hence

$$\cos\theta = \frac{1}{2}(e^{i\theta} + e^{i\theta}),$$
$$\sin\theta = \frac{1}{2i}(e^{i\theta} - e^{-i\theta}).$$

From these equations, it follows by elementary algebra that Formulas (5)–(7) can be rewritten (by rewriting each term separately) as

$$c_0 + \sum_{n=1}^{\infty} \{c_n e^{inx} + c_{-n} e^{-inx}\} \tag{8}$$

with c_n defined for all integers n by

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{f}(x) e^{-inx}\, dx. \tag{9}$$

The series in Eq. (8) is usually written in the form

$$\sum_{n=-\infty}^{\infty} c_n e^{inx}. \tag{10}$$

We now consider a couple of examples. First, let f_1 be defined over $[-\pi, \pi]$ by

$$\mathrm{f}_1(x) = \begin{cases} 1 & \text{if } |x| < \pi/2 \\ 0 & \text{if } \pi/2 \le |x| \le \pi \end{cases}$$

and have period 2π. The graph of f_1 is shown in Fig. 1; it is called a *square wave* in electric circuit theory. The constant c_0 is

$$c_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{f}_1(x)\, dx$$
$$= \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} 1\, dx = \frac{1}{2}.$$

While, for $n \neq 0$,

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{f}_1(x) e^{-inx}\, dx$$
$$= \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} e^{-inx}\, dx$$
$$= \frac{1}{2\pi} \frac{e^{-in\pi/2} - e^{in\pi/2}}{-in}$$
$$= \frac{\sin(n\pi/2)}{n\pi}.$$

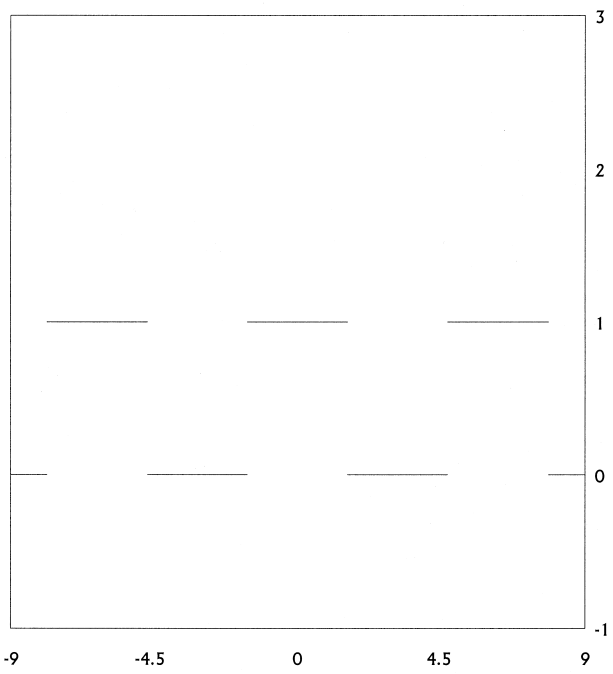


FIGURE 1 Square wave.

Thus, the Fourier series for this square wave is

$$\frac{1}{2} + \sum_{n=1}^{\infty} \frac{\sin(n\pi/2)}{n\pi} (e^{inx} + e^{-inx})$$
$$= \frac{1}{2} + \sum_{n=1}^{\infty} \frac{2\sin(n\pi/2)}{n\pi} \cos nx. \tag{11}$$

Second, let $\mathrm{f}_2(x) = x^2$ over $[-\pi, \pi]$ and have period 2π, see Fig. 2. We shall refer to this wave as a *parabolic wave*. This parabolic wave has $c_0 = \pi^2/3$ and c_n, for $n \neq 0$, is

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} x^2 e^{-inx}\, dx$$
$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} x^2 \cos nx\, dx - \frac{i}{2\pi} \int_{-\pi}^{\pi} x^2 \sin nx\, dx$$
$$= \frac{2(-1)^n}{n^2}$$

after an integration by parts. The Fourier series for this function is then

$$\frac{\pi^2}{3} + \sum_{n=1}^{\infty} \frac{2(-1)^n}{n^2} (e^{inx} + e^{-inx})$$
$$= \frac{\pi^2}{3} + \sum_{n=1}^{\infty} \frac{4(-1)^n}{n^2} \cos nx. \tag{12}$$

We will discuss the convergence of these Fourier series, to f_1 and f_2, respectively, in Section III.

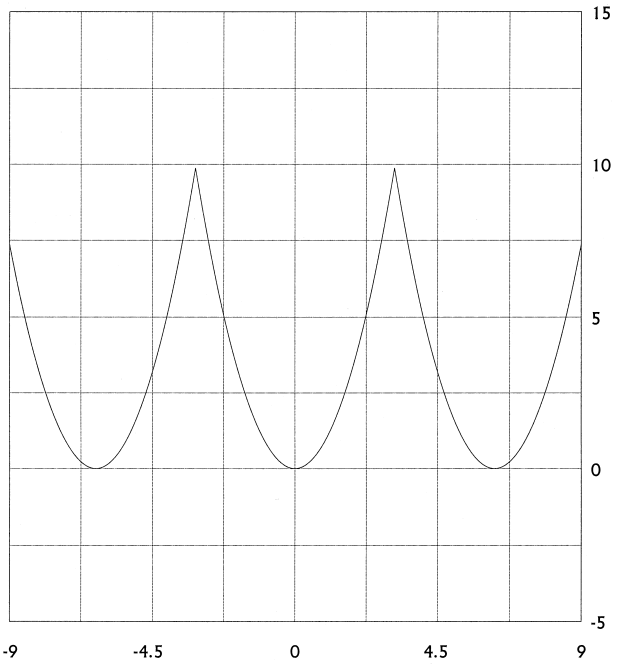


FIGURE 2 Parabolic wave.

Returning to the general Fourier series in Eq. (10), we shall now discuss some ways of interpreting this series. A complex exponential $e^{inx} = \cos nx + i \sin nx$ has a smallest period of $2\pi/n$. Consequently it is said to have a *frequency* of $n/2\pi$, because the form of its graph over the interval $[0, 2\pi/n]$ is repeated $n/2\pi$ times within each unit-length. Therefore, the integral in Eq. (9) that defines the Fourier coefficient c_n can be interpreted as a *correlation* between f and a complex exponential with a precisely located frequency of $n/2\pi$. Thus the whole collection of these integrals, for all integers n, specifies the *frequency content* of f over the set of frequencies $\{n/2\pi\}_{n=-\infty}^{\infty}$. If the series in Eq. (10) converges to f, i.e., if we can write

$$\mathrm{f}(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx}, \tag{13}$$

then f is being expressed as a superposition of elementary functions $c_n e^{inx}$ having frequency $n/2\pi$ and amplitude c_n. (The validity of Eq. (13) will be discussed in the next section.) Furthermore, the correlations in Eq. (9) are *independent* of each other in the sense that correlations between distinct exponentials are zero:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{inx} e^{-imx}\, dx = \begin{cases} 0 & \text{if } m \neq n \\ 1 & \text{if } m = n. \end{cases} \tag{14}$$

This equation is called the *orthogonality property* of complex exponentials.

The orthogonality property of complex exponentials can be used to give a derivation of Eq. (9). Multiplying Eq. (13) by e^{-imx} and integrating term-by-term from $-\pi$ to π, we obtain

$$\int_{-\pi}^{\pi} \mathrm{f}(x) e^{-imx}\, dx = \sum_{n=-\infty}^{\infty} c_n \int_{-\pi}^{\pi} e^{inx} e^{-imx}\, dx.$$

By the orthogonality property, this leads to

$$\int_{-\pi}^{\pi} \mathrm{f}(x) e^{-imx}\, dx = 2\pi c_m,$$

which justifies (in a formal, nonrigorous way) the definition of c_n in Eq. (9).

We close this section by discussing two important properties of Fourier coefficients, *Bessel's inequality* and the *Riemann-Lebesgue lemma.*

Theorem 1 (Bessel's Inequality): If $\int_{-\pi}^{\pi} |\mathrm{f}(x)|^2\, dx$ is finite, then

$$\sum_{n=-\infty}^{\infty} |c_n|^2 \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} |\mathrm{f}(x)|^2\, dx. \tag{15}$$

Bessel's inequality can be proved easily. In fact, we have

$$0 \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} \left| \mathrm{f}(x) - \sum_{n=-N}^{N} c_n e^{inx} \right|^2 dx$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(\mathrm{f}(x) - \sum_{m=-N}^{N} c_m e^{imx} \right) \left(\overline{\mathrm{f}(x)} - \sum_{n=-N}^{N} \overline{c_n} e^{-inx} \right) dx.$$

Multiplying out the last integrand above, and making use of Eqs. (9) and (14), we obtain

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \left| \mathrm{f}(x) - \sum_{n=-N}^{N} c_n e^{inx} \right|^2 dx$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} |\mathrm{f}(x)|^2\, dx - \sum_{n=-N}^{N} |c_n|^2. \tag{16}$$

Thus, for all N,

$$\sum_{n=-N}^{N} |c_n|^2 \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} |\mathrm{f}(x)|^2\, dx \tag{17}$$

and Bessel's inequality (15) follows by letting $N \to \infty$.

Bessel's inequality has a physical interpretation. If f has *finite energy,* in the sense that the right side of Eq. (15) is finite, then the sum of the moduli-squared of the Fourier coefficients is also finite. In Section IV, we shall see that the inequality in Eq. (15) is actually an equality, which says that the sum of the moduli-squared of the Fourier coefficients is precisely the same as the energy of f.

Because of Bessel's inequality, it follows that

$$\lim_{|n|\to\infty} c_n = 0 \tag{18}$$

holds whenever $\int_{-\pi}^{\pi} |f(x)|^2\,dx$ is finite. The Riemann-Lebesgue lemma says that Eq. (18) holds in the following more general case:

Theorem 2 (Riemann-Lebesgue Lemma): If $\int_{-\pi}^{\pi} |f(x)|\,dx$ is finite, then Eq. (18) holds.

One of the most important uses of the Riemann-Lebesgue lemma is in proofs of some basic pointwise convergence theorems, such as the ones described in the next section.

See Krantz and Walker (1998) for further discussions of the definition of Fourier series, Bessel's inequality, and the Riemann-Lebesgue lemma.

III. CONVERGENCE OF FOURIER SERIES

There are many ways to interpret the meaning of Eq. (13). Investigations into the types of functions allowed on the left side of Eq. (13), and the kinds of convergence considered for its right side, have fueled mathematical investigations by such luminaries as Dirichlet, Riemann, Weierstrass, Lipschitz, Lebesgue, Fejér, Gelfand, and Schwartz. In short, convergence questions for Fourier series have helped lay the foundations and much of the superstructure of mathematical analysis.

The three types of convergence that we shall describe here are *pointwise*, *uniform*, and *norm* convergence. We shall discuss the first two types in this section and take up the third type in the next section.

All convergence theorems are concerned with how the *partial sums*

$$S_N(x) := \sum_{n=-N}^{N} c_n\, e^{inx}$$

converge to $f(x)$. That is, *does* $\lim_{N\to\infty} S_N = f$ *hold in some sense*?

The question of pointwise convergence, for example, concerns whether $\lim_{N\to\infty} S_N(x_0) = f(x_0)$ for each fixed x-value x_0. If $\lim_{N\to\infty} S_N(x_0)$ does equal $f(x_0)$, then we say that the *Fourier series for* f *converges to* $f(x_0)$ *at* x_0.

We shall now state the simplest pointwise convergence theorem for which an elementary proof can be given. This theorem assumes that a function is Lipschitz at each point where convergence occurs. A function is said to be *Lipschitz at a point* x_0 if, for some positive constant A,

$$|f(x) - f(x_0)| \le A\,|x - x_0| \tag{19}$$

holds for all x near x_0 (i.e., $|x - x_0| < \delta$ for some $\delta > 0$). It is easy to see, for instance, that the square wave function f_1 is Lipschitz at all of its continuity points.

The inequality in Eq. (19) has a simple geometric interpretation. Since both sides are 0 when $x = x_0$, this inequality is equivalent to

$$\left|\frac{f(x) - f(x_0)}{x - x_0}\right| \le A \tag{20}$$

for all x near x_0 (and $x \ne x_0$). Inequality (20) simply says that the difference quotients of f (i.e., the slopes of its secants) near x_0 are bounded. With this interpretation, it is easy to see that the parabolic wave f_2 is Lipschitz at all points. More generally, if f has a derivative at x_0 (or even just left- and right-hand derivatives), then f is Lipschitz at x_0.

We can now state and prove a simple convergence theorem.

Theorem 3: Suppose f has period 2π, that $\int_{-\pi}^{\pi} |f(x)|\,dx$ is finite, and that f is Lipschitz at x_0. Then the Fourier series for f converges to $f(x_0)$ at x_0.

To prove this theorem, we assume that $f(x_0) = 0$. There is no loss of generality in doing so, since we can always subtract the constant $f(x_0)$ from $f(x)$. Define the function g by $g(x) = f(x)/(e^{ix} - e^{ix_0})$. This function g has period 2π. Furthermore, $\int_{-\pi}^{\pi} |g(x)|\,dx$ is finite, because the quotient $f(x)/(e^{ix} - e^{ix_0})$ is bounded in magnitude for x near x_0. In fact, for such x,

$$\left|\frac{f(x)}{e^{ix} - e^{x_0}}\right| = \left|\frac{f(x) - f(x_0)}{e^{ix} - e^{x_0}}\right| \le A\left|\frac{x - x_0}{e^{ix} - e^{x_0}}\right|$$

and $(x - x_0)/(e^{ix} - e^{x_0})$ is bounded in magnitude, because it tends to the reciprocal of the derivative of e^{ix} at x_0.

If we let d_n denote the nth Fourier coefficient for $g(x)$, then we have $c_n = d_{n-1} - d_n e^{ix_0}$ because $f(x) = g(x)(e^{ix} - e^{ix_0})$. The partial sum $S_N(x_0)$ then telescopes:

$$\begin{aligned} S_N(x_0) &= \sum_{n=-N}^{N} c_n e^{inx_0} \\ &= d_{-N-1} e^{-iNx_0} - d_N\, e^{i(N+1)x_0}. \end{aligned}$$

Since $d_n \to 0$ as $|n| \to \infty$, by the Riemann-Lebesgue lemma, we conclude that $S_N(x_0) \to 0$. This completes the proof.

It should be noted that for the square wave f_1 and the parabolic wave f_2, it is not necessary to use the general Riemann-Lebesgue lemma stated above. That is because for those functions it is easy to see that $\int_{-\pi}^{\pi} |g(x)|^2\,dx$ is

finite for the function g defined in the proof of Theorem 3. Consequently, $d_n \to 0$ as $|n| \to \infty$ follows from Bessel's inequality for g.

In any case, Theorem 3 implies that the Fourier series for the square wave f_1 converges to f_1 at all of its points of continuity. It also implies that the Fourier series for the parabolic wave f_2 converges to f_2 at all points. While this may settle matters (more or less) in a pure mathematical sense for these two waves, it is still important to examine specific partial sums in order to learn more about the nature of their convergence to these waves.

For example, in Fig. 3 we show a graph of the partial sum S_{100} superimposed on the square wave. Although Theorem 3 guarantees that $S_N \to \mathrm{f}_1$ as $N \to \infty$ at each continuity point, Fig. 3 indicates that this convergence is at a rather slow rate. The partial sum S_{100} differs significantly from f_1. Near the square wave's jump discontinuities, for example, there is a severe spiking behavior called *Gibbs' phenomenon* (see Fig. 4). This spiking behavior does *not* go away as $N \to \infty$, although the width of the spike does tend to zero. In fact, the peaks of the spikes overshoot the square wave's value of 1, tending to a limit of about 1.09. The partial sum also oscillates quite noticeably about the constant value of the square wave at points away from the discontinuities. This is known as *ringing.*

These defects do have practical implications. For instance, oscilloscopes—which generate wave forms as combinations of sinusoidal waves over a limited range of frequencies—cannot use S_{100}, or any partial sum S_N, to produce a square wave. We shall see, however, in Section V that a clever modification of a partial sum does produce an acceptable version of a square wave.

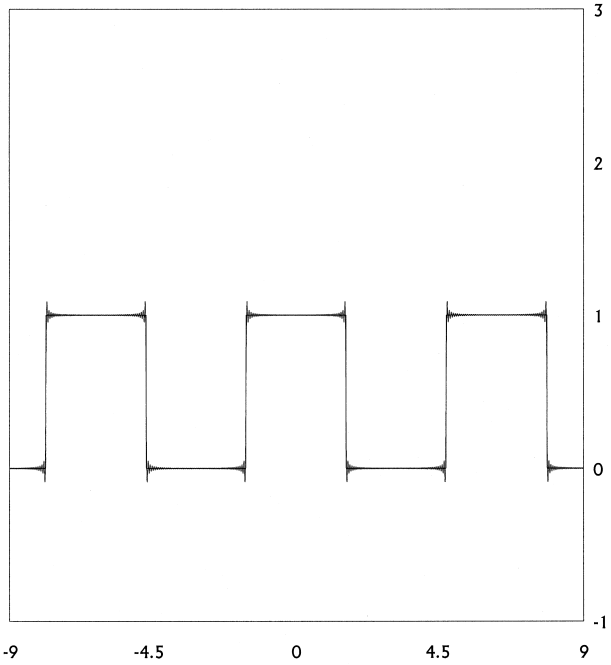


FIGURE 3 Fourier series partial sum S_{100} superimposed on square wave.

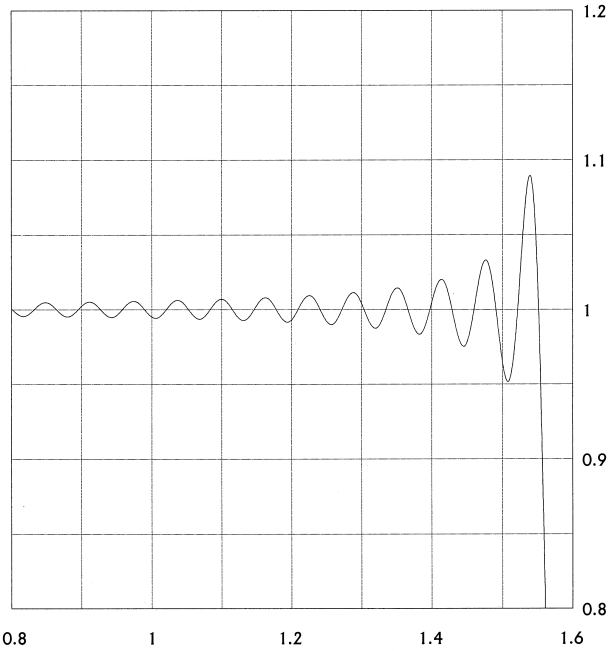


FIGURE 4 Gibbs' phenomenon and ringing for square wave.

The cause of ringing and Gibbs' phenomenon for the square wave is a rather slow convergence to zero of its Fourier coefficients (at a rate comparable to $|n|^{-1}$). In the next section, we shall interpret this in terms of energy and show that a partial sum like S_{100} does not capture a high enough percentage of the energy of the square wave f_1.

In contrast, the Fourier coefficients of the parabolic wave f_2 tend to zero more rapidly (at a rate comparable to n^{-2}). Because of this, the partial sum S_{100} for f_2 is a much better approximation to the parabolic wave (see Fig. 5). In fact, its partial sums S_N exhibit the phenomenon of *uniform convergence.*

We say that the Fourier series for a function f *converges uniformly* to f if

$$\lim_{N\to\infty} \left\{ \max_{x\in[-\pi,\pi]} |\mathrm{f}(x) - S_N(x)| \right\} = 0. \tag{21}$$

This equation says that, for large enough N, we can have the maximum distance between the graphs of f and S_N as small as we wish. Figure 5 is a good illustration of this for the parabolic wave.

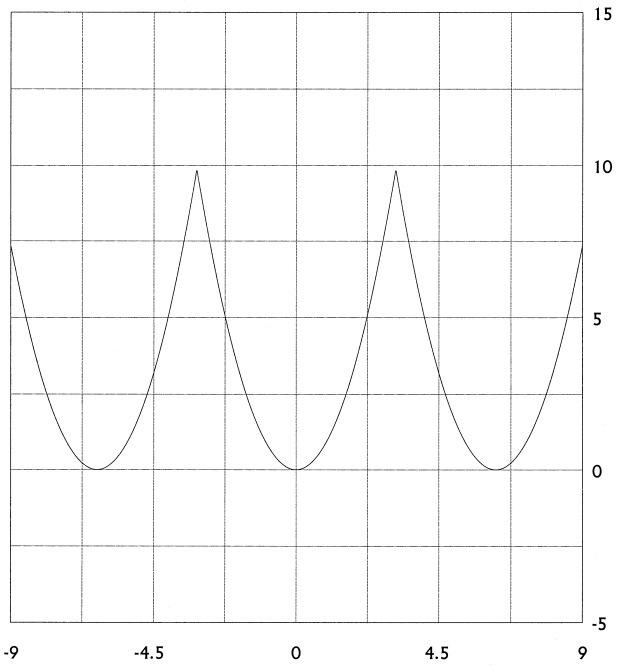


FIGURE 5 Fourier series partial sum S_{100} for parabolic wave.

We can verify Eq. (21) for the parabolic wave as follows. By Eq. (21) we have

$$\begin{aligned} |\mathrm{f}_2(x) - S_N(x)| &= \left| \sum_{n=N+1}^{\infty} \frac{4(-1)^n}{n^2} \cos nx \right| \\ &\le \sum_{n=N+1}^{\infty} \left| \frac{4(-1)^n}{n^2} \cos nx \right| \\ &\le \sum_{n=N+1}^{\infty} \frac{4}{n^2}. \end{aligned}$$

Consequently

$$\begin{aligned} \max_{x\in[-\pi,\pi]} |\mathrm{f}_2(x) - S_N(x)| &\le \sum_{n=N+1}^{\infty} \frac{4}{n^2} \\ &\to 0 \quad \text{as } N \to \infty \end{aligned}$$

and thus Eq. (21) holds for the parabolic wave f_2.

Uniform convergence for the parabolic wave is a special case of a more general theorem. We shall say that f is *uniformly Lipschitz* if Eq. (19) holds for all points using the same constant A. For instance, it is not hard to show that a continuously differentiable, periodic function is uniformly Lipschitz.

Theorem 4: Suppose that f has period 2π and is uniformly Lipschitz at all points, then the Fourier series for f converges uniformly to f.

A remarkably simple proof of this theorem is described in Jackson (1941). More general uniform convergence theorems are discussed in Walter (1994).

Theorem 4 applies to the parabolic wave f_2, but it does not apply to the square wave f_1. In fact, the Fourier series for f_1 cannot converge uniformly to f_1. That is because a famous theorem of Weierstrass says that a uniform limit of continuous functions (like the partial sums S_N) must be a continuous function (which f_1 is certainly not). The Gibbs' phenomenon for the square wave is a conspicuous failure of uniform convergence for its Fourier series.

Gibbs' phenomenon and ringing, as well as many other aspects of Fourier series, can be understood via an integral form for partial sums discovered by Dirichlet. This integral form is

$$S_N(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{f}(x-t) D_N(t)\, dt \tag{22}$$

with *kernel* D_N defined by

$$D_N(t) = \frac{\sin(N+1/2)t}{\sin(t/2)}. \tag{23}$$

This formula is proved in almost all books on Fourier series (see, for instance, Krantz (1999), Walker (1988), or Zygmund (1968)). The kernel D_N is called *Dirichlet's kernel.* In Fig. 6 we have graphed D_{20}.

The most important property of Dirichlet's kernel is that, for all N,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} D_N(t)\, dt = 1.$$

From Eq. (23) we can see that the value of 1 follows from cancellation of signed areas, and also that the contribution

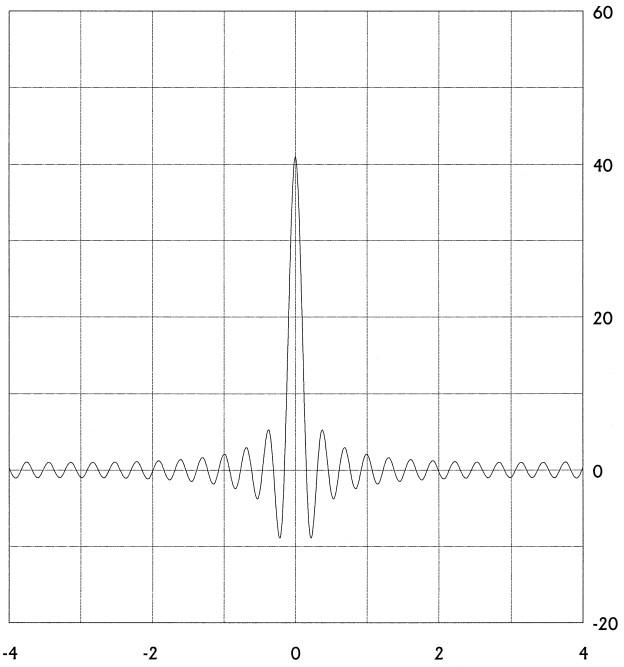


FIGURE 6 Dirichlet's kernel D_{20}.

of the main lobe centered at 0 (see Fig. 6) is significantly greater than 1 (about 1.09 in value).

From the facts just cited, we can explain the origin of ringing and Gibbs' phenomenon for the square wave. For the square wave function f_1, Eq. (22) becomes

$$S_N(x) = \frac{1}{2\pi}\int_{x-\pi/2}^{x+\pi/2} D_N(t)\,dt. \tag{24}$$

As x ranges from $-\pi$ to π, this formula shows that $S_N(x)$ is proportional to the signed area of D_N over an interval of length π centered at x. By examining Fig. 6, which is a typical graph for D_N, it is then easy to see why there is ringing in the partial sums S_N for the square wave. Gibbs' phenomenon is a bit more subtle, but also results from Eq. (24). When x nears a jump discontinuity, the central lobe of D_N is the dominant contributor to the integral in Eq. (24), resulting in a spike which overshoots the value of 1 for f_1 by about 9%.

Our final pointwise convergence theorem was, in essence, the first to be proved. It was established by Dirichlet using the integral form for partial sums in Eq. (22). We shall state this theorem in a stronger form first proved by Jordan.

Theorem 5: If f has period 2π and has bounded variation on $[0, 2\pi]$, then the Fourier series for f converges at all points. In fact, for all x-values,

$$\lim_{N\to\infty} S_N(x) = \tfrac{1}{2}[f(x+) + f(x-)].$$

This theorem is too difficult to prove in the limited space we have here (see Zygmund, 1968). A simple consequence of Theorem 5 is that the Fourier series for the square wave f_1 converges at its discontinuity points to $1/2$ (although this can also be shown directly by substitution of $x = \pm\pi/2$ into the series in (Eq. (11)).

We close by mentioning that the conditions for convergence, such as Lipschitz or bounded variation, cited in the theorems above cannot be dispensed with entirely. For instance, Kolmogorov gave an example of a period 2π function (for which $\int_{-\pi}^{\pi} |f(x)|\,dx$ is finite) that has a Fourier series which fails to converge at *every* point.

More discussion of pointwise convergence can be found in Walker (1998), Walter (1994), or Zygmund (1968).

IV. CONVERGENCE IN NORM

Perhaps the most satisfactory notion of convergence for Fourier series is convergence in L^2-norm (also called L^2-convergence), which we shall define in this section. One of the great triumphs of the Lebesgue theory of integration is that it yields necessary and sufficient conditions for L^2-convergence. There is also an interpretation of L^2-norm in terms of a generalized Euclidean distance and this gives a satisfying geometric flavor to L^2-convergence of Fourier series. By interpreting the square of L^2-norm as a type of energy, there is an equally satisfying physical interpretation of L^2-convergence. The theory of L^2-convergence has led to fruitful generalizations such as Hilbert space theory and norm convergence in a wide variety of function spaces.

To introduce the idea of L^2-convergence, we first examine a special case. By Theorem 4, the partial sums of a uniformly Lipschitz function f converge uniformly to f. Since that means that the maximum distance between the graphs of S_N and f tends to 0 as $N \to \infty$, it follows that

$$\lim_{N\to\infty} \frac{1}{2\pi}\int_{-\pi}^{\pi} |f(x) - S_N(x)|^2\,dx = 0. \tag{25}$$

This result motivates the definition of L^2-convergence.

If g is a function for which $|g|^2$ has a finite Lebesgue integral over $[-\pi, \pi]$, then we say that g is an L^2-function, and we define its L^2-norm $\|g\|_2$ by

$$\|g\|_2 = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} |g(x)|^2\,dx}\,.$$

We can then rephrase Eq. (25) as saying that $\|f - S_N\|_2 \to 0$ as $N \to \infty$. In other words, *the Fourier series for* f *converges to* f *in* L^2*-norm.* The following theorem generalizes this result to all L^2-functions (see Rudin (1986) for a proof).

Theorem 6: If f is an L^2-function, then its Fourier series converges to f in L^2-norm.

Theorem 6 says that Eq. (25) holds for every L^2-function f. Combining this with Eq. (16), we obtain *Parseval's equality:*

$$\sum_{n=-\infty}^{\infty} |c_n|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |f(x)|^2\,dx. \tag{26}$$

Parseval's equation has a useful interpretation in terms of energy. It says that the energy of the set of Fourier coefficients, defined to be equal to the left side of Eq. (26), is equal to the energy of the function f, defined by the right side of Eq. (26).

The L^2-norm can be interpreted as a generalized Euclidean distance. To see this take square roots of both sides of Eq. (26): $\sqrt{\sum |c_n|^2} = \|f\|_2$. The left side of this equation is interpreted as a Euclidean distance in an (infinite-dimensional) coordinate space, hence the L^2-norm $\|f\|_2$ is equivalent to such a distance.

As examples of these ideas, let's return to the square wave and parabolic wave. For the square wave f_1, we find that

$$\|f_1 - S_{100}\|_2^2 = \sum_{|n|>100} \frac{\sin^2(n\pi/2)}{(n\pi)^2}$$
$$= 1.0 \times 10^{-3}.$$

Likewise, for the parabolic wave f_2, we have $\|f_2 - S_{100}\|_2^2 = 2.6 \times 10^{-6}$. These facts show that the energy of the parabolic wave is almost entirely contained in the partial sum S_{100}; their energy difference is almost three orders of magnitude smaller than in the square wave case. In terms of generalized Euclidean distance, we have $\|f_2 - S_{100}\|_2 = 1.6 \times 10^{-3}$ and $\|f_1 - S_{100}\|_2 = 3.2 \times 10^{-2}$, showing that the partial sum is an order of magnitude closer for the parabolic wave.

Theorem 6 has a converse, known as the *Riesz-Fischer theorem.*

Theorem 7 (Riesz-Fischer): If $\sum |c_n|^2$ converges, then there exists an L^2-function f having $\{c_n\}$ as its Fourier coefficients.

This theorem is proved in Rudin (1986). Theorem and the Riesz-Fischer theorem combine to give necessary and sufficient conditions for L^2-convergence of Fourier series, conditions which are remarkably easy to apply. This has made L^2-convergence into the most commonly used notion of convergence for Fourier series.

These ideas for L^2-norms partially generalize to the case of L^p-norms. Let p be real number satisfying $p \geq 1$. If g is a function for which $|g|^p$ has a finite Lebesgue integral over $[-\pi, \pi]$, then we say that g is an L^p-function, and we define its L^p-norm $\|g\|_p$ by

$$\|g\|_p = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} |g(x)|^p \, dx\right]^{1/p}.$$

If $\|f - S_N\|_p \to 0$, then we say that the Fourier series for f converges to f in L^p-norm. The following theorem generalizes Theorem 6 (see Krantz (1999) for a proof).

Theorem 8: If f is an L^p-function for $p > 1$, then its Fourier series converges to f in L^p-norm.

Notice that the case of $p = 1$ is not included in Theorem 8. The example of Kolmogorov cited at the end of Section III shows that there exist L^1-functions whose Fourier series do not converge in L^1-norm. For $p \neq 2$, there are no simple analogs of either Parseval's equality or the Riesz-Fischer theorem (which say that we can characterize L^2-functions by the magnitude of their Fourier coefficients). Some partial analogs of these latter results for L^p-functions, when $p \neq 2$, are discussed in Zygmund (1968) (in the context of *Littlewood-Paley* theory).

We close this section by returning full circle to the notion of pointwise convergence. The following theorem was proved by Carleson for L^2-functions and by Hunt for L^p-functions ($p \neq 2$).

Theorem 9: If f is an L^p-function for $p > 1$, then its Fourier series converges to it at almost all points.

By *almost all points,* we mean that the set of points where divergence occurs has Lebesgue measure zero. References for the proof of Theorem 9 can be found in Krantz (1999) and Zygmund (1968). Its proof is undoubtedly the most difficult one in the theory of Fourier series.

V. SUMMABILITY OF FOURIER SERIES

In the previous sections, we noted some problems with convergence of Fourier series partial sums. Some of these problems include Kolmogorov's example of a Fourier series for an L^1-function that diverges everywhere, and Gibbs' phenomenon and ringing in the Fourier series partial sums for discontinuous functions. Another problem is Du Bois Reymond's example of a continuous function whose Fourier series diverges on a countably infinite set of points, see Walker (1968). It turns out that all of these difficulties simply disappear when new summation methods, based on appropriate modifications of the partial sums, are used.

The simplest modification of partial sums, and one of the first historically to be used, is to take *arithmetic means.* Define the Nth arithmetic mean σ_N by $\sigma_N = (S_0 + S_1 + \cdots + S_{N-1})/N$. From which it follows that

$$\sigma_N(x) = \sum_{n=-N}^{N} \left(1 - \frac{|n|}{N}\right) c_n \, e^{inx}. \tag{27}$$

The factors $(1 - |n|/N)$ are called *convergence factors.* They modify the Fourier coefficients c_n so that the amplitude of the higher frequency terms (for $|n|$ near N) are damped down toward zero. This produces a great improvement in convergence properties as shown by the following theorem.

Theorem 10: Let f be a periodic function. If f is an L^p-function for $p \geq 1$, then $\sigma_N \to f$ in L^p-norm as $N \to \infty$. If f is a continuous function, then $\sigma_N \to f$ uniformly as $N \to \infty$.

Notice that L^1-convergence is included in Theorem 10. Even for Kolmogorov's function, it is the case that $\|f - \sigma_N\|_1 \to 0$ as $N \to \infty$. It also should be noted that no assumption, other than continuity of the periodic function, is needed in order to ensure uniform convergence of its arithmetic means.

For a proof of Theorem 10, see Krantz (1999). The key to the proof is Fejér's integral form for σ_N:

$$\sigma_N(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{f}(x-t)F_N(t)\,dt \tag{28}$$

where *Fejér's kernel* F_N is defined by

$$F_N(t) = \frac{1}{N}\left(\frac{\sin Nt/2}{\sin t/2}\right)^2. \tag{29}$$

In Fig. 7 we show the graph of F_{20}. Compare this graph with the one of Dirichlet's kernel D_{20} in Fig. 6. Unlike Dirichlet's kernel, Fejér's kernel is positive $[F_N(t) \geq 0]$, and is close to 0 away from the origin. These two facts are the main reasons that Theorem 10 holds. The fact that Fejér's kernel satisfies

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} F_N(t)\,dt = 1$$

is also used in the proof.

An attractive feature of arithmetic means is that Gibbs' phenomenon and ringing do not occur. For example, in Fig. 8 we show σ_{100} for the square wave and it is plain that these two defects are absent. For the square wave function f_1, Eq. (28) reduces to

$$\sigma_N(x) = \frac{1}{2\pi}\int_{x-\pi/2}^{x+\pi/2} F_N(t)\,dt.$$

As x ranges from $-\pi$ to π, this formula shows that $\sigma_N(x)$ is proportional to the area of the positive function F_N over

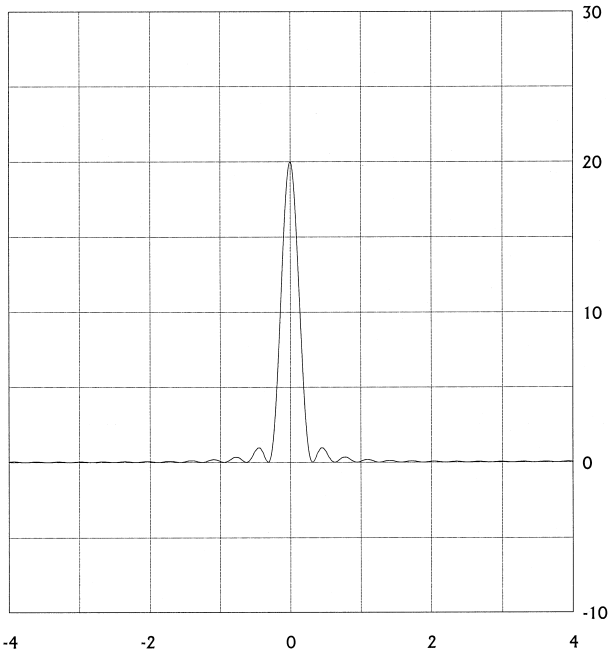


FIGURE 7 Fejér's kernel F_{20}.

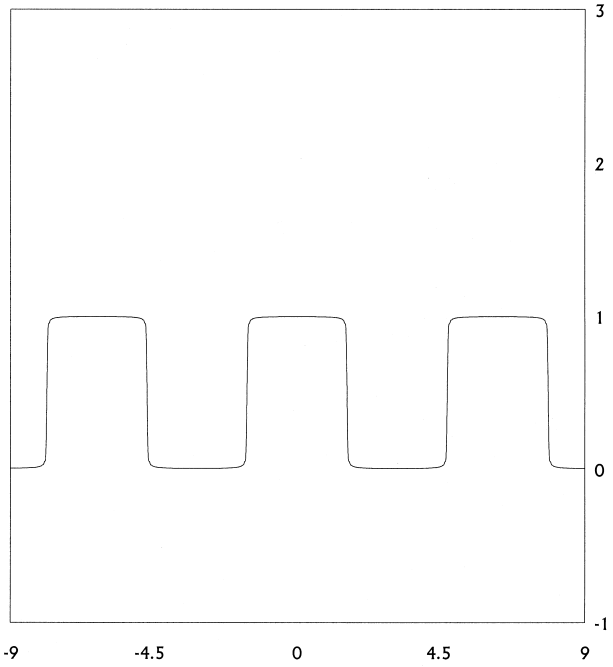


FIGURE 8 Arithmetic mean σ_{100} for square wave.

an interval of length π centered at x. By examining Fig. 7, which is a typical graph for F_N, it is easy to see why ringing and Gibbs' phenomenon do not occur for the arithmetic means of the square wave.

The method of arithmetic means is just one example from a wide range of summation methods for Fourier series. These summation methods are one of the major elements in the area of *finite impulse response filtering* in the fields of electrical engineering and signal processing.

A *summation kernel* K_N is defined by

$$K_N(x) = \sum_{n=-N}^{N} m_n\, e^{inx}. \tag{30}$$

The real numbers $\{m_n\}$ are the *convergence factors* for the kernel. We have already seen two examples: Dirichlet's kernel (where $m_n = 1$) and Fejér's kernel (where $m_n = 1 - |n|/N$).

When K_N is a summation kernel, then we define the modified partial sum of f to be $\sum_{n=-N}^{N} m_n\, c_n\, e^{inx}$. It then follows from Eqs. (14) and (30) that

$$\sum_{n=-N}^{N} m_n\, c_n\, e^{inx} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{f}(x-t)K_N(t)\,dt. \tag{31}$$

The function defined by both sides of Eq. (31) is denoted by $K_N * \mathrm{f}$. It is usually more convenient to use the left side of Eq. (31) to compute $K_N * \mathrm{f}$, while for theoretical purposes (such as proving Theorem 11 below), it is more convenient to use the right side of Eq. (31).

We say that a summation kernel K_N is *regular* if it satisfies the following three conditions.

1. For each N,
$$\frac{1}{2\pi}\int_{-\pi}^{\pi} K_N(x)\,dx = 1.$$
2. There is a positive constant C such that
$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |K_N(x)|\,dx \le C.$$
3. For each $0 < \delta < \pi$,
$$\lim_{N\to\infty}\left\{\max_{\delta\le|x|\le\pi} |K_N(x)|\right\} = 0.$$

There are many examples of regular summation kernels. Fejér's kernel, which has $m_n = 1 - |n|/N$, is regular. Another regular summation kernel is Hann's kernel, which has $m_n = 0.5 + 0.5\cos(n\pi/N)$. A third regular summation kernel is de le Vallée Poussin's kernel, for which $m_n = 1$ when $|n| \le N/2$, and $m_n = 2(1 - |m/N|)$ when $N/2 < |m| \le N$. The proofs that these summation kernels are regular are given in Walker (1996). It should be noted that Dirichlet's kernel is *not* regular, because properties 2 and 3 do not hold.

As with Fejér's kernel, all regular summation kernels significantly improve the convergence of Fourier series. In fact, the following theorem generalizes Theorem 10.

Theorem 11: Let f be a periodic function, and let K_N be a regular summation kernel. If f is an L^p-function for $p \ge 1$, then $K_N * \mathrm{f} \to \mathrm{f}$ in L^p-norm as $N \to \infty$. If f is a continuous function, then $K_N * \mathrm{f} \to \mathrm{f}$ uniformly as $N \to \infty$.

For an elegant proof of this theorem, see Krantz (1999).

From Theorem 11 we might be tempted to conclude that the convergence properties of regular summation kernels are all the same. They do differ, however, in the *rates* at which they converge. For example, in Fig. 9 we show $K_{100} * \mathrm{f}_1$ where the kernel is Hann's kernel and f_1 is the square wave. Notice that this graph is a much better approximation of a square wave than the arithmetic mean graph in Fig. 8. An oscilloscope, for example, can easily generate the graph in Fig. 9, thereby producing an acceptable version of a square wave.

Summation of Fourier series is discussed further in Krantz (1999), Walker (1996), Walter (1994), and Zygmund (1968).

VI. GENERALIZED FOURIER SERIES

The classical theory of Fourier series has undergone extensive generalizations during the last two hundred years.

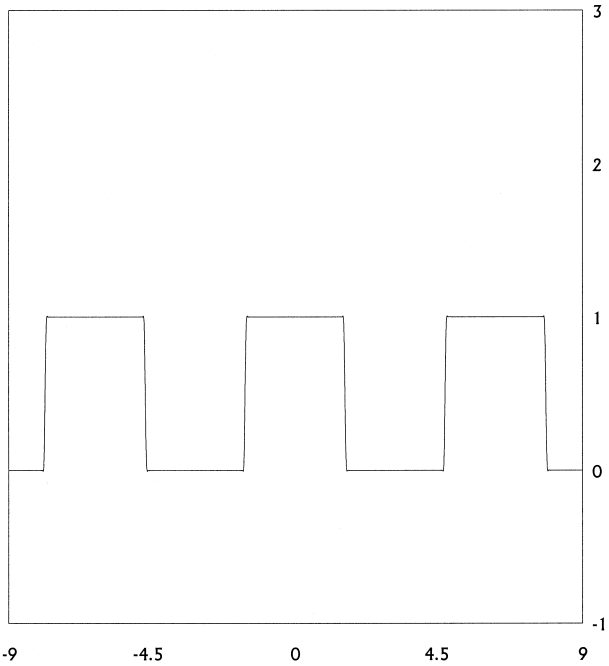


FIGURE 9 Approximate square wave using Hann's kernel.

For example, Fourier series can be viewed as one aspect of a general theory of *orthogonal series expansions.* In this section, we shall discuss a few of the more celebrated orthogonal series, such as Legendre series, Haar series, and wavelet series.

We begin with Legendre series. The first two *Legendre polynomials* are $P_0(x) = 1$, and $P_1(x) = x$. For $n = 2, 3, 4, \ldots$, the nth Legendre polynomial P_n is defined by the recursion relation

$$n P_n(x) = (2n-1)x P_{n-1}(x) + (n-1)P_{n-2}(x).$$

These polynomials satisfy the following *orthogonality relation*

$$\int_{-1}^{1} P_n(x)\,P_m(x)\,dx = \begin{cases} 0 & \text{if } m \ne n \\ (2n+1)/2 & \text{if } m = n. \end{cases} \tag{32}$$

This equation is quite similar to Eq. (14). Because of Eq. (32)—recall how we used Eq. (14) to derive Eq. (9)—the *Legendre series* for a function f over the interval $[-1, 1]$ is defined to be

$$\sum_{n=0}^{\infty} c_n P_n(x) \tag{33}$$

with

$$c_n = \frac{2}{2n+1}\int_{-1}^{1} \mathrm{f}(x)\,P_n(x)\,dx. \tag{34}$$

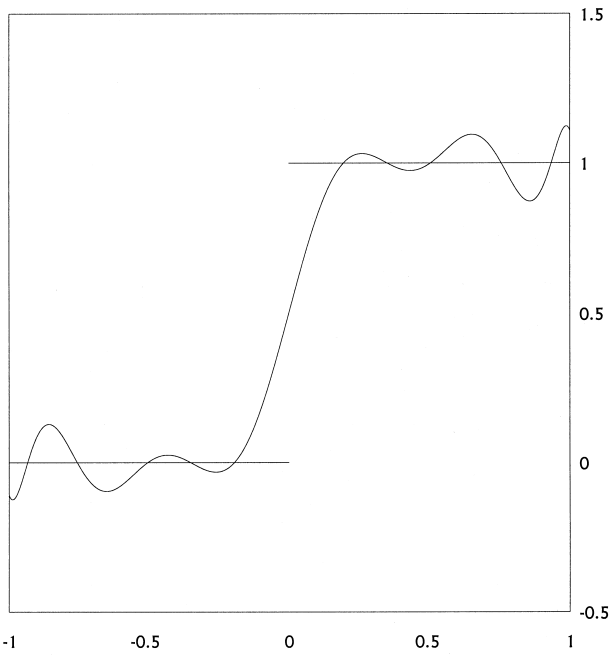


FIGURE 10 Step function and its Legendre series partial sum S_{11}.

The partial sum S_N of the series in Eq. (33) is defined to be

$$S_N(x) = \sum_{n=0}^{N} c_n P_n(x).$$

As an example, let $f(x) = 1$ for $0 \le x \le 1$ and $f(x) = 0$ for $-1 \le x < 0$. The Legendre series for this step function is [see Walker (1988)]:

$$\frac{1}{2} + \sum_{k=0}^{\infty} \frac{(-1)^k (4k+3)(2k)!}{4^{k+1}(k+1)!k!} P_{2k+1}(x).$$

In Fig. 10 we show the partial sum S_{11} for this series. The graph of S_{11} is reminiscent of a Fourier series partial sum for a step function. In fact, the following theorem is true.

Theorem 12: If $\int_{-1}^{1} |f(x)|^2 \, dx$ is finite, then the partial sums S_N for the Legendre series for f satisfy

$$\lim_{N \to \infty} \int_{-1}^{1} |f(x) - S_N(x)|^2 \, dx = 0.$$

Moreover, if f is Lipschitz at a point x_0, then $S_N(x_0) \to f(x_0)$ as $N \to \infty$.

This theorem is proved in Walter (1994) and Jackson (1941). Further details and other examples of *orthogonal polynomial series* can be found in either Davis (1975), Jackson (1941), or Walter (1994). There are many important orthogonal series—such as Hermite, Laguerre, and Tchebysheff—which we cannot examine here because of space limitations.

We now turn to another type of orthogonal series, the Haar series. The defects, such as Gibbs' phenomenon and ringing, that occur with Fourier series expansions can be traced to the unlocalized nature of the functions used for expansions. The complex exponentials used in classical Fourier series, and the polynomials used in Legendre series, are all non-zero (except possibly for a finite number of points) over their domains. In contrast, Haar series make use of localized functions, which are non-zero only over tiny regions within their domains.

In order to define Haar series, we first define the *fundamental Haar wavelet* $H(x)$ by

$$H(x) = \begin{cases} 1 & \text{if } 0 \le x < 1/2 \\ -1 & \text{if } 1/2 \le x \le 1. \end{cases}$$

The *Haar wavelets* $\{H_{j,k}(x)\}$ are then defined by

$$H_{j,k}(x) = 2^{j/2} H(2^j x - k)$$

for $j = 0, 1, 2, \ldots$; $k = 0, 1, \ldots, 2^j - 1$. Notice that $H_{j,k}(x)$ is non-zero only on the interval $[k2^{-j}, (k+1)2^{-j}]$, which for large j is a tiny subinterval of $[0, 1]$. As k ranges between 0 and $2^j - 1$, these subintervals partition the interval $[0, 1]$, and the partition becomes finer (shorter subintervals) with increasing j.

The *Haar series* for a function f is defined by

$$b + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j - 1} c_{j,k} H_{j,k}(x) \tag{35}$$

with $b = \int_0^1 f(x) \, dx$ and

$$c_{j,k} = \int_0^1 f(x) H_{j,k}(x) \, dx.$$

The definitions of b and $c_{j,k}$ are justified by orthogonality relations between the Haar functions (similar to the orthogonality relations that we used above to justify Fourier series and Legendre series).

A partial sum S_N for the Haar series in Eq. () is defined by

$$S_N(x) = b + \sum_{\{j,k \,|\, 2^j + k \le N\}} c_{j,k} H_{j,k}(x).$$

For example, let f be the function on $[0, 1]$ defined as follows

$$f(x) = \begin{cases} x - 1/2 & \text{if } 1/4 < x < 3/4 \\ 0 & \text{if } x \le 1/4 \text{ or } 3/4 \le x. \end{cases}$$

In Fig. 11 we show the Haar series partial sum S_{256} for this function. Notice that there is no Gibbs' phenomenon with this partial sum. This contrasts sharply with the Fourier

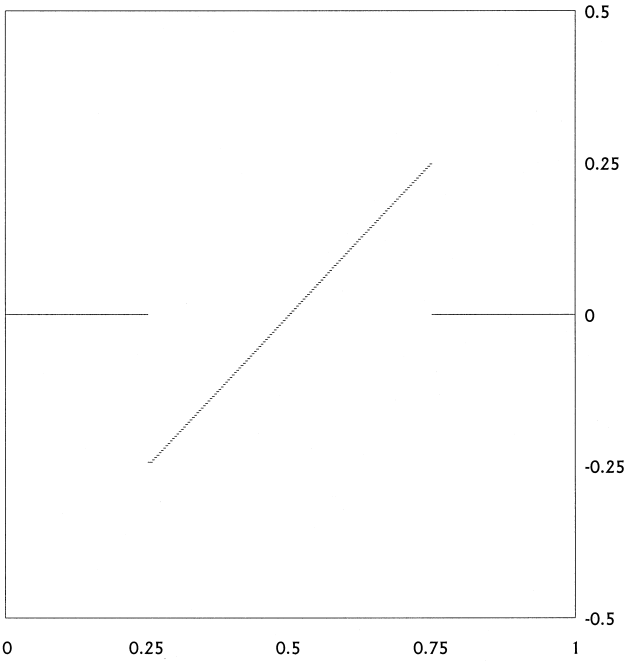


FIGURE 11 Haar series partial sum S_{256}, which has 257 terms.

series partial sum, also using 257 terms, which we show in Fig. 12.

The Haar series partial sums satisfy the following theorem [proved in Daubechies (1992) and in Meyer (1992)].

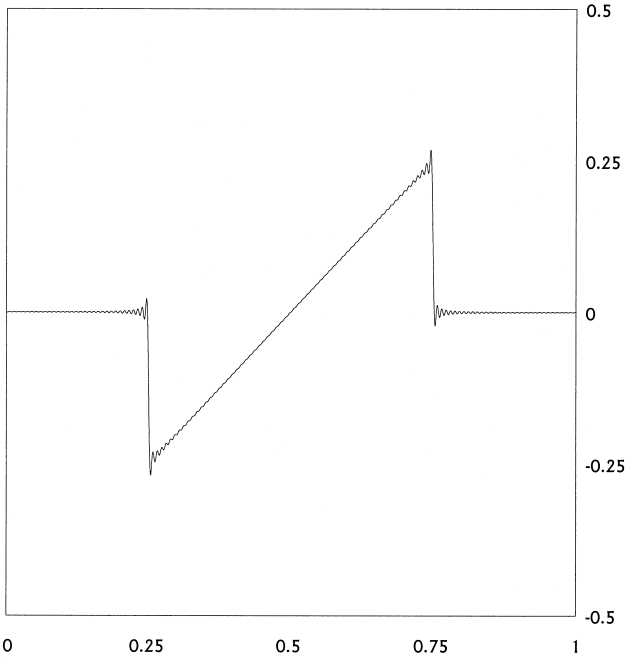


FIGURE 12 Fourier series partial sum S_{128}, which has 257 terms.

Theorem 13: Suppose that $\int_0^1 |f(x)|^p\,dx$ is finite, for $p \geq 1$. Then the Haar series partial sums for f satisfy

$$\lim_{N\to\infty}\left[\int_0^1 |f(x) - S_N(x)|^p\,dx\right]^{1/p} = 0.$$

If f is continuous on [0, 1], then S_N converges uniformly to f on [0, 1].

This theorem is reminiscent of Theorems 10 and 11 for the modified Fourier series partial sums obtained by arithmetic means or by a regular summation kernel. The difference here, however, is that for the Haar series no modifications of the partial sums are needed.

One glaring defect of Haar series is that the partial sums are discontinuous functions. This defect is remedied by the wavelet series discovered by Meyer, Daubechies, and others. The fundamental Haar wavelet is replaced by some new fundamental wavelet Ψ and the set of wavelets $\{\Psi_{j,k}\}$ is then defined by $\Psi_{j,k}(x) = 2^{-j/2}\Psi[2^j x - k]$. (The bracket symbolism $\Psi[2^j x - k]$ means that the value, $2^j x - k \mod 1$, is evaluated by Ψ. This technicality is needed in order to ensure periodicity of $\Psi_{j,k}$.) For example, in Fig. 13, we show graphs of $\Psi_{4,1}$ and $\Psi_{6,46}$ for one of the Daubechies wavelets (a Coif18 wavelet), which is *continuously differentiable.* For a complete discussion of the definition of these wavelet functions, see Daubechies (1992) or Mallat (1998).

The *wavelet series,* generated by the fundamental wavelet Ψ, is defined by

$$b + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} c_{j,k}\Psi_{j,k}(x) \tag{36}$$

with $b = \int_0^1 f(x)\,dx$ and

$$c_{j,k} = \int_0^1 f(x)\Psi_{j,k}(x)\,dx. \tag{37}$$

This wavelet series has partial sums S_N defined by

$$S_N(x) = b + \sum_{\{j,k\,|\,2^j+k\leq N\}} c_{j,k}\Psi_{j,k}(x).$$

Notice that when Ψ is continuously differentiable, then so is each partial sum S_N. These wavelet series partial sums satisfy the following theorem, which generalizes Theorem 13 for Haar series, for a proof, see Daubechies (1992) or Meyer (1992).

Theorem 14: Suppose that $\int_0^1 |f(x)|^p\,dx$ is finite, for $p \geq 1$. Then the Daubechies wavelet series partial sums for f satisfy

$$\lim_{N\to\infty}\left[\int_0^1 |f(x) - S_N(x)|^p\,dx\right]^{1/p} = 0.$$

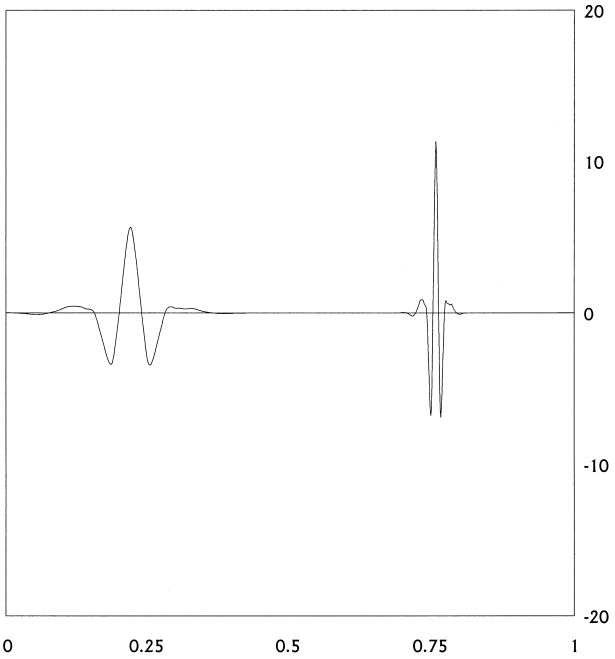


FIGURE 13 Two Daubechies wavelets.

If f is continuous on [0, 1], then S_N converges uniformly to f on [0, 1].

Theorem 14 does not reveal the full power of wavelet series. In almost all cases, it is possible to rearrange the terms in the wavelet series *in any manner whatsoever* and convergence will still hold. One reason for doing a rearrangement is in order to add the terms in the series with coefficients of largest magnitude (thus largest energy) *first* so as to speed up convergence to the function. Here is a convergence theorem for such permuted series.

Theorem 15: Suppose that $\int_0^1 |\mathrm{f}(x)|^p\,dx$ is finite, for $p > 1$. If the terms of a Daubechies wavelet series are permuted (in any manner whatsoever), then the partial sums S_N of the permuted series satisfy

$$\lim_{N\to\infty}\left[\int_0^1 |\mathrm{f}(x) - S_N(x)|^p\,dx\right]^{1/p} = 0.$$

If f is uniformly Lipschitz, then the partial sums S_N of the permuted series converge uniformly to f.

This theorem is proved in Daubechies (1992) and Meyer (1992). This type of convergence of wavelet series is called *unconditional convergence.* It is known [see Mallat (1998)] that unconditional convergence of wavelet series ensures an optimality of compression of signals. For details about compression of signals and other applications of wavelet series, see Walker (1999) for a simple introduction and Mallat (1998) for a thorough treatment.

VII. DISCRETE FOURIER SERIES

The digital computer has revolutionized the practice of science in the latter half of the twentieth century. The methods of computerized Fourier series, based upon the *fast Fourier transform* algorithms for digital approximation of Fourier series, have completely transformed the application of Fourier series to scientific problems. In this section, we shall briefly outline the main facts in the theory of discrete Fourier series.

The Fourier series coefficients $\{c_n\}$ can be discretely approximated via Riemann sums for the integrals in Eq. (9). For a (large) positive integer M, let $x_k = -\pi + 2\pi k/M$ for $k = 0, 1, 2, \ldots, M-1$ and let $\Delta x = 2\pi/M$. Then the nth Fourier coefficient c_n for a function f is approximated as follows:

$$c_n \approx \frac{1}{2\pi}\sum_{k=0}^{M-1} \mathrm{f}(x_k)\,e^{-i2\pi n x_k}\Delta x$$

$$= \frac{e^{-in\pi}}{M}\sum_{k=0}^{M-1} \mathrm{f}(x_k)e^{-i2\pi kn/M}.$$

The last sum above is called the *Discrete Fourier Transform* (DFT) of the finite sequence of numbers $\{\mathrm{f}(x_k)\}$. That is, we define the DFT of a sequence $\{g_k\}_{k=0}^{M-1}$ of numbers by

$$G_n = \sum_{k=0}^{M-1} g_k\,e^{-i2\pi kn/M}. \tag{38}$$

The DFT is the set of numbers $\{G_n\}$, and we see from the discussion above that the Fourier coefficients of a function f can be approximated by a DFT (multiplied by the factors $e^{-in\pi}/M$). For example, in Fig. 14 we show a graph of approximations of the Fourier coefficients $\{c_n\}_{n=-50}^{50}$ of the square wave f_1 obtained via a DFT (using $M = 1024$). For all values, these approximate Fourier coefficients differ from the exact coefficients by no more than 10^{-3}. By taking M even larger, the error can be reduced still further.

The two principal properties of DFTs are that they can be inverted and they preserve energy (up to a scale factor). The inversion formula for the DFT is

$$g_k = \sum_{n=0}^{M-1} G_n\,e^{i2\pi kn/M}. \tag{39}$$

And the conservation of energy property is

$$\sum_{k=0}^{M-1} |g_k|^2 = \frac{1}{N}\sum_{n=0}^{M-1} |G_n|^2. \tag{40}$$

Interpreting a sum of squares as energy, Eq. (40) says that, up to multiplication by the factor $1/N$, the energy of the discrete signal $\{g_k\}$ and its DFT $\{G_n\}$ are the same. These

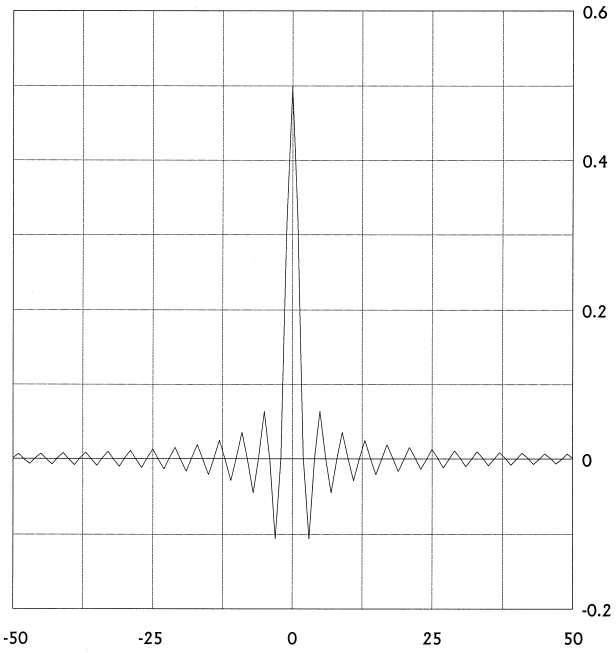


FIGURE 14 Fourier coefficients for square wave, $n = -50$ to 50. Successive values are connected with line segments.

facts are proved in Briggs and Henson (1995) and Walker (1996).

An application of inversion of DFTs is to the calculation of Fourier series partial sums. If we substitute $x_k = -\pi + 2\pi k/M$ into the Fourier series partial sum $S_N(x)$ we obtain (assuming that $N < M/2$ and after making a change of indices $m = n + N$):

$$\begin{aligned} S_N(x_k) &= \sum_{n=-N}^{N} c_n\, e^{in(-\pi+2\pi k/M)} \\ &= \sum_{n=-N}^{N} c_n\,(-1)^n\, e^{i2\pi nk/M} \\ &= \sum_{m=0}^{2N} c_{m-N}\,(-1)^{m-N}\, e^{-i2\pi kN/M}\, e^{i2\pi km/M}. \end{aligned}$$

Thus, if we let $g_m = c_{m-N}$ for $m = 0, 1, \ldots, 2N$ and $g_m = 0$ for $m = 2N+1, \ldots, M-1$, we have

$$S_M(x_k) = e^{-i2\pi kN/M} \sum_{m=0}^{M-1} g_m\,(-1)^{m-N}\, e^{i2\pi km/M}.$$

This equation shows that $S_M(x_k)$ can be computed using a DFT inversion (along with multiplications by exponential factors). By combining DFT approximations of Fourier coefficients with this last equation, it is also possible to approximate Fourier series partial sums, or arithmetic means, or other modified partial sums. See Briggs and Henson (1995) or Walker (1996) for further details.

These calculations with DFTs are facilitated on a computer using various algorithms which are all referred to as *fast Fourier transforms* (FFTs). Using FFTs, the process of computing DFTs, and hence Fourier coefficients and Fourier series, is now practically instantaneous. This allows for rapid, so-called *real-time,* calculation of the frequency content of signals. One of the most widely used applications is in calculating *spectrograms.* A spectrogram is calculated by dividing a signal (typically a recorded, digitally sampled, audio signal) into a successive series of short duration subsignals, and performing an FFT on each subsignal. This gives a portrait of the main frequencies present in the signal as time proceeds. For example, in Fig. 15a we analyze discrete samples of the function

$$\begin{aligned} &\sin(2\nu_1\pi x)e^{-100\pi(x-0.2)^2} + [\sin(2\nu_1\pi x) + \cos(2\nu_2\pi x)] \\ &\quad \times e^{-50\pi(x-0.5)^2} + \sin(2\nu_2\pi x)e^{-100\pi(x-0.8)^2} \end{aligned} \tag{41}$$

where the frequencies ν_1 and ν_2 of the sinusoidal factors are 128 and 256, respectively. The signal is graphed at the bottom of Fig. 15a and the magnitudes of the values of its spectrogram are graphed at the top. The more intense spectrogram magnitudes are shaded more darkly, while white regions indicate magnitudes that are essentially zero. The dark blobs in the graph of the spectrogram magnitudes clearly correspond to the regions of highest energy in the signal and are centered on the frequencies 128 and 256, the two frequencies used in Eq. (41).

As a second example, we show in Fig. 15b the spectrogram magnitudes for the signal

$$e^{-5\pi[(x-0.5)/0.4]^{10}}[\sin(400\pi x^2) + \sin(200\pi x^2)\,]. \tag{42}$$

This signal is a combination of two tones with sharply increasing frequency of oscillations. When run through a sound generator, it produces a sharply rising pitch. Signals like this bear some similarity to certain bird calls, and are also used in radar. The spectrogram magnitudes for this signal are shown in Fig. 15b. We can see two, somewhat blurred, line segments corresponding to the factors $400\pi x$ and $200\pi x$ multiplying x in the two sine factors in Eq. (42).

One important area of application of spectrograms is in *speech coding.* As an example, in Fig. 16 we show spectrogram magnitudes for two audio recordings. The spectrogram magnitudes in Fig. 16a come from a recording of a four-year-old girl singing the phrase "twinkle, twinkle, little star," and the spectrogram magnitudes in Fig. 16b come from a recording of the author of this article singing the same phrase. The main frequencies are seen to be in harmonic progression (integer multiples of a lowest, fundamental frequency) in both cases, but the young girl's main frequencies are higher (higher in pitch) than the adult male's. The slightly curved ribbons of frequency content are known as *formants* in linguistics. For more

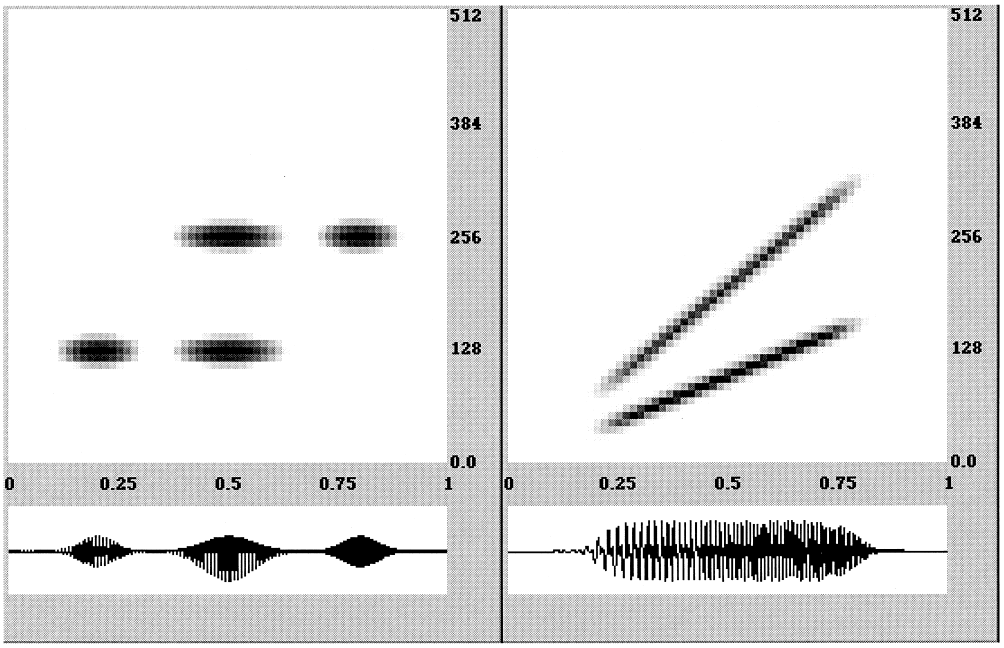


FIGURE 15 Spectrograms of test signals. (a) Bottom graph is the signal in Eq. (41). Top graph is the spectrogram magnitudes for this signal. (b) Signal and spectrogram magnitudes for the signal in (42). Horizontal axes are time values (in sec); vertical axes are frequency values (in Hz). Darker pixels denote larger magnitudes, white pixels are near zero in magnitude.

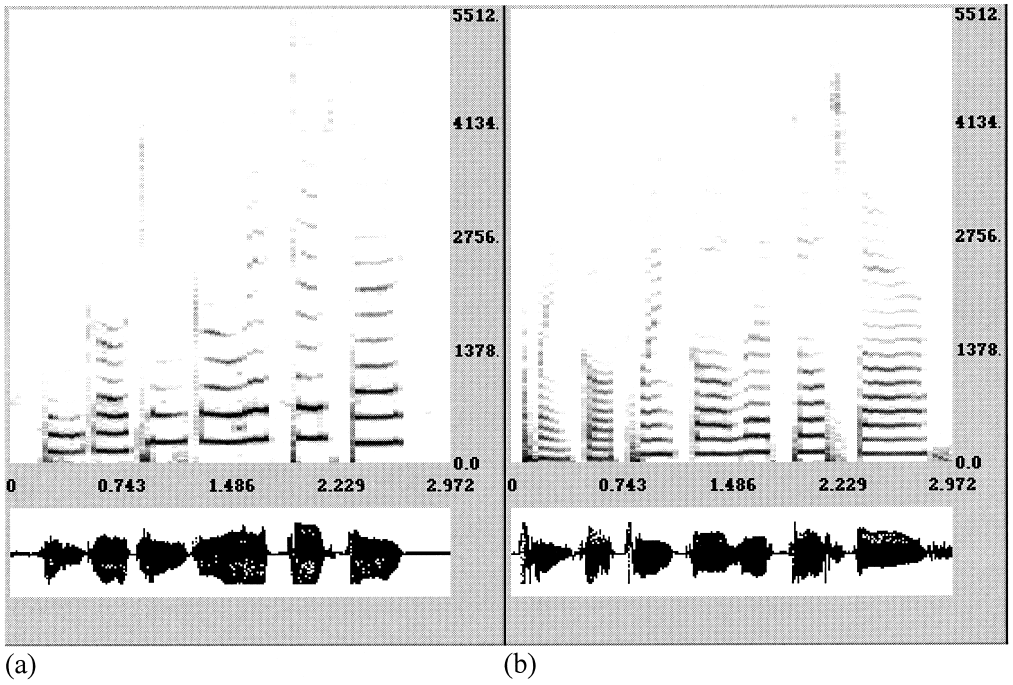


FIGURE 16 Spectrograms of audio signals. (a) Bottom graph displays data from a recording of a young girl singing "twinkle, twinkle, little star." Top graph displays the spectrogram magnitudes for this recording. (b) Similar graphs for the author's rendition of "twinkle, twinkle, little star."

details on the use of spectrograms in signal analysis, see Mallat (1998).

It is possible to invert spectrograms. In other words, we can recover the original signal by inverting the succession of DFTs that make up its spectrogram. One application of this inverse procedure is to the *compression of audio signals.* After discarding (setting to zero) all the values in the spectrogram with magnitudes below a threshold value, the inverse procedure creates an approximation to the signal which uses significantly less data than the original signal. For example, by discarding all of the spectrogram values having magnitudes less than $1/320$ times the largest magnitude spectrogram value, the young girl's version of "twinkle, twinkle, little star" can be approximated, *without noticeable degradation of quality,* using about one-eighth the amount of data as the original recording. Some of the best results in audio compression are based on sophisticated generalizations of this spectrogram technique—referred to either as *lapped transforms or as local cosine expansions*, see Malvar (1992) and Mallat (1998).

VIII. CONCLUSION

In this article, we have outlined the main features of the theory and application of one-variable Fourier series. Much additional information, however, can be found in the references. In particular, we did not have sufficient space to discuss the intricacies of multivariable Fourier series which, for example, have important applications in crystallography and molecular structure determination. For a mathematical introduction to multivariable Fourier series, see Krantz (1999), and for an introduction to their applications, see Walker (1988).

SEE ALSO THE FOLLOWING ARTICLES

FUNCTIONAL ANALYSIS • GENERALIZED FUNCTIONS • MEASURE AND INTEGRATION • NUMERICAL ANALYSIS • SIGNAL PROCESSING • WAVELETS

BIBLIOGRAPHY

Briggs, W. L., and Henson, V. E. (1995). "The DFT. An Owner's Manual," SIAM, Philadelphia.
Daubechies, I. (1992). "Ten Lectures on Wavelets," SIAM, Philadelphia.
Davis, P. J. (1975). "Interpolation and Approximation," Dover, New York.
Davis, P. J., and Hersh, R. (1982). "The Mathematical Experience," Houghton Mifflin, Boston.
Fourier, J. (1955). "The Analytical Theory of Heat," Dover, New York.
Jackson, D. (1941). "Fourier Series and Orthogonal Polynomials," Math. Assoc. of America, Washington, DC.
Krantz, S. G. (1999). "A Panorama of Harmonic Analysis," Math. Assoc. of America, Washington, DC.
Mallat, S. (1998). "A Wavelet Tour of Signal Processing," Academic Press, New York.
Malvar, H. S. (1992). "Signal Processing with Lapped Transforms," Artech House, Norwood.
Meyer, Y. (1992). "Wavelets and Operators," Cambridge Univ. Press, Cambridge.
Rudin, W. (1986). "Real and Complex Analysis," 3rd edition, McGraw-Hill, New York.
Walker, J. S. (1988). "Fourier Analysis," Oxford Univ. Press, Oxford.
Walker, J. S. (1996). "Fast Fourier Transforms," 2nd edition, CRC Press, Boca Raton.
Walker, J. S. (1999). "A Primer on Wavelets and their Scientific Applications," CRC Press, Boca Raton.
Walter, G. G. (1994). "Wavelets and Other Orthogonal Systems with Applications," CRC Press, Boca Raton.
Zygmund, A. (1968). "Trigonometric Series," Cambridge Univ. Press, Cambridge.



Fractals

Benoit B. Mandelbrot
Michael Frame
Yale University

I. Scale Invariance
II. The Generic Notion of Fractal Dimension and a Few Specific Implementations
III. Algebra of Dimensions and Latent Dimensions
IV. Methods of Computing Dimension in Mathematical Fractals
V. Methods of Measuring Dimension in Physical Systems
VI. Lacunarity
VII. Fractal Graphs and Self-Affinity
VIII. Fractal Attractors and Repellers of Dynamical Systems
IX. Fractals and Differential or Partial Differential Equations
X. Fractals in the Arts and in Teaching

GLOSSARY

Dimension An exponent characterizing how some aspect—mass, number of boxes in a covering, etc.—of an object scales with the size of the object.
Lacunarity A measure of the distribution of hole sizes of a fractal. The prefactor in the mass–radius scaling is one such measure.
Self-affine fractal A shape consisting of smaller copies of itself, all scaled by affinities, linear transformations with different contraction ratios in different directions.
Self-similar fractal A shape consisting of smaller copies of itself, all scaled by similitudes, linear transformations with the same contraction ratios in every direction.

FRACTALS have a long history: after they became the object of intensive study in 1975, it became clear that they had been used worldwide for millenia as decorative patterns. About a century ago, their appearance in pure mathematics had two effects. It led to the development of tools like fractal dimensions, but marked a turn toward abstraction that contributed to a deep and long divide between mathematics and physics. Quite independently from fundamental mathematical physics as presently defined, fractal geometry arose in equal parts from an awareness of past mathematics and a concern for practical, mundane questions long left aside for lack of proper tools.

The mathematical input ran along the lines described by John von Neumann: "A large part of mathematics which became useful developed with absolutely no desire to be useful . . . This is true for all science. Successes were largely due to . . . relying on . . . intellectual elegance. It was by following this rule that one actually got ahead in the long run, much better than any strictly utilitarian course would have permitted . . . The principle of laissez-faire has led to strange and wonderful results."

The influence of mundane questions grew to take on far more importance than was originally expected, and recently revealed itself as illustrating a theme that is common

in science. Every science started as a way to organize a large collection of messages our brain receives from our senses. The difficulty is that most of these messages are very complex, and a science can take off only after it succeeds in identifying special cases that allow a workable first step. For example, acoustics did not take its first step with chirps or drums but with idealized vibrating strings. These led to sinusoids and constants or other functions invariant under translation in time. For the notion of roughness, no proper measure was available only 20 years ago. The claim put forward forcibly in Mandelbrot (1982) is that a workable entry is provided by rough shapes that are dilation invariant. These are fractals.

Fractal roughness proves to be ubiquitous in the works of nature and man. Those works of man range from mathematics and the arts to the Internet and the financial markets. Those works of nature range from the cosmos to carbon deposits in diesel engines. A sketchy list would be useless and a complete list, overwhelming. The reader is referred to Frame and Mandelbrot (2001) and to a Panorama mentioned therein, available on the web. This essay is organized around the mathematics of fractals, and concrete examples as illustrations of it.

To avoid the need to discuss the same topic twice, mathematical complexity is allowed to fluctuate up and down. The reader who encounters paragraphs of oppressive difficulty is urged to skip ahead until the difficulty becomes manageable.

I. SCALE INVARIANCE

A. On Choosing a "Symmetry" Appropriate to the Study of Roughness

The organization of experimental data into simple theoretical models is one of the central works of every science; invariances and the associated symmetries are powerful tools for uncovering these models. The most common invariances are those under Euclidean motions: translations, rotations, reflections. The corresponding ideal physics is that of uniform or uniformly accelerated motion, uniform or smoothly varying pressure and density, smooth submanifolds of Euclidean physical or phase space. The geometric alphabet is Euclidean, the analytical tool is calculus, the statistics is stationary and Gaussian.

Few aspects of nature or man match these idealizations: turbulent flows are grossly nonuniform; solid rocks are conspicuously cracked and porous; in nature and the stock market, curves are nowhere smooth. One approach to this discrepancy, successful for many problems, is to treat observed objects and processes as "roughened" versions of an underlying smooth ideal. The underlying geometry is Euclidean or locally Euclidean, and observed nature is written in the language of noisy Euclidean geometry.

Fractal geometry was invented to approach roughness in a very different way. Under magnification, smooth shapes are more and more closely approximated by their tangent spaces. The more they are magnified, the simpler ("better") they look. Over some range of magnifications, looking more closely at a rock or a coastline does not reveal a simpler picture, but rather more of the same kind of detail. Fractal geometry is based on this ubiquitous *scale invariance*. "A fractal is an object that doesn't look any better when you blow it up." Scale invariance is also called "symmetry under magnification."

A manifestation is that fractals are sets (or measures) that can be broken up into pieces, each of which closely resembles the whole, except it is smaller. If the pieces scale isotropically, the shape is called *self-similar*; if different scalings are used in different directions, the shape is called *self-affine*.

There are deep relations between the geometry of fractal sets and the renormalization approach to critical phenomena in statistical physics.

B. Examples of Self-Similar Fractals

1. Exact Linear Self-Similarity

A shape S is called *exactly (linearly) self-similar* if the whole S splits into the union of parts S_i: $S = S_1 \cup S_2 \cup \ldots \cup S_n$. The parts satisfy two restrictions: (a) each part S_i is a copy of the whole S scaled by a linear contraction factor r_i, and (b) the intersections between parts are empty or "small" in the sense of dimension. Anticipating Section II, if $i \neq j$, the fractal dimension of the intersection $S_i \cap S_j$ must be lower than that of S. The roughness of these sets is characterized by the *similarity dimension* d. In the special equiscaling case $r_1 = \cdots = r_n = r$, $d = \log(n)/\log(1/r)$. In general, d is the solution of the *Moran equation*

$$\sum_{i=1}^{n} r_i^d = 1.$$

More details are given in Section II.

Exactly self-similar fractals can be constructed by several elegant mathematical approaches.

a. Initiator and generator. An *initiator* is a starting shape; a *generator* is a juxtaposition of scaled copies of the initiator. Replacing the smaller copies of the initiator in the generator with scaled copies of the generator sets in motion a process whose limit is an exactly self-similar fractal. Stages before reaching the limit are called *protofractals*. Each copy is anchored by a fixed point, and one may have to specify the orientation of each replacement. The



FIGURE 1 Construction of the Sierpinksi gasket. The initiator is a filled-in equilateral triangle, and the generator (on the left) is made of $N=3$ triangles, each obtained from the initiator by a contraction map T_i of reduction ratio $r=1/2$. The contractions' fixed points are the vertices of the initiator. The middle shows the second stage, replacing each copy of the initiator with a scaled copy of the generator. On the right is the seventh stage of the construction.

Sierpinski gasket (Fig. 1) is an example. The eye spontaneously splits the whole S into parts. The simplest split yields $N=3$ parts S_i, each a copy of the whole reduced by a similitude of ratio $1/2$ and with fixed point at a vertex of the initiator. In finer subdivisions, $S_i \cap S_j$ is either empty or a single point, for which $d=0$. In this example, but not always, it can be made empty by simply erasing the topmost point of every triangle in the construction.

Some of the most familiar fractals were orignally constructed to provide instances of curves that exemplify properties deemed counterintuitive: classical curves may have one multiple point (like Fig. 8) or a few. To the contrary, the Sierpinski gasket (Fig. 2, far left) is a curve with dense multiple points. The Sierpinski carpet (Fig. 2, mid left) is a universal curve in the sense that one can embed in the carpet every plane curve, irrespective of the collection of its multiple points. The Peano curve [initiator the diagonal segment from (0, 0) to (1, 1), generator in Fig. 2 mid right] is actually not a curve but a motion. It is plane-filling: a continuous onto map $[0,1] \to [0,1] \times [0,1]$.

b. Iterated function systems. Iterated function systems (IFS) are a formalism for generating exactly self-similar fractals based on work of Hutchinson (1981) and Mandelbrot (1982), and popularized by Barnsley (1988). IFS are the foundation of a substantial industry of image compression. The basis is a (usually) finite collection $\{T_1, \ldots, T_n\}$ of contraction maps $T_i: \mathbf{R}^n \to \mathbf{R}^n$ with contraction ratios $r_i < 1$. Each T_i is assigned a probability p_i that serves, at each (discrete) instant of time, to select the next map to be used. An IFS attractor also can be viewed as the limit set of the orbit $\mathcal{O}^+(x_0)$ of any point x_0 under the action of the semigroup generated by $\{T_1, \ldots, T_n\}$.

The formal definition of IFS, which is delicate and technical, proceeds as follows. Denoting by $\mathcal{K}$ the set of nonempty compact subsets of $\mathbf{R}^n$ and by h the Hausdorff metric on $\mathcal{K}$ [$h(A, B) = \inf\{\delta: A \subset B_\delta \text{ and } B \subset A_\delta\}$, where $A_\delta = \{x \in \mathbf{R}^n: d(x, y) \le \delta \text{ for some } y \in A\}$ is the δ-*thickening* of A, and d is the Euclidean metric], the T_i together define a transformation $\mathcal{T}: \mathcal{K} \to \mathcal{K}$ by $\mathcal{T}(A) = \bigcup_{i=1}^n \{T_i(x): x \in A\}$, a contraction in the Hausdorff metric with contraction ratio $r = \max\{r_1, \ldots, r_n\}$. Because $(\mathcal{K}, h)$ is complete, the contraction mapping principle guarantees there is a unique fixed point C of $\mathcal{T}$. This fixed point is the *attractor* of the IFS $\{T_1, \ldots, T_n\}$. Moreover, for any $K \in \mathcal{K}$, the sequence $K, \mathcal{T}(K), \mathcal{T}^2(K), \ldots$ converges to C in the sense that $\lim_{n\to\infty} h(\mathcal{T}^n(K), C) = 0$.

The IFS *inverse problem* is, for a given compact set A and given tolerance $\delta > 0$, to find a set of transformations $\{T_1, \ldots, T_n\}$ with attractor C satisfying $h(A, C) < \delta$. The search for efficient algorithms to solve the inverse problem is the heart of *fractal image compression*. Detailed discussions can be found in Barnsley and Hurd (1993) and Fischer (1995).

2. Exact Nonlinear Self-Similarity

A broader class of fractals is produced if the decomposition of S into the union $S = S_1 \cup S_2 \cup \ldots \cup S_n$ allows the S_i to be the images of S under nonlinear transformations.

a. Quadratic Julia sets. For fixed complex number c, the "quadratic orbit" of the starting complex number z is a sequence of numbers that begins with $f_c(z) = z^2 + c$, then $f_c^2(z) = (f_c(z))^2 + c$ and continues by following the rule $f_c^n(z) = f_c(f_c^{n-1}(z))$. The *filled-in (quadratic) Julia set* consists of the starting points that do not iterate to infinity, formally, the points $\{z: f_c^n(z)$ remains bounded as $n \to \infty\}$. The *(quadratic) Julia set* J_c is the boundary of the filled-in Julia set. Figure 3 shows the Julia set J_c for $c = 0.4 + 0 \cdot i$ and the filled-in Julia set for $c = -0.544 + 0.576 \cdot i$. The latter has an attracting 5-cycle, the black region is the basin of attraction of the

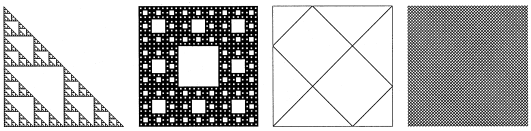


FIGURE 2 The Sierpinski gasket, Sierpinski carpet, the Peano curve generator, and the fourth stage of the Peano curve.

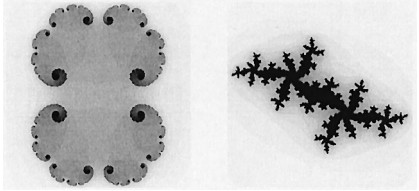


FIGURE 3 The Julia set of $z^2 + 0.4$ (left) and the filled-in Julia set for $z^2 - 0.544 + 0.576 \cdot i$ (right).

5-cycle, and the Julia set is the boundary of the black region. Certainly, J_c is invariant under f_c and under the inverses of f_c, $f_{c+}^{-1}(z) = \sqrt{z-c}$ and $f_{c-}^{-1}(z) = -\sqrt{z-c}$. Polynomial functions allow several equivalent characterizations: J_c is the closure of the set of repelling periodic points of $f_c(z)$ and J_c is the attractor of the nonlinear IFS $\{f_{c+}^{-1}, f_{c-}^{-1}\}$.

Much is known about Julia sets of quadratic functions. For example, McMullen proved that at a point whose rotation number has periodic continued-fraction expansion, the J set is asymptotically self-similar about the critical point.

The J sets are defined for functions more general than polynomials. Visually striking and technically interesting examples correspond to the Newton function $N_f(z) = z - f(z)/f'(z)$ for polynomial families $f(z)$, or entire functions like $\lambda \sin z$, $\lambda \cos z$, or $\lambda \exp z$ (see Section VIII.B). Discussions can be found in Blanchard (1994), Curry *et al.* (1983), Devaney (1994), Keen (1994), and Peitgen (1989).

b. The Mandelbrot set. The quadratic orbit $f_c^n(z)$ always converges to infinity for large enough values of z. Mandelbrot attempted a computer study of the set M^0 of those values of c for which the orbit does *not* converge to infinity, but to a stable cycle. This approach having proved unrewarding, he moved on to a set that promised an easier calculation and proved spectacular. Julia and Fatou, building on fundamental work of Montel, had shown that the Julia set J_c of $f_c(z) = z^2 + c$ must be either connected or totally disconnected. Moreover, J_c is connected if, and only if, the orbit $\mathcal{O}^+(0)$ of the critical point $z = 0$ remains bounded. The set M defined by $\{c\colon f_c^n(0)$ remains bounded$\}$ is now called the *Mandelbrot set* (see the left side of Fig. 4). Mandelbrot (1980) performed a computer investigation of its structure and reported several observations. As is now well known, small copies of the M set are infinitely numerous and dense in its boundary. The right side of Fig. 4 shows one such small copy, a nonlinearly distorted copy of the whole set. Although the small copy on the right side of Fig. 4 appears to be an isolated "island," Mandelbrot conjectured and Douady and Hubbard (1984) proved that the M set is connected. Sharpening an observation by Mandelbrot, Tan Lei (1984) proved the convergence of appropriate magnifications of Julia sets and the M set at certain points named after Misiurewicz. Shishikura (1994) proved Mandelbrot's (1985) and Milnor's (1989) conjecture that the boundary of the M set has Hausdorff dimension 2. Lyubich proved that the boundary of the M set is asymptotically self-similar about the Feigenbaum point.

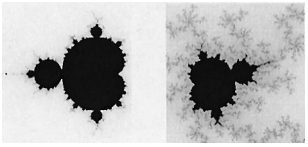


FIGURE 4 Left: The Mandelbrot set. Right: A detail of the Mandelbrot set showing a small copy of the whole. Note the nonlinear relation between the whole and the copy.

Mandelbrot's first conjecture, that the interior of the M set consists entirely of components (called *hyperbolic*) for which there is a stable cycle, remains unproved in general, though McMullen (1994) proved it for all such components that intersect the real axis. Mandelbrot's notion that M may be the closure of M^0 is equivalent to the assertion that the M set is locally connected. Despite intense efforts, that assertion remains a conjecture, though Yoccoz and others have made progress.

Other developments include the theory of quadratic-like maps (Douady and Hubbard, 1985), implying the universality and ubiquity of the M set. This result was presaged by the discovery (Curry *et al.*, 1983) of a Mandelbrot set in the parameter space of Newton's method for a family of cubic polynomials.

The recent book by Tan Lei (2000) surveys current results and attests to the vitality of this field.

c. Circle inversion limit sets. Inversion I_C in a circle C with center O and radius r transforms a point P into the point P' lying on the ray OP and with $d(O, P) \cdot d(O, P') = r^2$. This is the orientation-reversing involution defined on $\mathbf{R}^2 \cup \{\infty\}$ by $P \to I_C(P) = P'$. Inversion in C leaves C fixed, and interchanges the interior and exterior of C. It contracts the "outer" component not containing O, but the contraction ratio is not bounded by any $r < 1$.

Poincaré generalized from inversion in one circle to a collection of more than one inversion. As an example, consider a collection of circles $C_1, \ldots, C_N$ each of which is external to all the others. That is, for all $j \neq i$, the disks bounded by C_i and C_j have disjoint interiors. The *limit set* $\Lambda(C_1, \ldots, C_N)$ of inversion in these circles is the set of limit points of the orbit $\mathcal{O}^+(P)$ of any point P, external to $C_1, \ldots, C_N$, under the group generated by $I_{C_1}, \ldots, I_{C_N}$. Equivalently, it is the set left invariant by every one of the inversions $I_{C_1}, \ldots, I_{C_N}$.

The limit set Λ is nearly always fractal but the nonlinearity of inversion guarantees that Λ is nonlinearly self-similar. An example is shown in Fig. 5: the part of the limit set inside C_1 is easily seen to be the transform by I_1 of the part of the limit set inside C_2, C_3, C_4, and C_5.

How can one draw the limit set Λ when the arrangement of the circles $C_1, \ldots, C_N$ is more involved? Poincaré's original algorithm converges extraordinarily slowly. The first alternative algorithm was advanced in Mandelbrot

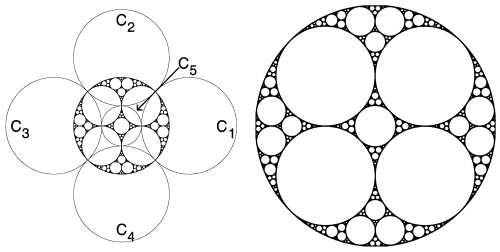


FIGURE 5 Left: The limit set generated by inversion in the five circles $C_1, \ldots, C_5$. Right: A magnification of the limit set.

(1982, Chapter 18); it is intuitive and the large-scale features of Λ appear very rapidly, followed by increasingly fine features down to any level a computer's memory can support.

d. Kleinian group limit sets. A Kleinian group (Beardon, 1983; Maskit, 1988) is a discrete group of Möbius transformations

$$z \to \frac{az+b}{cz+d}$$

acting on the Riemann sphere $\hat{\mathbf{C}}$, the sphere at infinity of hyperbolic 3-space $\mathbf{H}^3$. The isometries of $\mathbf{H}^3$ can be represented by complex matrices

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

[More precisely, by their equivalence classes in $PSL_2(\mathbf{C})$.] Sullivan's side-by-side dictionary (Sullivan, 1985) between Kleinian groups and iterates of rational maps is another deep mathematical realm informed, at least in part, by fractal geometry. Thurston's "geometrization program" for 3-manifolds (Thurston, 1997) involves giving many 3-manifolds hyperbolic structures by viewing them as quotients of $\mathbf{H}^3$ by the action of a Kleinian group G (Epstein, 1986). The corresponding action of G on $\hat{\mathbf{C}}$ determines the limit set $\Lambda(G)$, defined as the intersection of all nonempty G-invariant subsets of $\hat{\mathbf{C}}$. For many G, the limit set is a fractal. An example gives the flavor of typical results: the limit set of a finitely generated Kleinian group is either totally disconnected, a circle, or has Hausdorff dimension greater than 1 (Bishop and Jones, 1997). The Hausdorff dimension of the limit set has been studied by Beardon, Bishop, Bowen, Canary, Jones, Keen, Mantica, Maskit, McMullen, Mumford, Parker, Patterson, Sullivan, Tricot, Tukia, and many others. Poincaré exponents, eigenvalues of the Laplacian, and entropy of geodesic flows are among the tools used.

Figure 5 brings forth a relation between some limit sets of inversions or Kleinian groups and Apollonian packings (Keen *et al.*, 1993). In fact, the limit set of the Kleinian groups that are in the Maskit embedding (Keen and Series, 1993) of the Teichmüller space of any finite-type Riemann surface are Apollonian packings. These correspond to hyperbolic 3-manifolds having totally geodesic boundaries. McShane *et al.* (1994) used automatic group theory to produce efficient pictures of these limit sets, and Parker (1995) showed that in many cases the Hausdorff dimension of the limit set equals the circle packing exponent, easily estimated as the slope of the log–log plot of the number of circles of radius $\geq r$ (y axis) versus r (x axis).

Limit sets of Kleinian group actions are an excellent example of a deep, subtle, and very active area of pure mathematics in which fractals play a central role.

3. Statistical Self-Similarity

A tree's branches are not exact shrunken copies of that tree, inlets in a bay are not exact shrunken copies of that bay, nor is each cloud made up of exact smaller copies of that cloud. To justify the role of fractal geometry as a geometry of nature, one must take a step beyond exact self-similarity (linear or otherwise). Some element of randomness appears to be present in many natural objects and processes. To accommodate this, the notions of self-similarity and self-affinity are made statistical.

a. Wiener brownian motion: its graphs and trails. The first example is classical. It is one-dimensional Brownian motion, the random process $X(t)$ defined by these properties: (1) with probability 1, $X(0) = 0$ and $X(t)$ is continuous, and (2) the increments $X(t+\Delta t) - X(t)$ of $X(t)$ are Gaussian with mean 0 and variance Δt. That is,

$$Pr\{X(t+\Delta t) - X(t) \leq x\} = \frac{1}{\sqrt{2\pi\,\Delta t}} \times \int_{-\infty}^{x} \exp\left(\frac{-u^2}{2\Delta t}\right) du.$$

An immediate consequence is independence of increments over disjoint intervals. A fundamental property of Brownian motion is statistical self-affinity: for all $s > 0$,

$$Pr\{X(s(t+\Delta t)) - X(st) \leq \sqrt{s}x\} = Pr\{X(t+\Delta t) - X(t) \leq x\}.$$

That is, rescaling t by a factor of s, and of x by a factor of $\sqrt{s}$, leaves the distribution unchanged. This correct rescaling is shown on the left panel of Fig. 6: t (on the horizontal axis) is scaled by 4, x (on the vertical axis) is scaled by $2 = 4^{1/2}$. Note that this magnification has about the same degree of roughness as the full picture. In the center panel, t is scaled by 4, x by 4/3; the magnification is flatter than the original. In the right panel, both t and

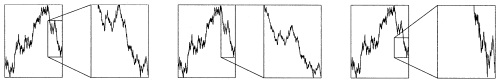


FIGURE 6 Left panel: Correct rescaling illustrating the self-affinity of Brownian motion. Center and right panels: Two incorrect rescalings.

x are scaled by 4; the magnification is steeper than the original.

A sequence of increments of Brownian motion is called Gaussian white noise. Even casual inspection of the graph reveals some fundamental features. The width of an old pen-plotter line being equal to the spacing between successive difference values, the bulk of the difference plot merges into a "band" with the following properties (see Fig. 7):

- The band's width is approximately constant.
- The values beyond that band stay close to it (this is due to the fact that the Gaussian has "short tails").
- The values beyond that band do not cluster.

Positioning E independent one-dimensional Brownian motions along E coordinate axes gives a higher dimensional Brownian motion: $B(t) = \{X_1(t), \ldots, X_E(t)\}$. Plotted as a curve in E-dimensional space, the collection of points that $B(t)$ visits between $t = 0$ and $t = 1$ defines a *Brownian trail*.

When $E > 2$, this is an example of a statistically self-similar fractal. To split the Brownian trail into N reduced-scale parts, pick $N - 1$ arbitrary instants t_n with $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = 1$. The Brownian trail for $0 \le t \le 1$ splits into N subtrails B_n for the interval $t_{n-1} < t < t_n$. The parts B_n follow the same statistical distribution as the whole, after the size is expanded by $(t_n - t_{n-1})^{-1/2}$ in every direction.

Due to the definition of self-similarity, this example reveals a pesky complication: for $i \neq j$, $B_i \cap B_j$ must be of dimension less than B. This is indeed the case if $E > 2$, but not in the plane $E = 2$. However, the overall idea can be illustrated for $E = 2$. The right side of Fig. 8 shows $B_1(t)$ for $0 \le t \le 1/4$, expanded by a factor of $2 = (1/4 - 0)^{-1/2}$ and with additional points interpolated so the part and the whole exhibit about the same number of turns. The details for $E = 2$ are unexpectedly complex, as shown in Mandelbrot (2001b, Chapter 3).

The dimensions (see Sections II.B and II.E) of some Brownian constructions are well-known, at least in most cases. For example, with probability 1:



FIGURE 7 Plot of 4000 successive Brownian increments.

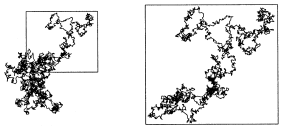


FIGURE 8 A Brownian trail. Right: The first quarter of the left trail, magnified and with additional turns interpolated so the left and right pictures have about the same number of turns.

- For $E \ge 2$ a Brownian trail $B: [0, 1] \to \mathbf{R}^E$ has Hausdorff and box dimensions $d_{\mathrm{H}} = d_{\mathrm{box}} = 2$, respectively.
- The graph of one-dimensional Brownian motion $B: [0, 1] \to \mathbf{R}$ has $d_{\mathrm{H}} = d_{\mathrm{box}} = 3/2$.

Some related constructions have been more resistant to theoretical analysis. Mandelbrot's *planar Brownian cluster* is the graph of the complex $B(t)$ constrained to satisfy $B(0) = B(1)$. It can be constructed by linearly detrending the x- and y-coordinate functions: $(X(t) - tX(1), Y(t) - tY(1))$. See Fig. 9. The cluster is known to have dimension 2. Visual inspection supported by computer experiments led to the $4/3$ *conjecture*, which asserts that the boundary of the cluster has dimension $4/3$ (Mandelbrot, 1982, p. 243). This has been proved by Lawler *et al.* (2000).

Brownian motion is the unique stationary random process with increments independent over disjoint intervals and with finite variance. For many applications, these conditions are too restrictive, drawing attention to other random processes that retain scaling but abandon either independent increments or finite variance.

b. Fractional Brownian motion. For fixed $0 < H < 1$, *fractional Brownian motion* (FBM) of exponent H is a random process $X(t)$ with increments $X(t + \Delta t) - X(t)$ following the Gaussian distribution with mean 0 and standard deviation $(\Delta t)^H$. Statistical self-affinity is straightforward: for all $s > 0$

$$Pr\{X(s(t + \Delta t)) - X(st) \le s^H x\}$$
$$= Pr\{X(t + \Delta t) - X(t) \le x\}.$$

The correlation is the expected value of the product of successive increments. It equals



FIGURE 9 A Brownian cluster.

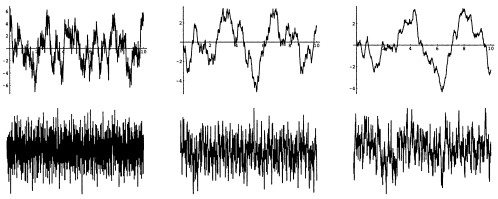


FIGURE 10 Top: Fractional Brownian motion simulations with $H = 0.25$, $H = 0.5$, and $H = 0.75$. Bottom: Difference plots $X(t+1) - X(t)$ of the graphs above.

$$E((X(t) - X(0)) \cdot (X(t+h) - X(t))) = \tfrac{1}{2}((t+h)^{2H} - t^{2H} - h^{2H}).$$

If $H = 1/2$, this correlation vanishes and the increments are independent. In fact, FBM reduces to Brownian motion. If $H > 1/2$, the correlation is positive, so the increments tend to have the same sign. This is *persistent FBM*. If $H < 1/2$, the correlation is negative, so the increments tend to have opposite signs. This is *antipersistent FBM*. See Fig. 10. The exponent determines the dimension of the graph of FBM: with probability 1, $d_{\mathrm{H}} = d_{\mathrm{box}} = 2 - H$. Notice that for $H > 1/2$, the central band of the difference plot moves up and down, a sign of long-range correlation, but the outliers still are small. Figure 11 shows the trails of these three flavors of FBM. FBM is the main topic of Mandelbrot (2001c).

c. Lévy stable processes. While FBM introduces correlations, its increments remain Gaussian and so have small outliers. The Gaussian distribution is characterized by its first two moments (mean and variance), but some natural phenomena appear to have distributions for which these are not useful indicators. For example, at the critical point of percolation there are clusters of all sizes and the expected cluster size diverges.

Paul Lévy studied random walks for which the jump distributions follow the power law $Pr\{X > x\} \approx x^{-\alpha}$. There is a geometrical approach for generating examples of Lévy processes.

The *unit step function* $\xi(t)$ is defined by

$$\xi(t) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$$

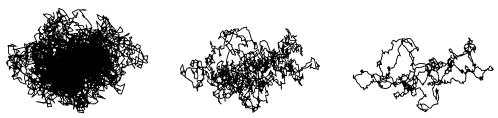


FIGURE 11 Top: Fractional Brownian motion simulations with $H = 0.25$, $H = 0.5$, and $H = 0.75$. Bottom: Difference plots $X(t+1) - X(t)$ of the graphs above.



FIGURE 12 Lévy flight on the line. Left: the graph as a function of time. Right, the increments.

and a (one-dimensional) Lévy stable process is defined as a sum

$$f(t) = \sum_{k=1}^{\infty} \lambda_k \xi(t - t_k),$$

where the pulse times t_n and amplitudes λ_n are chosen according to the following Lévy measure: given t and λ, the probability of choosing (t_i, λ_i) in the rectangle $t < t_i < t + dt$, $\lambda < \lambda_i < \lambda + d\lambda$ is $C\lambda^{-\alpha-1}\, d\lambda\, dt$. Figure 12 shows the graph of a Lévy process or flight, and a graph of its increments.

Comparing Figs. 7, 10, and 12 illustrates the power of the increment plot for revealing both global correlations (FBM) and long tails (Lévy processes).

The effect of large excursions in Lévy processess is more visible in the plane. See Fig. 13. These Lévy flights were used in Mandelbrot (1982, Chapter 32) to mimic the statistical properties of galaxy distributions.

Using fractional Brownian motion and Lévy processes, Mandelbrot (in 1965 and 1963) improved upon Bachelier's Brownian model of the stock market. The former corrects the independence of Brownian motion, the latter corrects its short tails. The original and corrected processes in the preceding sentence are statistically self-affine random fractal processes. This demonstrates the power of invariances in financial modeling; see Mandelbrot (1997a,b).

d. Self-affine cartoons with mild to wild randomness. Many natural processes exhibit long tails or global dependence or both, so it was a pleasant surprise that both can be incorporated in an elegant family of simple cartoons. (Mandelbrot, 1997a, Chapter 6; 1999, Chapter N1; 2001a). Like for self-similar curves (Section I.B.1.a), the basic construction of the cartoon involves an initiator and a

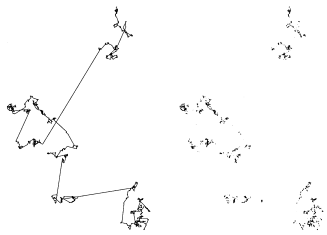


FIGURE 13 Left: Trail of the Lévy flight in the plane. Right: The Lévy dust formed by the turning points.



FIGURE 14 The initiator (left), generator (middle), and first generation (right) of the Brownian cartoon.

generator. The process used to generate the graph consists in replacing each copy of the initiator with an appropriately rescaled copy of the generator. For a Brownian cartoon, the initiator can be the diagonal of the unit square, and the generator, the broken line with vertices (0, 0), (4/9, 2/3), (5/9, 1/3), and (1, 1). Pictured in Fig. 14 are the initiator (left), generator (middle), and first iteration of the process (right).

To get an appreciation for how quickly the local roughness of these pictures increases, the left side of Fig. 15 shows the sixth iterate of the process.

Self-affinity is built in because each piece is an appropriately scaled version of the whole. In Fig. 14, the scaling ratios have been selected to match the "square root" property of Brownian motion: for each segment of the generator we have $|\Delta x_i| = (\Delta t_i)^{1/2}$.

More generally, a cartoon is called *unifractal* if there is a constant H with $|\Delta x_i| = (\Delta t_i)^H$ for each generator segment, where $0 < H < 1$. If different H are needed for different segments, the cartoon is *multifractal*.

The left side of Figure 15 is too symmetric to mimic any real data, but this problem is palliated by shuffling the order in which the three pieces of the generator are put into each scaled copy. The right side of Fig. 15 shows a Brownian cartoon randomized in this way.

Figure 16 illustrates how the statistical properties of the increments can be modified by adjusting the generator in a symmetrical fashion. Keeping fixed the endpoints (0, 0) and (1, 1), the middle turning points are changed into $(a, 2/3)$ and $(1-a, 1/3)$ for $0 < a \leq 1/2$.

e. Percolation clusters (Stauffer and Aharony, 1992). Given a square lattice of side length L and a number $p \in [0, 1]$, assign a random number $x \in [0, 1]$ to each lattice cell and fill the cell if $x \leq p$. A *cluster* is a maximal collection of filled cells, connected by sharing common edges. Three examples are shown in Fig. 17. A

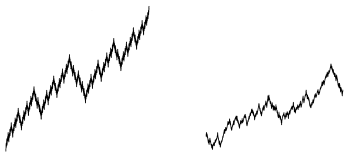


FIGURE 15 Left: The sixth iterate of the process of Fig. 14. Right: A sixth iterate of a randomized Brownian cartoon.

spanning cluster connects opposite sides of the lattice. For large L there is a *critical probability* or *percolation threshold* p_c; spanning clusters do not arise for $p < p_c$. Numerical experiments suggest $p_c \approx 0.59275$. In Fig. 17, $p = 0.4$, 0.6, and 0.8. Every lattice has its own p_c.

At $p = p_c$ the masses of the spanning clusters scale with the lattice size L as L^d, independently of the lattices. Experiment yields $d = 1.89 \pm 0.03$, and theory yields $d = 93/49$. This d is the mass dimension of Section II.C. In addition, spanning clusters have holes of all sizes; they are statistically self-similar fractals.

Many fractals are defined as part of a percolation cluster. The *backbone* is the subset of the spanning cluster that remains after removing all parts that can be separated from both spanned sides by removing a single filled cell from the spanning cluster. Numerical estimates suggest the backbone has dimension 1.61. The backbone is the path followed by a fluid diffusing through the lattice.

The *hull* of a spanning cluster is its boundary. It was observed by R. F. Voss in 1984 and proven by B. Duplantier that the hull's dimension is 7/4.

A more demanding definition of the boundary yields the *perimeter*. It was observed by T. Grossman and proven by B. Duplantier that the perimeter's dimension is 4/3.

Sapoval *et al.* (1985) examined discrete diffusion and showed that it involves a fractal diffusion front that can be modeled by the hull and the perimeter of a percolation cluster.

f. Diffusion-limited aggregation (DLA; Vicsek, 1992). DLA was proposed by Witten and Sander (1981, 1983) to simulate the aggregates that carbon particles form in a diesel engine. On a grid of square cells, a cartoon of DLA begins by occupying the center of the grid with a "seed particle." Next, place a particle in a square selected at random on the edge of a large circle centered on the seed square and let it perform a simple random walk. With each tick of the clock, with equal probabilities it will move to an adjacent square, left, right, above, or below. If the moving particle wanders too far from the seed, it falls off the edge of the grid and another wandering particle is started at a randomly chosen edge point. When a wandering particle reaches one of the four squares adjacent to the seed, it sticks to form a cluster of two particles, and another moving particle is released. When a moving particle reaches a square adjacent to the cluster, it sticks there. Continuing in this way builds an arbitrarily large object called a diffusion-limited aggregate (DLA) because the growth of the cluster is governed by the particles' diffusing across the grid. Figure 18 shows a moderate-size DLA cluster.

Early computer experiments on clusters of up to the 10^4 particles showed the mass $M(r)$ of the part of the cluster a distance r from the seed point scales as $M(r) \approx k \cdot r^d$,

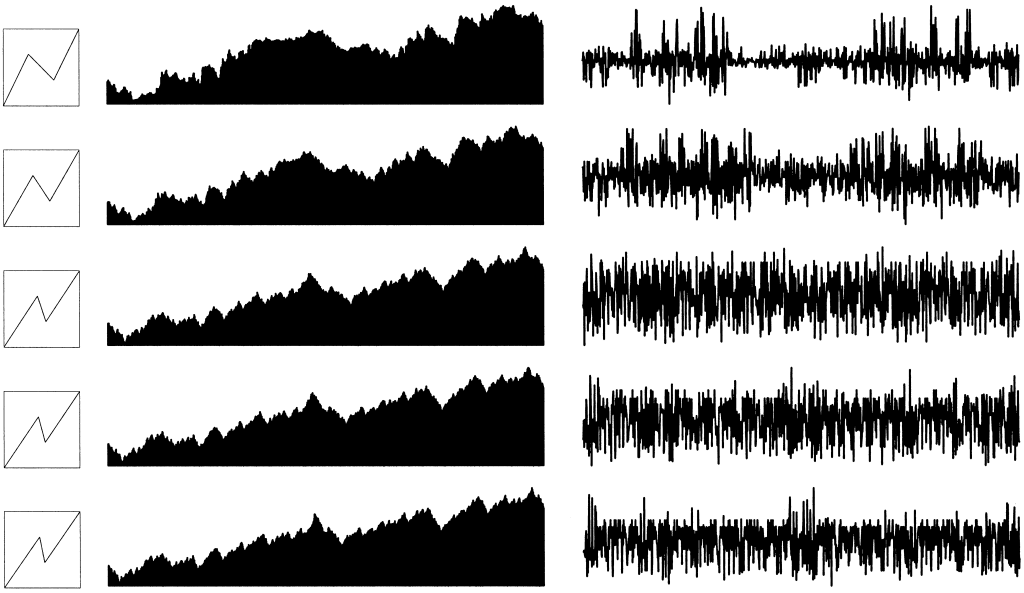


FIGURE 16 Generators, cartoons and difference graphs for symmetric cartoons with turning points $(a, 2/3)$ and $(1-a, 1/3)$, for $a = 0.333, 0.389, 0.444, 0.456$, and 0.467. The same random number seed is used in all graphs.

with $d \approx 1.71$ for clusters in the plane and $d \approx 2.5$ for clusters in space. This exponent d is the mass dimension of the cluster. (See Sections II.C and V.) These values match measured scalings of physical objects moderately, but not terribly well. A careful examination of much larger clusters revealed discrepancies that added in due time to a very complex picture of DLA. Mandelbrot *et al.* (1995) investigated clusters in the 10^7 range; careful measurement reveals an additional dimension of 1.65 ± 0.01. This suggests the clusters become more compact as they grow. Also, as the cluster grows, more arms develop and the largest gaps decrease in size; i.e., the lacunarity decreases. (See Section VI.)

II. THE GENERIC NOTION OF FRACTAL DIMENSION AND A FEW SPECIFIC IMPLEMENTATIONS

The first, but certainly not the last, step in quantifying fractals is the computation of a dimension. The notion of Euclidean dimension has many aspects and therefore extends in several fashions. The extensions are distinct in the most general cases but coincide for exactly self-similar fractals. Many other dimensions cannot be mentioned here.

A more general approach to quantifying degrees of roughness is found in the article on *Multifractals*.



FIGURE 17 Percolation lattices well below, near, and well above the percolation threshold.

A. Similarity Dimension

The definition of similarity dimension is rooted in the fact that the unit cube in D-dimensional Euclidean space is self-similar: for any positive integer b the cube can be decomposed into $N = b^D$ cubes, each scaled by the similarity ratio $r = 1/b$, and overlapping at most along $(D-1)$-dimensional cubes.

The equiscaling or isoscaling case. Provided the pieces do not overlap significantly, the power-law relation $N = (1/r)^D$ between the number N and scaling factor r of the pieces generalizes to all exactly self-similar sets with all pieces scaled by the factor r. The *similarity dimension* d_{sim} is

$$d_{\text{sim}} = \frac{\log(N)}{\log(1/r)}.$$



FIGURE 18 A moderate-size DLA cluster.

The pluriscaling case. More generally, for self-similar sets where each piece is scaled by a possibly different factor r_i, the similarity dimension is the unique positive root d of the *Moran equation*

$$\sum_{i=1}^{N} r_i^d = 1.$$

The relation $0 \le d_{\text{sim}} \le E$. If the fractal is a subset of E-dimensional Euclidean space, E is called the *embedding dimension*.

So long as the overlap of the parts is not too great (technically, under the *open set condition*), we have $d_{\text{sim}} \le E$. If at least two of the r_i are positive, we have $d_{\text{sim}} > 0$. However, Section III.F shows that some circumstances introduce a latent dimension d, related indirectly to the similarity dimension, and that can satisfy $d < 0$ or $d > E$.

B. Box Dimension

The similarity dimension is meaningful only for exactly self-similar sets. For more general sets, including experimental data, it is often replaced by the box dimension. For any bounded (nonempty) set A in E-dimensional Euclidean space, and for any $\delta > 0$, a *δ-cover* of A is a collection of sets of diameter δ whose union contains A. Denote by $N_\delta(A)$ the smallest number of sets in a δ-cover of A. Then the *box dimension* d_{box} of A is

$$d_{\text{box}} = \lim_{\delta \to 0} \frac{\log(N_\delta(A))}{\log(1/\delta)}$$

when the limit exists. When the limit does not exist, the replacement of lim with lim sup and lim inf defines the *upper* and *lower box dimensions*:

$$\overline{d_{\text{box}}} = \limsup_{\delta \to 0} \frac{\log(N_\delta(A))}{\log(1/\delta)},$$

$$\underline{d_{\text{box}}} = \liminf_{\delta \to 0} \frac{\log(N_\delta(A))}{\log(1/\delta)}.$$

The box dimension can be thought of as measuring how well a set can be covered with small boxes of equal size, because the limit (or lim sup and lim inf) remain unchanged if $N_\delta(A)$ is replaced by the smallest number of E-dimensional cubes of side δ needed to cover A, or even the number of cubes of a δ lattice that intersect A.

Section V describes methods of measuring the box dimension for physical datasets.

C. Mass Dimension

The mass $M(r)$ of a d-dimensional Euclidean ball of constant density ρ and radius r is given by

$$M(r) = \rho \cdot V(d) \cdot r^d$$

$$\text{with} \quad V(d) = \left(\Gamma\left(\tfrac{1}{2}\right)^d\right) \Big/ \Gamma\left(d + \tfrac{1}{2}\right),$$

where $V(d)$ is the volume of the d-dimensional unit sphere. That is, for constant-density Euclidean objects, the ordinary dimension—among many other roles—is the exponent relating mass to size. This role motivated the definition of mass dimension for a fractal. The definition of mass is delicate. For example, the mass of a Sierpinski gasket cannot be defined by starting with a triangle of uniform density and removing middle triangles; this process would converge to a mass reduced to 0. One must, instead, proceed as on the left side of Fig. 1: take as initiator a triangle of mass 1, and as generator three triangles each scaled by $1/2$ and of mass $1/3$. Moreover, two very new facts come up.

Firstly, the $=$ sign in the formula for $M(r)$ must be replaced by $\approx$. That is, $M(r)$ fluctuates around a multiple of r^d. For example, as mentioned in Sections I.B.3.e and I.B.3.f, the masses of spanning percolation clusters and diffusion-limited aggregates scale as a power-law function of size. Consequently, the exponent in the relation $M(r) \approx k \cdot r^{d_{\text{mass}}}$ is called the *mass dimension*.

Second, in the Euclidean case the center is arbitrary but in the fractal case it must belong to the set under consideration. As an example, Fig. 19 illustrates attempts to measure the mass dimension of the Sierpinski gasket. Suppose we take circles centered at the lower left vertex of the initiator, and having radii $1/2$, $1/4, 1/8, \ldots$. We obtain $M(1/2^i) = 1/3^i = (1/2^i)^d$, where $d_{\text{mass}} = \log 3 / \log 2$. See the left side of Fig. 19. That is, the mass dimension agrees with the similarity dimension.

On the other hand, if the circle's center is randomly selected in the interior of the initiator, the passage to the limit $r \to 0$ almost surely eventually stops with circles bounding no part of the gasket. See the middle of Fig. 19.

Taking a family of circles with center c a point of the gasket, the mass–radius relation becomes $M(r) = k(r, c) \cdot r^d$, where the prefactor $k(r, c)$ fluctuates and depends on both r and c. See the right side of Fig. 19. Even in this case, the exponent is the mass dimension. The prefactor is no longer a constant density, but a random variable, depending on the choice of the origin.

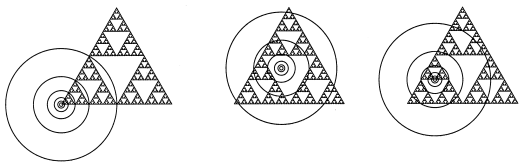


FIGURE 19 Attempts at measuring the mass dimension of a Sierpinski gasket using three families of circles.

Section V describes methods of measuring the mass dimension for physical datasets.

D. Minkowski–Bouligand Dimension

Given a set $A \subset \mathbf{R}^E$ and $\delta > 0$, the *Minkowski sausage* of A, also called the *δ-thickening* or *δ-neighborhood* of A, is defined as $A_\delta = \{x \in \mathbf{R}^E : d(x, y) \le \delta \text{ for some } y \in A\}$. (See Section I.B.b.) In the Euclidean case when A is a smooth m-dimensional manifold imbedded in $\mathbf{R}^E$, one has $vol(A_\delta) \sim \Lambda \cdot \delta^{E-m}$. That is, the E-dimensional volume of A_δ scales as δ to the codimension of A. This concept extends to fractal sets A: if the limit exists,

$$E - \lim_{\delta \to 0} \frac{\log(vol(A_\delta))}{\log(\delta)}$$

defines the *Minkowski–Bouligand dimension*, $d_{\mathrm{MB}}(A)$ (see Mandelbrot, 1982, p. 358). In fact, it is not difficult to see that $d_{\mathrm{MB}}(A) = d_{\mathrm{box}}(A)$. If the limit does not exist, lim sup gives $\underline{d_{\mathrm{box}}}(A)$ and lim inf gives $\overline{d_{\mathrm{box}}}(A)$.

In the privileged case when the limit

$$\lim_{\delta \to 0} \frac{vol(A_\delta)}{\delta^{E-m}}$$

exists, it generalizes the notion of Minkowski content for smooth manifolds A. Section VI will use this prefactor to measure lacunarity.

E. Hausdorff–Besicovitch Dimension

For a set A in Euclidean space, given $s \ge 0$ and $\delta > 0$, consider the quantity

$$\mathcal{H}^s_\delta(A) = \inf\left\{ \sum_i |U_i|^s : \{U_i\} \text{ is a } \delta\text{-cover of } A \right\}.$$

A decrease of δ reduces the collection of δ-covers of A, therefore $\mathcal{H}^s_\delta(A)$ increases as $\delta \to 0$ and $\mathcal{H}^s(A) = \lim_{\delta \to 0} \mathcal{H}^s_\delta(A)$ exists. This limit defines the *s-dimensional Hausdorff measure* of A. For $t > s$, $\mathcal{H}^t_\delta(A) \le \delta^{t-s} \mathcal{H}^s_\delta(A)$. It follows that a unique number d_{H} has the property that

$$s < d_{\mathrm{H}} \quad \text{implies} \quad \mathcal{H}^s(A) = \infty$$

and

$$s > d_{\mathrm{H}} \quad \text{implies} \quad \mathcal{H}^s(A) = 0.$$

That is,

$$d_{\mathrm{H}}(A) = \inf\{s : \mathcal{H}^s(A) = 0\} = \sup\{s : \mathcal{H}^s(A) = \infty\}.$$

This quantity d_{H} is the *Hausdorff–Besicovitch dimension* of A. It is of substantial theoretical significance, but in most cases is quite challenging to compute, even though it suffices to use coverings by disks. An upper bound often is relatively easy to obtain, but the lower bound can be much more difficult because the inf is taken over the collection of all δ-covers. Because of the inf that enters in its definition, the Hausdorff–Besicovitch dimension cannot be measured for any physical object.

Note: If A can be covered by $N_\delta(A)$ sets of diameter at most δ, then $\mathcal{H}^s_\delta(A) \le N_\delta(A) \cdot \delta^s$. From this it follows $d_{\mathrm{H}}(A) \le \underline{d_{\mathrm{box}}}(A)$, so $d_{\mathrm{H}}(A) \le d_{\mathrm{box}}(A)$ if $d_{\mathrm{box}}(A)$ exists. This inequality can be strict. For example, if A is any countable set, $d_{\mathrm{H}}(A) = 0$ and yet $d_{\mathrm{box}}(\text{rationals in } [0, 1]) = 1$.

F. Packing Dimension

Hausdorff dimension measures the efficiency of covering a set by disks of varying radius. Tricot (1982) introduced packing dimension to measure the efficiency of packing a set with disjoint disks of varying radius. Specifically, for $\delta > 0$ a *δ-packing* of A is a countable collection of disjoint disks $\{B_i\}$ with radii $r_i < \delta$ and with centers in A. In analogy with Hausdorff measure, define

$$\mathcal{P}^s_\delta(A) = \sup\left\{ \sum_i |B_i|^s : \{B_i\} \text{ is a } \delta\text{-packing of } A \right\}.$$

As δ decreases, so does the collection of δ-packings of A. Thus $\mathcal{P}^s_\delta(A)$ decreases as δ decreases and the limit

$$\mathcal{P}^s_0(A) = \lim_{\delta \to 0} \mathcal{P}^s_\delta(A)$$

exists. A technical complication requires an additional step. The *s-dimensional packing measure* of A is defined as

$$\mathcal{P}^s(A) = \inf\left\{ \sum_i \mathcal{P}^s_0(A_i) : A \subset \bigcup_{i=1}^{\infty} A_i \right\}.$$

Then the *packing dimension* $d_{\mathrm{pack}}(A)$ is

$$d_{\mathrm{pack}}(A) = \inf\{s : \mathcal{P}^s(A) = 0\} = \sup\{s : \mathcal{P}^s(A) = \infty\}.$$

Packing, Hausdorff, and box dimensions are related:

$$d_{\mathrm{H}}(A) \le d_{\mathrm{pack}}(A) \le \overline{d_{\mathrm{box}}}(A).$$

For appropriate A, each inequality is strict.

III. ALGEBRA OF DIMENSIONS AND LATENT DIMENSIONS

The dimensions of ordinary Euclidean sets obey several rules of thumb that are widely used, though rarely stated explicitly. For example, the union of two sets of dimension d and d' usually has dimension $\max\{d, d'\}$. The projection of a set of dimension d to a set of dimension d' usually gives a set of dimension $\min\{d, d'\}$. Also, for Cartesian products, the dimensions usually add:

$\dim(A \times B) = \dim(A) + \dim(B)$. For the intersection of subsets A and B of $\mathbf{R}^E$, it is the codimensions that usually add: $E - \dim(A \cap B) = (E - \dim(A)) + (E - \dim(B))$, but only so long as the sum of the codimensions is nonnegative. If this sum is negative, the intersection is empty. Mandelbrot (1984, Part II) generalized those rules to fractals and (see Section III.G) interpreted negative dimensions as measures of "degree of emptiness."

For simplicity, we restrict our attention to generaling these properties to the Hausdorff and box dimensions of fractals.

A. Dimension of Unions and Subsets

Simple applications of the definition of Hausdorff dimension give

$$A \subseteq B \quad \text{implies} \quad d_H(A) \leq d_H(B)$$

and

$$d_H(A \cup B) = \max\{d_H(A), d_H(B)\}.$$

Replacing max with sup, this property holds for countable collections of sets. The subset and finite union properties hold for box dimension, but the countable union property fails.

B. Product and Sums of Dimensions

For all subsets A and B of Euclidean space, $d_H(A \times B) \geq d_H(A) + d_H(B)$. Equality holds if one of the sets is sufficiently regular. For example, if $d_H(A) = \overline{d_H}(A)$, then $d_H(A \times B) = d_H(A) + d_H(B)$. Equality does not always hold: Besicovitch and Moran (1945) give an example of subsets A and B of $\mathbf{R}$ with $d_H(A) = d_H(B) = 0$, yet $d_H(A \times B) = 1$.

For upper box dimensions, the inequality is reversed: $\overline{d_{box}}(A \times B) \leq \overline{d_{box}}(A) + \overline{d_{box}}(B)$.

C. Projection

Denote by $\text{proj}_P(A)$ the projection of a set $A \subset \mathbf{R}^3$ to a plane $P \subset \mathbf{R}^3$ through the origin. If A is a one-dimensional Euclidean object, then for almost all choices of the plane P, $\text{proj}_P(A)$ is one-dimensional. If A is a two- or three-dimensional Euclidean object, then for almost all choices of the plane P, $\text{proj}_P(A)$ is two-dimensional of positive area. That is, $\dim(\text{proj}_P(A)) = \min\{\dim(A), \dim(P)\}$.

The analogous properties hold for fractal sets A. If $d_H(A) < 2$, then for almost all choices of the plane P, $d_H(\text{proj}_P(A)) = d_H(A)$. If $d_H(A) \geq 2$, then for almost all choices of the plane P, $d_H(\text{proj}_P(A)) = 2$ and $\text{proj}_P(A)$ has positive area. So again, $d_H(\text{proj}_P(A)) = \min\{d_H(A), d_H(P)\}$.

The obvious generalization holds for fractals $A \subset \mathbf{R}^E$ and projections to k-dimensional hyperplanes through the origin.

Projections of fractals can be very complicated. There are fractal sets $A \subset \mathbf{R}^3$ with the surprising property that for almost every plane P through the origin, the projection $\text{proj}_P(A)$ is any prescribed shape, to within a set of area 0. Consequently, as Falconer (1987) points out, in principle we could build a fractal digital sundial.

D. Subordination and Products of Dimension

We have already seen operations realizing the sum, max, and min of dimensions, and in the next subsection we shall examine the sum of codimensions. For certain types of fractals, multiplication of dimensions is achieved through "subordination," a process introduced in Bochner (1955) and elaborated in Mandelbrot (1982). Examples are constructed easily from the Koch curve generator (Fig. 20a). The initiator (the unit interval) is unchanged, but the new generator is a subset of the original generator. Figure 20 shows three examples.

In Fig. 20, generator (b) gives a fractal dust (B) of dimension $\log 3/\log 3 = 1$. Generator (c) gives the standard Cantor dust (C) of dimension $\log 2/\log 3$. Generator (d) gives a fractal dust (D) also of dimension $\log 2/\log 3$. Thinking of the Koch curve K as the graph of a function $f\colon [0, 1] \to K \subset \mathbf{R}^2$, the fractal (B) can be obtained by restricting f to the Cantor set with initiator [0, 1] and generator the intervals [0, 1/4], [1/4, 1/2], and [1/2, 3/4]. In this case, the *subordinand* is a Koch curve, the *subordinator* is a Cantor set, and the *subordinate* is the fractal (B). The identity

$$\frac{\log 3}{\log 3} = \frac{\log 4}{\log 3} \cdot \frac{\log 3}{\log 4}$$

expresses that the dimensions multiply,

$$\dim(\text{subordinate}) = \dim(\text{subordinand}) \cdot \dim(\text{subordinator}).$$

Figure 20, (C) and (D) give other illustrations of this multiplicative relation. The *seeded universe* model of the distribution of galaxies (Section IX.D.1) uses subordination to obtain fractal dusts; see Mandelbrot (1982, plate 298).

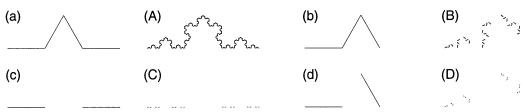


FIGURE 20 The Koch curve (A) and its generator (a); (b), (c), and (d) are subordinators, and the corresponding subordinates of the subordinand (A) are (B), (C), and (D).

E. Intersection and Sums of Codimension

The dimension of the intersection of two sets obviously depends on their relative placement. When $A \cap B = \emptyset$, the dimension vanishes. The following is a typical result. For Borel subsets A and B of $\mathbf{R}^E$, and for almost all $x \in \mathbf{R}^E$,

$$d_{\mathrm{H}}(A \cap (B + x)) \leq \max\{0, d_{\mathrm{H}}(A \times B) - E\}.$$

If $d_{\mathrm{H}}(A \times B) = d_{\mathrm{H}}(A) + d_{\mathrm{H}}(B)$, this reduces to

$$d_{\mathrm{H}}(A \cap (B + x)) \leq \max\{0, d_{\mathrm{H}}(A) + d_{\mathrm{H}}(B) - E\}.$$

This is reminiscent of the transversality relation for intersections of smooth manifolds.

Corresponding lower bounds are known in more restricted circumstances. For example, there is a positive measure set M of similarity transformations of $\mathbf{R}^E$ with

$$d_{\mathrm{H}}(A \cap T(B)) \geq d_{\mathrm{H}}(A) + d_H(B) - E$$

for all $T \in M$. Note $d_{\mathrm{H}}(A \cap T(B)) = d_{\mathrm{H}}(A) + d_{\mathrm{H}}(B) - E$ is equivalent to the addition of codimensions: $E - d_{\mathrm{H}}(A \cap T(B)) = (E - d_{\mathrm{H}}(A)) + (E - d_{\mathrm{H}}(B))$.

F. Latent Dimensions below 0 or above E

A blind application of the rule that codimensions are additive easily yields results that seem nonsensical, yet become useful if they are properly interpreted and the Hausdorff dimension is replaced by a suitable new alternative.

1. Negative Latent Dimensions as Measures of the "Degree of Emptiness"

Section E noted that if the codimension addition rule gives a negative dimension, the actual dimension is 0. This exception is an irritating complication and hides a feature worth underlining.

As background relative to the plane, consider the following intersections of two Euclidean objects: two points, a point and a line, and two lines. Naive intuition tells us that the intersection of two points is emptier than the intersection of a point and a line, and that the latter in turn is emptier than the intersection of two lines (which is almost surely a point). This informal intuition fails to be expressed by either a Euclidean or a Hausdorff dimension. On the other hand, the formal addition of codimensions suggests that the three intersections in question have the respective dimensions -2, -1, and 0. The inequalities between those values conform with the above-mentioned naive intuition. Therefore, they ushered in the search for a new mathematical definition of dimension that can be measured and for which negative values are legitimate and intuitive. This search produced several publications leading to Mandelbrot (1995). Two notions should be mentioned.

Embedding. A problem that concerns $\mathbf{R}^2$ can often be reinterpreted as a problem that really concerns $\mathbf{R}^E$, with $E > 2$, but must be approached within planar intuitions by $\mathbf{R}^2$. Conversely, if a given problem can be embedded into a problem concerning $\mathbf{R}^E$, the question arises, "which is the 'critical' value of $E - 2$, defined as the smallest value for which the intersection ceases to be empty, and precisely reduces to a point?" In the example of a line and a point, the critical $E - 2$ is precisely 1: once embedded in $\mathbf{R}^3$, the problem transforms into the intersection of a plane and a line, which is a point.

Approximation and pre-asymptotics in mathematics and the sciences. Consider a set defined as the limit of a sequence of decreasing approximations. When the limit is not empty, all the usual dimensions are defined as being properties of the limit, but when the limit is empty and all the dimensions vanish, it is possible to consider instead the limits of the properties of the approximations. The Minkowski–Bouligand formal definition of dimension generalizes to fit the naive intuitive values that may be either positive or negative.

2. Latent Dimensions That Exceed That of the Embedding Space

For a strictly self-similar set in $\mathbf{R}^E$, the Moran equation defines a similarity dimension that obeys $d_{\mathrm{sim}} \leq E$. On the other hand, a generator that is a self-avoiding broken line can easily yield $\log(N)/\log(1/r) = d_{\mathrm{sim}} > E$. Recursive application of this generator defines a parametrized motion, but the union of the positions of the motion is neither a self-similar curve nor any other self-similar set. It is, instead, a set whose points are covered infinitely often. Its box dimension is $\leq E$, which *a fortiori* is $< d_{\mathrm{sim}}$. However, one can load a mass on this set by following the route that applies in the absence of multiple points. Mass is distributed on the generator's intervals in proportion to the values of $r_i^{d_{\mathrm{sim}}}$. By infinite recursion, the difference between the times t' and t'' when points P' and P'' are visited is defined as the mass supported by the portion of the curve that links these points.

If so and $d_{\mathrm{sim}} > E$, the similarity dimension acquires a useful role as a latent dimension. For example, consider the multiplication of dimensions in Section III.D. Suppose that our recursively constructed set is not lighted for all instants of time, but only intermittently when time falls within a fractal dust of dimension d''. Then, the rule of thumb is that the latent dimension of the lighted points is $d_{\mathrm{sim}}d''$. When $d_{\mathrm{sim}}d'' < E$, the rule of thumb is that the true dimension is also $d_{\mathrm{sim}}d''$.

Figure 21 shows an example. The generator has $N = 6$ segments, each with scaling ratio $r = 1/2$, hence latent dimension $d_{\mathrm{sim}} = \log 6/\log 2 > 2$. Taking as subordinator

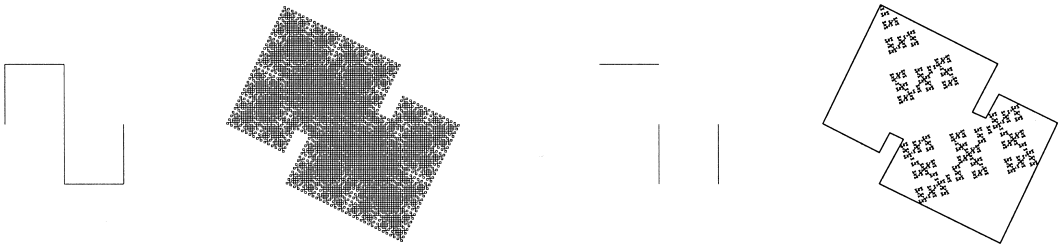


FIGURE 21 Left: Generator and limiting shape with latent dimension exceeding 2. Right: generator and limiting shape of a subordinate with dimension <2. For comparison, this limiting shape is enclosed in the outline of the left limiting shape.

a Cantor set with generator having $N = 3$ segments, each with scaling ratio $r = 1/2$, yields a self-similar fractal with dimension $\log 3/\log 2$.

G. Mapping

Recall f satisfies the Hölder condition with exponent H if there is a positive constant c for which $|f(x) - f(y)| \le c|x - y|^H$. For such functions, $d_{\mathrm{H}}(f(A)) \le (1/H)d_{\mathrm{H}}(A)$. If $H = 1$, f is called a Lipschitz function; f is bi-Lipschitz if there are constants c_1 and c_2 with $c_1|x - y| \le |f(x) - f(y)| \le c_2|x - y|$. Hausdorff dimension is invariant under bi-Lipschitz maps. The analogous properties hold for box-counting dimension.

IV. METHODS OF COMPUTING DIMENSION IN MATHEMATICAL FRACTALS

Upper bounds for the Hausdorff dimension can be relatively straightforward: it suffices to consider a specific family of coverings of the set. Lower bounds are more delicate. We list and describe briefly some methods for computing dimension.

A. Mass Distribution Methods

A *mass distribution* on a set A is a measure μ with $\mathrm{supp}(\mu) \subset A$ and $0 < \mu(A) < \infty$. The *mass distribution principle* (Falconer, 1990, p. 55) establishes a lower bound for the Hausdorff dimension: Let μ be a mass distribution on A and suppose for some s there are constants $c > 0$ and $\delta > 0$ with $\mu(U) \le c \cdot |U|^s$ for all sets U with $|U| \le \delta$. Then $\delta \le d_{\mathrm{H}}(A)$.

Suitable choice of mass distribution can show that no individual set of a cover can cover too much of A. This can eliminate the problems caused by covers by sets of a wide range of diameters.

B. Potential Theory Methods

Given a mass distribution μ, the *s-potential* is defined by Frostman (1935) as

$$\phi_s(x) = \int \frac{d\mu(x)}{|x - y|^s}.$$

If there is a mass distribution μ on a set A with $\int \phi_s(x)\, d\mu(x) < \infty$, then $d_{\mathrm{H}}(A) \ge s$. Potential theory has been useful for computing dimension of many sets, for example, Brownian paths.

C. Implicit Methods

McLaughlin (1987) introduced a geometrical method, based on local approximate self-similarities, which succeeds in proving that $d_{\mathrm{H}}(A) = \overline{d_{\mathrm{box}}}(A)$, without first determining $d_{\mathrm{H}}(A)$. If small parts of A can be mapped to large parts of A without too much distortion, or if A can be mapped to small parts of A without too much distortion, then $d_{\mathrm{H}}(A) = \overline{d_{\mathrm{box}}}(A) = s$ and $\mathcal{H}^s(A) > 0$ (in the former case) or $\mathcal{H}^s(A) < \infty$ (in the latter case). Details and examples can be found in Falconer (1997, Section 3.1).

D. Thermodynamic Formalism

Sinai (1972), Bowen (1975), and Ruelle (1978) adapted methods of statistical mechanics to determine the dimensions of fractals arising from some nonlinear processes. Roughly, for a fractal defined as the attractor A of a family of nonlinear contractions F_i with an inverse function f defined on A, the *topological pressure* $P(\phi)$ of a Lipschitz function $\phi\colon A \to \mathbf{R}$ is

$$P(\phi) = \lim_{k\to\infty} \frac{1}{k} \log \left\{ \sum_{x \in \mathrm{Fix}(f^k)} \exp[\phi(x) + \phi(f(x)) + \cdots \phi(f^{k-1}(x))] \right\},$$

where $\mathrm{Fix}(f^k)$ denotes the set of fixed points of f^k. The sum plays the role of the partition function in statistical mechanics, part of the motivation for the name "thermodynamic formalism." There is a unique s for which

$P(-s\log|f'|)=0$, and $s=d_{\mathrm{H}}(A)$. Under these conditions, $0<\mathcal{H}^s(A)<\infty$, $\mathcal{H}^s(A)$ is a Gibbs measure on A, and many other results can be deduced. Among other places, this method has been applied effectively to the study of Julia sets.

V. METHODS OF MEASURING DIMENSION IN PHYSICAL SYSTEMS

For shapes represented in the plane—for example, coastlines, rivers, mountain profiles, earthquake faultlines, fracture and cracking patterns, viscous fingering, dielectric breakdown, growth of bacteria in stressed environments—box dimension is often relatively easy to compute. Select a sequence $\epsilon_1>\epsilon_2>\cdots>\epsilon_n$ of sizes of boxes to be used to cover the shape, and denote by $N(\epsilon_i)$ the number of boxes of size ϵ_i needed to cover the shape. A plot of $\log(N(\epsilon_i))$ against $\log(1/\epsilon_i)$ often reveals a scaling range over which the points fall close to a straight line. In the presence of other evidence (hierarchical visual complexity, for example), this indicates a fractal structure with box dimension given by the slope of the line. Interpreting the box dimension in terms of underlying physical, chemical, and biological processes has yielded productive insights.

For physical objects in three-dimensional space—for example, aggregates, dustballs, physiological branchings (respiratory, circulatory, and neural), soot particles, protein clusters, terrain maps—it is often easier to compute mass dimension. Select a sequence of radii $r_1>r_2>\cdots>r_n$ and cover the object with concentric spheres of those radii. Denoting by $M(r_i)$ the mass of the part of the object contained inside the sphere of radius r_i, a plot of $\log(M(r_i))$ against $\log(r_i)$ often reveals a scaling range over which the points fall close to a straight line. In the presence of other evidence (hierarchical arrangements of hole sizes, for example), this indicates a fractal structure with mass dimension given by the slope of the line. Mass dimension is relevant for calculating how density scales with size, and this in turn has implications for how the object is coupled to its environment.

VI. LACUNARITY

Examples abound of fractals sharing the same dimension but looking quite different. For instance, both Sierpinski carpets in Fig. 22 have dimension $\log 40/\log 7$. The holes' distribution is more uniform on the left than on the right. The quantification of this difference was undertaken in Mandelbrot (1982, Chapter 34). It introduced *lacunarity* as one expression of this difference, and took another step in characterizing fractals through associated numbers. How can the distribution of a fractal's holes or gaps ("lacunae") be quantified?

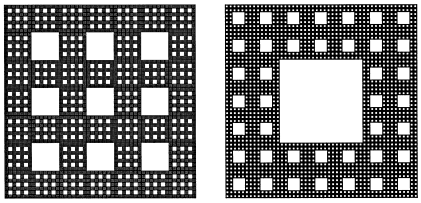


FIGURE 22 Two Sierpinski carpet fractals with the same dimension.

A. The Prefactor

Suppose A is either carpet in Fig. 22, and let A_δ denote the δ-thickening of A. As mentioned in Section II.D, $\mathrm{area}(A_\delta)\sim\Lambda\cdot\delta^{2-\log 40/\log 7}$. One measure of lacunarity is $1/\Lambda$, if the appropriate limit exists.

It is well known that for the box dimension, the limit as $\epsilon\to 0$ can be replaced by the sequential limit $\epsilon_n\to 0$, for ϵ_n satisfying mild conditions. For these carpets, natural choices are those ϵ_n just filling successive generations of holes. Applied to Fig. 22, these ϵ_n give $1/\Lambda\approx 0.707589$ and 0.793487, agreeing with the notion that higher lacunarity corresponds to a more uneven distribution of holes.

Unfortunately, the prefactor is much more sensitive than the exponent: different sequences of ϵ_n give different limits. Logarithmic averages can be used, but this is work in progress.

B. The Crosscuts Structure

An object is often best studied through its *crosscuts* by straight lines, concentric circles, or spheres. For a fractal of dimension d in the plane, the rule of thumb is that the crosscuts are Cantor-like objects of dimension $d-1$. The case when the gaps between points in the crosscut are statistically independent was singled out by Mandelbrot as defining "neutral lacunarity." If the crosscut is also self-similar, it is a Lévy dust.

Hovi *et al.* (1996) studied the intersection of lines (linear crosscuts) with two- and three-dimensional critical percolation clusters, and found the gaps are close to being statistically independent, thus a Lévy dust.

In studying very large DLA clusters, Mandelbrot *et al.* (1995) obtained a crosscut dimension of $d_c=0.65\pm 0.01$, different from the value 0.71 anticipated if DLA clusters were statistically self-similar objects with mass dimension $d_{\mathrm{mass}}=1.71$. The difference can be explained by asserting the number of particles $N_c(r/l)$ on a crosscut

of radius r scales as $N_c(r/l) = \Lambda(r)(r/l)^{d_c}$. Here l is the scaling length, and the lacunarity prefactor varies with r. Assuming slow variation of $\Lambda(r)$ with r, the observed linear log–log fit requires $\Lambda(r) \sim r^{\delta d}$, where $\delta d = d_{\text{mass}} - 1 - d_c = 0.06 \pm 0.01$. Transverse crosscut analysis reveals lacunarity decreases with r for large DLA clusters.

C. Antipodal Correlations

Select an occupied point p well inside a random fractal cluster, so that the $R \times R$ square centered at p lies within the cluster. Now select two vectors V and W based at p and separated by the angle θ. Finally, denote by x and y the number of occupied sites within the wedges with apexes at p, apex angles ϕ much less than θ, and centered about the vectors V and W. The *angular correlation function* is

$$C(\theta) = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\langle x^2 \rangle - \langle x \rangle \langle x \rangle},$$

where $\langle \cdots \rangle$ denotes an average over many realizations of the random fractal. *Antipodal correlations* concern $\theta = \pi$. Negative and positive antipodal correlations are interpreted as indicating high and low lacunarity; vanishing correlation is a weakened form of neutral lacunarity.

Mandelbrot and Stauffer (1994) used antipodal correlations to study the lacunarity of critical percolation clusters. On smaller central subclusters, they found the antipodes are uncorrelated.

Trema random fractals. These are formed by removing randomly centered discs, tremas, with radii obeying a power-law scaling. For them, $C(\pi) \to 0$ with ϕ because a circular hole that overlaps a sector cannot overlap the opposite sector. But nonconvex tremas introduce positive antipodal correlations. For θ close to π, needle-shaped tremas, though still convex, yield $C(\theta)$ much higher than for circular trema sets. From this more refined viewpoint, needle tremas' lacunarity is much lower.

VII. FRACTAL GRAPHS AND SELF-AFFINITY

A. Weierstrass Functions

Smooth functions' graphs, as seen under sufficient magnification, are approximated by their tangents. Unless the function itself is linear, the existence of a tangent contradicts the scale invariance that characterizes fractals. The early example of a continuous, nowhere-differentiable function devised in 1834 by Bolzano remained unpublished until the 1920s. The first example to become widely known was constructed by Weierstrass in 1872. The *Weierstrass sine function* is

$$W(t) = \sum_{n=0}^{\infty} b^{-Hn} \sin(2\pi b^n t),$$

and the complex *Weierstrass function* is

$$W_0(t) = \sum_{n=0}^{\infty} b^{-Hn} \exp(2\pi i b^n t).$$

Hardy (1916) showed $W(t)$ is continuous and nowhere-differentiable if and only if $b > 1$ and $0 < H < 1$.

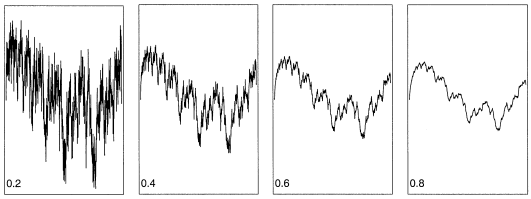


FIGURE 23 The effect of H on Weierstrass graph roughness. In all pictures, $b = 1.5$ and H has the indicated value.

As shown in Fig. 23, the parameter H determines the roughness of the graph. In this case, H is not a perspicuous "roughness exponent." Indeed, as b increases, the amplitudes of the higher frequency terms decrease and the graph is more clearly dominated by the lowest frequency terms. This effect of b is a little-explored aspect of lacunarity.

B. Weierstrass–Mandelbrot Functions

The Weierstrass function revolutionized mathematics but did not enter physics until it was modified in a series of steps described in Mandelbrot (1982, pp. 388–390; (2001d, Chapter H4). The step from $W_0(t)$ to $W_1(t)$ added low frequencies in order to insure self-affinity. The step from $W_1(t)$ to $W_2(t)$ added to each addend a random phase φ_n uniformly distributed on [0, 1]. The step from $W_1(t)$ to $W_3(t)$ added a random amplitude $A_n = \sqrt{-2 \log V}$, where V is uniform on [0, 1]. A function $W_4(t)$ that need not be written down combines a phase and an amplitude. The latest step leads to another function that need not be written down: it is $W_5(t) = W_4(t) + W_4(-t)$, where the two addends are statistically independent. Contrary to all earlier extensions, $W_5(t)$ is not chiral. We have

$$W_1(t) = \sum_{n=-\infty}^{\infty} b^{-Hn}(\exp(2\pi i b^n t) - 1),$$

$$W_2(t) = \sum_{n=-\infty}^{\infty} b^{-Hn}(\exp(2\pi i b^n t) - 1)\exp(i\varphi_n),$$

$$W_3(t) = \sum_{n=-\infty}^{\infty} A_n b^{-Hn}(\exp(2\pi i b^n t) - 1).$$

C. The Hölder Exponent

A function $f: [a, b] \to \mathbf{R}$ has *Hölder exponent* H if there is a constant $c > 0$ for which

$$|f(x) - f(y)| \le c \cdot |x - y|^H$$

for all x and y in $[a, b]$ (recall Section III.G). If f is continuous and has Hölder exponent H satisfying $0 < H \le 1$, then the graph of f has box dimension $d_{\text{box}} \le 2 - H$.

The Weierstrass function $W(t)$ has Hölder exponent H, hence its graph has $d_{\text{box}} \le 2 - H$. For large enough b, $d_{\text{box}} = 2 - H$, so one can think of the Hölder exponent as a measure of roughness of the graph.

VIII. FRACTAL ATTRACTORS AND REPELLERS OF DYNAMICAL SYSTEMS

The modern renaissance in dynamical systems is associated most often with chaos theory. Consequently, the relations between fractal geometry and chaotic dynamics, mediated by symbolic dynamics, are relevant to our discussion. In addition, we consider fractal basin boundaries, which generalize Julia sets to much wider contexts including mechanical systems.

A. The Smale Horseshoe

If they exist, intersections of the stable and unstable manifolds of a fixed point are called *homoclinic points*. Poincaré (1890) recognized that homoclinic points cause great complications in dynamics. Yet much can be understood by labeling an appropriate coarse-graining of a neighborhood of a homoclinic point and translating the corresponding dynamics into a string of symbols (the coarse-grain bin labels). The notion of symbolic dynamics first appears in Hadamard (1898), and Birkhoff (1927) proved every neighborhood of a homoclinic point contains infinitely many periodic points.

Motivated by work of Cartwright and Littlewood (1945) and Levinson (1949) on the forced van der Pol oscillator, Smale (1963) constructed the *horseshoe map*. This is a map from the unit square into the plane with completely invariant set a Cantor set Λ, roughly the Cartesian product of two Cantor middle-thirds sets. Restricted to Λ, with the obvious symbolic dynamics encoding, the horseshoe map is conjugate to the shift map on two symbols, the archetype of a chaotic map.

This construction is universal in the sense that it occurs in every transverse homoclinic point to a hyperbolic saddle point. The Conley–Moser theorem (see Wiggins, 1990) establishes the existence of chaos by conjugating the dynamics to a shift map on a Cantor set under general conditions. In this sense, chaos often equivalent to simple dynamics on an underlying fractal.

B. Fractal Basin Boundaries

For any point c belonging to a hyperbolic component of the Mandelbrot set, the Julia set is the boundary of the basins of attraction of the attracting cycle and the attracting fixed point at infinity. See the right side of Fig. 3.

Another example favored by Julia is found in Newton's method for finding the roots of a polynomial $f(z)$ of degree at least 3. It leads to the dynamical system $z_{n+1} = N_f(z_n) = z_n - f(z_n)/f'(z_n)$. The roots of $f(z)$ are attracting fixed points of $N_f(z)$, and the boundary of the basins of attraction of these fixed points is a fractal; an example is shown on the left side of Fig. 24. If contaminated by even small uncertainties, the fate of initial points near the basin boundary cannot be predicted. Sensitive dependence on initial conditions is a signature of chaos, but here we deal with something different. The eventual behavior is completely predictable, except for initial points taken exactly on the basin boundary, usually of two-dimensional Lebesgue measure 0.

The same complication enters mechanical engineering problems for systems with multiple attractors. Moon (1984) exhibited an early example. Extensive theoretical and computer studies by Yorke and coworkers are described in Alligood and Yorke (1992). The driven harmonic oscillator with two-well potential

$$\frac{d^2x}{dt^2} + f\frac{dx}{dt} - \frac{1}{2}x(1 - x^2) = A\cos(\omega t)$$

is a simple example. The undriven system has two equilibria, $x = -1$ and $x = +1$. Initial values (x, x') are painted white if the trajectory from that point eventually stays in the left basin, black if it eventually stays in the right basin. The right side of Fig. 24 shows the initial condition portrait for the system with $f = 0.15$, $\omega = 0.8$, and $A = 0.094$.

IX. FRACTALS AND DIFFERENTIAL OR PARTIAL DIFFERENTIAL EQUATIONS

The daunting task to which a large portion of Mandelbrot (1982) is devoted was to establish that many works of nature and man [as shown in Mandelbrot (1997), the latter includes the stock market!] are fractal. New and often important examples keep being discovered, but the hardest present challenge is to discover the *causes* of fractality. Some cases remain obscure, but others are reasonably clear.

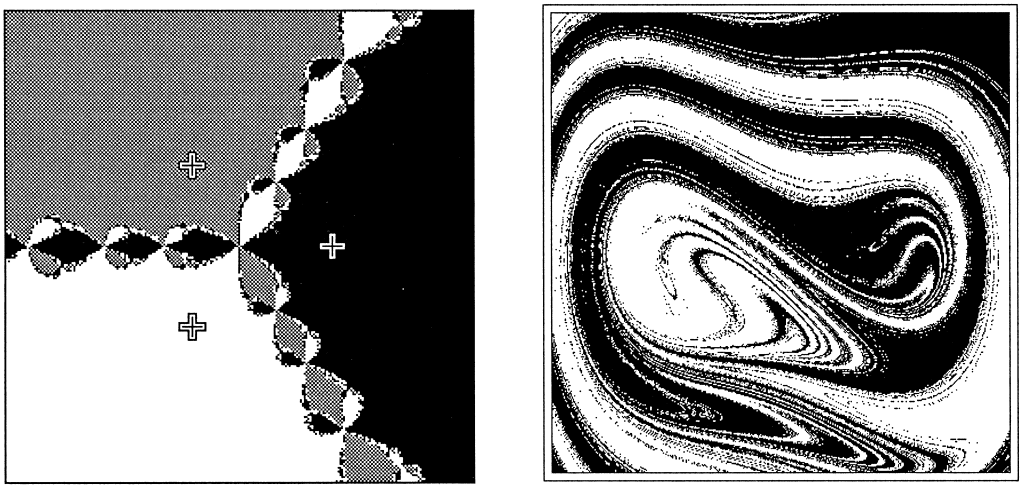


FIGURE 24 Left: The basins of attraction of Newton's method for finding the roots of $z^3 - 1$. Right: The basins of attraction for a damped, driven two-well harmonic oscillator.

Thus, the fractality of the physical percolation clusters (Section I.B.3.e) is the geometric counterpart of scaling and renormalization: the analytic properties of those objects follow a wealth of power-law relations. Many mathematical issues, some of them already mentioned, remain open, but the overall renormalization framework is firmly rooted. Renormalization and the resulting fractality also occur in the structure of attractors and repellers of dynamical systems. Best understood is renormalization for quadratic maps. Feigenbaum and others considered the real case. For the complex case, renormalization establishes that the Mandelbrot set contains infinitely many small copies of itself.

Unfortunately, additional examples of fractality proved to be beyond the scope of the usual renormalization. A notorious case concerns DLA (Section I.B.3.f).

A. Fractal Attractors of Ordinary Differential Equations

The Lorenz equations for fluid convection in a two-dimensional layer heated from below are

$$\frac{dx}{dt} = \sigma(y-x), \quad \frac{dy}{dt} = -xz+rx-y, \quad \frac{dz}{dt} = xy-bz.$$

Here x denotes the rate of convective overturning, y the horizontal temperature difference, and z the departure from a linear vertical temperature gradient. For the parameters $\sigma = 10$, $b = 8/3$, and $r = 28$, Lorenz (1963) suggested that trajectories in a bounded region converge to an attractor that is a fractal, with dimension about 2.06, as estimated by Liapunov exponents. The Lorenz equations are very suggestive but do not represent weather systems very well. However, Haken established a connection with lasers. The sensitivity to initial conditions common to chaotic dynamics is mediated by the intricate fractal interleaving of the multiple layers of the attractor. In addition, Birman and Williams (1983) showed an abundance of knotted periodic orbits embedded in the Lorenz attractor, though Williams (1983) showed all such knots are prime. Grist (1997) constructed a *universal template*, a branched 2-manifold in which all knots are embedded. Note the interesting parallel with the universal aspects of the Sierpinski carpet (Section I.B.1.a). It is not yet known if the attractor of any differential equation contains a universal template. The Poincaré–Bendixson theorem prohibits fractal attractors for differential equations in the plane, but many other classical ordinary differential equations in at least three dimensions exhibit similar fractal attractors in certain parameter ranges.

B. Partial Differential Equations on Domains with Fractal Boundaries ("Can One Hear the Shape of a Fractal Drum?")

Suppose $D \subset \mathbf{R}^n$ is an open region with boundary ∂D. Further, suppose the eigenvalue problem $\nabla^2 u = -\lambda u$ with boundary conditions $u(x) = 0$ for all $x \in \partial D$ has real eigenvalues $0 < \lambda_1 < \lambda_2 < \cdots$. For D with sufficiently smooth boundary, a theorem of Weyl (1912) shows $N(\lambda) \sim \lambda^{n/2}$, where the *eigenvalue counting function* $N(\lambda) = \{\text{the number of } \lambda_i \text{ for which } \lambda_i \leq \lambda\}$. If the boundary ∂D is a fractal, Berry (1979, pp. 51–53) postulated that some form of the dimension of ∂D appears in the second term in the expansion of $N(\lambda)$, therefore can be recovered from the eigenvalues. This *could not* be the Hausdorff dimension, but Lapidus (1995) showed that it

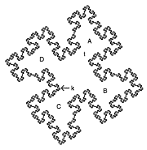


FIGURE 25 Perimeter of an extensively studied fractal drum.

is the Minkowski–Bouligard dimension. The analysis is subtle, involving some deep number theory.

For regions with fractal boundaries, the heat equation $\nabla^2 u = (\partial/\partial t)u$ shows heat flow across a fractal boundary is related to the dimension of the boundary. Sapoval (1989) and Sapoval *et al.* (1991) conducted elegant experimenths to study the modes of fractal drums. The perimeter of Fig. 25 has dimension $\log 8/\log 4 = 3/2$. A membrane stretched across this fractal curve was excited acoustically and the resulting modes observed by sprinkling powder on the membrane and shining laser light transverse to the surface. Sapoval observed modes localized to bounded regions A, B, C, and D shown in Fig. 25. By carefully displacing the acoustic source, he was able to excite each separately.

Theoretical and computer-graphic analyses of the wave equation on domains with fractal boundaries have been carried out by Lapidus *et al.* (1996), among others.

C. Partial Differential Equations on Fractals

The problem is complicated by the fact that a fractal is not a smooth manifold. How is the Laplacian to be defined on such a space? One promising approach was put forward by physicists in the 1980s and made rigorous in Kigami (1989): approximate the fractal domain by a sequence of graphs representing successive protofractals, and define the fractal Laplacian as the limit of a suitably renormalized sequence of Laplacians on the graphs. Figure 26 shows the first four graphs for the equilateral Sierpinski gasket. The values at the boundary vertices are specified by the boundary conditions at any nonboundary vertex x_0. The mth approximate Laplacian of a function $f(x)$ is the product of a renormalization factor by $\sum (f(y) - f(x_0))$, where the sum is taken over all vertices y in the mth protofractal graph corresponding to the mth-stage reduction of the whole graph.

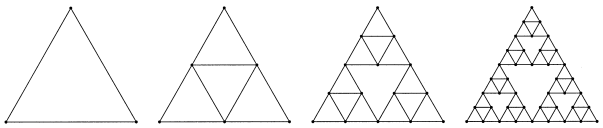


FIGURE 26 Graphs corresponding to protofractal approximations of the equilateral Sierpinski gasket.

With this Laplacian, the heat and wave equations can be defined on fractals. Among other things, the wave equation on domains with fractal boundaries admits localized solutions, as we saw for the wave equation on fractal drums. A major challenge is to extend these ideas to fractals more complicated than the Sierpinski gasket and its relatvives.

D. How Partial Differential Equations Generate Fractals

A quandary: It is universally granted that physics is ruled by diverse partial differential equations, such as those of Laplace, Poisson, and Navier–Stokes. A differential equation necessarily implies a great degree of local smoothness, even though close examination shows isolated singularities or "catastrophes." To the contrary, fractality implies everywhere-dense roughness or fragmentation. This is one of the several reasons that fractal models in diverse fields were initially perceived as being "anomalies" contradicting one of the firmest foundations of science.

A conjecture–challenge responding to the preceding quandry. There is no contradiction at all: fractals arise unavoidably in the long-time behavior of the solution of very familiar and innocuous-looking equations. In particular, many concrete situations where fractals are observed involve equations that allow free and moving boundaries, interfaces, or singularities. As a suggestive "principle," Mandelbrot (1982, Chapter 11) described the following possibility: under broad conditions that largely remain to be specified, these free boundaries, interfaces, and singularities converge to suitable fractals. Many equations have been examined from this viewpoint, but we limit ourselves to two examples of central importance.

1. The Large-Scale Distribution of Galaxies

Chapters 9 and 33–35 of Mandelbrot (1982) conjecture that the distribution of galaxies is fractal. This conjecture results from a search for invariants that was central to every aspect of the construction of fractal geometry. Granted that the distribution of galaxies certainly deviates from homogeneity, one broad approach consists in correcting for local inhomogeneity by using local "patches." The next simplest global assumption is that the distribution is nonhomogeneous but scale invariant, therefore fractal.

Excluding the strict hierarchies, two concrete constructions of random fractal sets were subjected to detailed mathematical and visual investigation. These constructions being random, self-similarity can only be statistical. But a strong counteracting asset is that the self-similarity ratio can be chosen freely, is not restricted to powers of a prescribed r_0. A surprising and noteworthy finding came forth. These constructions exhibited a strong hierarchical

structure that is not a deliberate and largely arbitrary input. Details are given in Mandelbrot (1982).

The first construction is the *seeded universe* based on a Lévy flight. Its Hausdorff-dimensional properties were well known. Its correlation properties (Mandelbrot 1975) proved to be nearly identical to those of actual galaxy maps. The second construction is the *parted universe* obtained by subtracting from space a random collection of overlapping tremas. Either construction yields sets that are highly irregular and involve no special center, yet, with no deliberate design, exhibit a clear-cut clustering, "filaments" and "walls." These structures were little known when these constructions were designed.

Conjecture: Could it be that the observed "clusters," "filaments," and "walls" need not be explained separately? They may not result from unidentified features of specific models, but represent unavoidable consequences of a variety of unconstrained forms of random fractality, as interpreted by a human brain.

A problem arose when careful examination of the simulations revealed a clearly incorrect prediction. The simulations in the *seeded universe* proved to be visually far more "lacunar" than the real world. That is, the simulations' holes are larger than in reality. The *parted universe* model fared better, since its lacunarity can be adjusted at will and fit to the actual distribution. A lowered lacunarity is expressed by a positive correlation between masses in antipodal directions. Testing this specific conjecture is a challenge for those who analyze the data.

Does dynamics make us expect the distribution of galaxies to be fractal? Position a large array of point masses in a cubic box in which opposite sides are identified to form a three-dimensional torus. The evolution of this array obeys the Laplace equation, with the novelty that the singularities of the solution are the positions of the points, therefore movable. All simulations we know (starting with those performed at IBM around 1960) suggest that, even when the pattern of the singularities begins by being uniform or Poisson, it gradually creates clusters and a semblance of hierarchy, and appears to tend toward fractality. It is against the preceding background that the limit distribution of galaxies is conjectured to be fractal, and fractality is viewed as compatible with Newton's equations.

2. The Navier–Stokes Equation

The first concrete use of a Cantor dust in real spaces is found in Berger and Mandelbrot (1963), a paper on noise records. This was nearly simultaneous with Kolmogorov's work on the intermittence of turbulence. After numerous experimental tests designed to create an intuitive feeling for this phenomenon (e.g., listening to turbulent velocity records that were made audible), the fractal viewpoint was extended to turbulence, and circa 1964 led to the following conjecture.

Conjecture. The property of being "turbulently dissipative" should not be viewed as attached to domains in a fluid with significant interior points, but as attached to fractal sets. In a first approximation, those sets' intersection with a straight line is a Cantor-like fractal dust having a dimension in the range from 0.5 to 0.6. The corresponding full sets in space should therefore be expected to be fractals with Hausdorff dimension in the range from 2.5 to 2.6.

Actually, Cantor dust and Hausdorff dimension are not the proper notions in the context of viscous fluids because viscosity necessarily erases the fine detail essential to fractals. Hence the following conjecture (Mandelbrot, 1982, Chapter 11; 1976). The dissipation in a viscous fluid occurs in the neighborhood of a singularity of a nonviscous approximation following Euler's equations, and the motion of a nonviscous fluid acquires singularities that are sets of dimension about 2.5–2.6. Several numerical tests agree with this conjecture (e.g., Chorin, 1981).

A related conjecture, that the Navier–Stokes equations have fractal singularities of much smaller dimension, has led to extensive work by V. Scheffer, R. Teman, and C. Foias, and many others. But this topic is not exhausted.

Finally, we mention that fractals in phase space entered the transition from laminar to turbulent flow through the work of Ruelle and Takens (1971) and their followers. The task of unifying the real- and phase-space roles of fractals is challenging and far from being completed.

X. FRACTALS IN THE ARTS AND IN TEACHING

The Greeks asserted art reflects nature, so it is little surprise that the many fractal aspects of nature should find their way into the arts—beyond the fact that a representational painting of a tree exhibits the same fractal branching as a physical tree. Voss and Clarke (1975) found fractal power-law scaling in music, and self-similarity is designed in the music of the composers György Ligeti and Charles Wuorinen. Pollard-Gott (1986) established the presence of fractal repetition patterns in the poetry of Wallace Stevens. Computer artists use fractals to create both abstract aesthetic images and realistic landscapes. Larry Poons' paintings since the 1980s have had rich fractal textures. The "decalcomania" of the 1830s and the 1930s and 1940s used viscous fingering to provide a level of visual complexity. Before that, Giacometti's Alpine wildflower paintings are unquestionably fractal. Earlier still, relatives of the Sierpinski gasket occur as decorative motifs in Islamic and Renaissance art. Fractals abound in architecture, for example, in the cascades of spires in Indian temples, Bramante's

plan for St. Peter's, Malevich's Architektonics, and some of Frank Lloyd Wright's designs. Fractals occur in the writing of Clarke, Crichton, Hoag, Powers, Updike, and Wilhelm, among others, and in at least one play, Stoppard's *Arcadia*. Postmodern literary theory has used some concepts informed by fractal geometry, though this application has been criticized for its overly free interpretations of precise scientific language. Some have seen evidence of power-law scaling in historical records, the distribution of the magnitudes of wars and of natural disasters, for example. In popular culture, fractals have appeared on t-shirts, totebags, book covers, MTV logos, been mentioned on public radio's *A Prairie Home Companion*, and been seen on television programs from *Nova* and *Murphy Brown*, through several incarnations of *Star Trek*, to *The X-Files* and *The Simpsons*. While Barnsley's (1988) slogan, "fractals everywhere," is too strong, the degree to which fractals surround us outside of science and engineering is striking.

A corollary of this last point is a good conclusion to this high-speed survey. In our increasingly technological world, science education is very important. Yet all too often humanities students are presented with limited choices: the first course in a standard introductory sequence, or a survey course diluted to the level of journalism. The former builds toward major points not revealed until later courses, the latter discusses results from science without showing how science is done. In addition, many efforts to incorporate computer-aided instruction attempt to replace parts of standard lectures rather than engage students in exploration and discovery.

Basic fractal geometry courses for non-science students provide a radical departure from this mode. The subject of fractal geometry operates at human scale. Though new to most, the notion of self-similarity is easy to grasp, and (once understood) handles familiar objects from a genuinely novel perspective. Students can explore fractals with the aid of readily available software. These instances of computer-aided instruction are perfectly natural because computers are so central to the entire field of fractal geometry. The contemporary nature of the field is revealed by a supply of mathematical problems that are simple to state but remain unsolved. Altogether, many fields of interest to non-science students have surprising examples of fractal structures. Fractal geometry is a powerful tool for imparting to non-science students some of the excitement for science often invisible to them. Several views of this are presented in Frame and Mandelbrot (2001).

The importance of fractals in the practice of science and engineering is undeniable. But fractals are also a proven force in science education. Certainly, the boundaries of fractal geometry have not yet been reached.

SEE ALSO THE FOLLOWING ARTICLES

CHAOS • PERCOLATION • TECTONOPHYSICS

BIBLIOGRAPHY

Alligood, K., and Yorke, J. (1992). *Ergodic Theory Dynam. Syst.* **12,** 377–400.

Alligood, K., Sauer, T., and Yorke, J. (1997). "Chaos. An Introduction to Dynamical Systems," Springer-Verlag, New York.

Barnsley, M. (1988). "Fractals Everywhere," 2nd ed., Academic Press, Orlando, FL.

Barnsley, M., and Demko, S. (1986). "Chaotic Dynamics and Fractals," Academic Press, Orlando, FL.

Barnsley, M., and Hurd, L. (1993). "Fractal Image Compression," Peters, Wellesley, MA.

Batty, M., and Longley, P. (1994). "Fractal Cities," Academic Press, London.

Beardon, A. (1983). "The Geometry of Discrete Groups," Springer-Verlag, New York.

Beck, C., and Schlögl, F. (1993). "Thermodynamics of Chaotic Systems: An Introduction," Cambridge University Press, Cambridge.

Berger, J., and Mandelbrot, B. (1963). *IBM J. Res. Dev.* **7,** 224–236.

Berry, M. (1979). "Structural Stability in Physics," Springer-Verlag, New York.

Bertoin, J. (1996). "Lévy Processes," Cambridge University Press, Cambridge.

Besicovitch, A., and Moran, P. (1945). *J. Lond. Math. Soc.* **20,** 110–120.

Birkhoff, G. (1927). "Dynamical Systems, American Mathematical Society, Providence, RI.

Birman, J., and Williams, R. (1983). *Topology* **22,** 47–82.

Bishop, C., and Jones, P. (1997). *Acta Math.* **179,** 1–39.

Blanchard, P. (1994). *In* "Complex Dynamical Systems. The Mathematics behind the Mandelbrot and Julia Sets" (Devaney, R., ed.), (pp. 139–154, American Mathematical Society, Providence, RI.

Bochner, S. (1955). "Harmonic Analysis and the Theory of Probability," University of California Press, Berkeley, CA.

Bowen, R. (1975). "Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms," Springer-Verlag, New York.

Bunde, A., and Havlin, S. (1991). "Fractals and Disordered Systems," Springer-Verlag, New York.

Cartwright, M., and Littlewood, L. (1945). *J. Lond. Math. Soc.* **20,** 180–189.

Cherbit, G. (1987). "Fractals. Non-integral Dimensions and Applications," Wiley, Chichester, UK.

Chorin, J. (1981). *Commun. Pure Appl. Math.* **34,** 853–866.

Crilly, A., Earnshaw, R., and Jones, H. (1991). "Fractals and Chaos," Springer-Verlag, New York.

Crilly, A., Earnshaw, R., and Jones, H. (1993). "Applications of Fractals and Chaos," Springer-Verlag, New York.

Curry, J., Garnett, L., and Sullivan, D. (1983). *Commun. Math. Phys.* **91,** 267–277.

Dekking, F. M. (1982). *Adv. Math.* **44,** 78–104.

Devaney, R. (1989). "An Introduction to Chaotic Dynamical Systems," 2nd ed., Addison-Wesley, Reading, MA.

Devaney, R. (1990). "Chaos, Fractals, and Dynamics. Computer Experiments in Mathematics," Addison-Wesley, Reading, MA.

Devaney, R. (1992). "A First Course in Chaotic Dynamical Systems. Theory and Experiment," Addison-Wesley, Reading, MA.

Devaney, R. (ed.). (1994). "Complex Dynamical Systems. The Mathematics Behind the Mandelbrot and Julia Sets," American Mathematical Society, Providence, RI.

Devaney, R., and Keen, L. (1989). "Chaos and Fractals. The Mathematics Behind the Computer Graphics," American Mathematical Society, Providence, RI.
Douady, A., and Hubbard, J. (1984). "Étude dynamique des polynômes complexes. I, II," Publications Mathematiques d'Orsay, Orsay, France.
Douady, A., and Hubbard, J. (1985). *Ann. Sci. Ecole Norm. Sup*. **18,** 287–343.
Edgar, G. (1990). "Measure, Topology, and Fractal Geometry," Springer-Verlag, New York.
Edgar, G. (1993). "Classics on Fractals," Addison-Wesley, Reading, MA.
Edgar, G. (1998). "Integral, Probability, and Fractal Measures," Springer-Verlag, New York.
Eglash, R. (1999). "African Fractals. Modern Computing and Indigenous Design," Rutgers University Press, New Brunswick, NJ.
Encarncação, J., Peitgen, H.-O., Sakas, G., and Englert, G. (1992). "Fractal Geometry and Computer Graphics," Springer-Verlag, New York.
Epstein, D. (1986). "Low-Dimensional Topology and Kleinian Groups," Cambridge University Press, Cambridge.
Evertsz, C., Peitgen, H.-O., and Voss, R. (eds.). (1996). "Fractal Geometry and Analysis. The Mandelbrot Festschrift, Curaçao 1995," World Scientific, Singapore.
Falconer, K. (1985). "The Geometry of Fractal Sets," Cambridge University Press, Cambridge.
Falconer, K. (1987). *Math. Intelligencer* **9,** 24–27.
Falconer, K. (1990). "Fractal Geometry. Mathematical Foundations and Applications," Wiley, Chichester, UK.
Falconer, K. (1997). "Techniques in Fractal Geometry," Wiley, Chichester, UK.
Family, F., and Vicsek, T. (1991). "Dynamics of Fractal Surfaces," World Scientific, Singapore.
Feder, J. (1988). "Fractals," Plenum Press, New York.
Feder, J., and Aharony, A. (1990). "Fractals in Physics. Essays in Honor of B. B. Mandelbrot," North-Holland, Amsterdam.
Fisher, Y. (1995). "Fractal Image Compression. Theory and Application," Springer-Verlag, New York.
Flake, G. (1998). "The Computational Beauty of Nature. Computer Explorations of Fractals, Chaos, Complex Systems, and Adaptation," MIT Press, Cambridge, MA.
Fleischmann, M., Tildesley, D., and Ball, R. (1990). "Fractals in the Natural Sciences," Princeton University Press, Princeton, NJ.
Frame, M., and Mandelbrot, B. (2001). "Fractals, Graphics, and Mathematics Education," Mathematical Association of America, Washington, DC.
Frostman, O. (1935). *Meddel. Lunds. Univ. Math. Sem*. **3,** 1–118.
Gazalé, M., (1990). "Gnomon. From Pharaohs to Fractals," Princeton University Press, Princeton, NJ.
Grist, R. (1997). *Topology* **36,** 423–448.
Gulick, D. (1992). "Encounters with Chaos," McGraw-Hill, New York.
Hadamard, J. (1898). *J. Mathematiques* **5,** 27–73.
Hardy, G. (1916). *Trans. Am. Math. Soc.* **17,** 322–323.
Hastings, H., and Sugihara, G. (1993). "Fractals. A User's Guide for the Natural Sciences," Oxford University Press, Oxford.
Hovi, J.-P., Aharony, A., Stauffer, D., and Mandelbrot, B. B. (1996). *Phys. Rev. Lett.* **77,** 877–880.
Hutchinson, J. E. (1981). *Ind. Univ. J. Math.* **30,** 713–747.
Keen, L. (1994). *In* "Complex Dynamical Systems. The Mathematics behind the Mandelbrot and Julia Sets" (Devaney, R., ed.), pp. 139–154, American Mathematical Society, Providence, RI.
Keen, L., and Series, C. (1993). *Topology* **32,** 719–749.
Keen, L., Maskit, B., and Series, C. (1993). *J. Reine Angew. Math.* **436,** 209–219.
Kigami, J. (1989). *Jpn. J. Appl. Math.* **8,** 259–290.
Lapidus, M. (1995). *Fractals* **3,** 725–736.
Lapidus, M., Neuberger, J., Renka, R., and Griffith, C. (1996). *Int. J. Bifurcation Chaos* **6,** 1185–1210.
Lasota, A., and Mackey, M. (1994). "Chaos, Fractals, and Noise. Stochastic Aspects of Dynamics," 2nd ed., Springer-Verlag, New York.
Lawler, G., Schramm, O., and Warner, W. (2000). *Acta Math.*, to appear [xxx.lanl.gov./abs/math. PR/0010165].
Lei, T. (2000). "The Mandelbrot Set, Theme and Variations," Cambridge University Press, Cambridge.
Le Méhauté, A. (1990). "Fractal Geometries. Theory and Applications," CRC Press, Boca Raton, FL.
Levinson, N. (1949). *Ann. Math.* **50,** 127–153.
Lorenz, E. (1963). *J. Atmos. Sci.* **20,** 130–141.
Lu, N. (1997). "Fractal Imaging," Academic Press, San Diego.
Lyubich, M. (2001). *Ann. Math.,* to appear.
Mandelbrot, B. (1975). *C. R. Acad. Sci. Paris* **280A,** 1075–1078.
Mandelbrot, B. (1975, 1984, 1989, 1995). "Les objects fractals," Flammarion, Paris.
Mandelbrot, B. (1976). *C. R. Acad. Sci. Paris* **282A,** 119–120.
Mandelbrot, B. (1980). *Ann. N. Y. Acad. Sci.* **357,** 249–259.
Mandelbrot, B. (1982). "The Fractal Geometry of Nature," Freeman, New York.
Mandelbrot, B. (1984). *J. Stat. Phys.* **34,** 895–930.
Mandelbrot, B. (1985). *In* "Chaos, Fractals, and Dynamics" (Fischer, P., and Smith, W., eds.), pp. 235–238, Marcel Dekker, New York.
Mandelbrot, B. (1995). *J. Fourier Anal. Appl.* **1995,** 409–432.
Mandelbrot, B. (1997a). "Fractals and Scaling in Finance. Discontinuity, Concentration, Risk," Springer-Verlag, New York.
Mandelbrot, B. (1997b). "Fractales, Hasard et Finance," Flammarion, Paris.
Mandelbrot, B. (1999). "Multifractals and $1/f$ Noise. Wild Self-Affinity in Physics," Springer-Verlag, New York.
Mandelbrot, B. (2001a). *Quant. Finance* **1,** 113–123, 124–130.
Mandelbrot, B. (2001b). "Gaussian Self-Affinity and Fractals: Globality, the Earth, $1/f$ Noise, & R/S," Springer-Verlag, New York.
Mandelbrot, B. (2001c). "Fractals and Chaos and Statistical Physics," Springer-Verlag, New York.
Mandelbrot, B. (2001d). "Fractals Tools," Springer-Verlag, New York.
Mandelbrot, B. B., and Stauffer, D. (1994). *J. Phys. A* **27,** L237–L242.
Mandelbrot, B. B., Vespignani, A., and Kaufman, H. (1995). *Europhy. Lett.* **32,** 199–204.
Maksit, B. (1988). "Kleinian Groups," Springer-Verlag, New York.
Massopust, P. (1994). "Fractal Functions, Fractal Surfaces, and Wavelets," Academic Press, San Diego, CA.
Mattila, P. (1995). "Geometry of Sets and Measures in Euclidean Space. Fractals and Rectifiability," Cambridge University Press, Cambridge.
McCauley, J. (1993). "Chaos, Dynamics and Fractals. An Algorithmic Approach to Deterministic Chaos," Cambridge University Press, Cambridge.
McLaughlin, J. (1987). *Proc. Am. Math. Soc.* **100,** 183–186.
McMullen, C. (1994). "Complex Dynamics and Renormalization," Princeton University Press, Princeton, NJ.
McShane, G., Parker, J., and Redfern, I. (1994). *Exp. Math.* **3,** 153–170.
Meakin, P. (1996). "Fractals, Scaling and Growth Far from Equilibrium," Cambridge University Press, Cambridge.
Milnor, J. (1989). *In* "Computers in Geometry and Topology" (Tangora, M., ed.), pp. 211–257, Marcel Dekker, New York.
Moon, F. (1984). *Phys. Rev. Lett.* **53,** 962–964.
Moon, F. (1992). "Chaotic and Fractal Dynamics. An Introduction for Applied Scientists and Engineers," Wiley-Interscience, New York.
Parker, J. (1995). *Topology* **34,** 489–496.
Peak, D., and Frame, M. (1994). "Chaos Under Control. The Art and Science of Complexity," Freeman, New York.
Peitgen, H.-O. (1989). "Newton's Method and Dynamical Systems," Kluwer, Dordrecht.

Peitgen, H.-O., and Richter, P. H. (1986). "The Beauty of Fractals," Springer-Verlag, New York.
Peitgen, H.-O., and Saupe, D. (1988). "The Science of Fractal Images," Plenum Press, New York.
Peitgen, H.-O., Jürgens, H., and Saupe, D. (1992). "Chaos and Fractals: New Frontiers of Science," Springer-Verlag, New York.
Peitgen, H.-O., Rodenhausen, A., and Skordev, G. (1998). *Fractals* **6,** 371–394.
Pietronero, L. (1989). "Fractals Physical Origins and Properties," North-Holland, Amsterdam.
Pietronero, L., and Tosatti, E. (1986). "Fractals in Physics," North-Holland, Amsterdam.
Poincaré, H. (1890). *Acta Math.* **13,** 1–271.
Pollard-Gott, L. (1986). *Language Style* **18,** 233–249.
Rogers, C. (1970). "Hausdorff Measures," Cambridge University Press, Cambridge.
Ruelle, D. (1978). "Thermodynamic Formalism: The Mathematical Structures of Classical Equilibrium Statistical Mechanics," Addison-Wesley, Reading, MA.
Ruelle, D., and Takens, F. (1971). *Commun. Math. Phys.* **20,** 167–192.
Samorodnitsky, G., and Taqqu, M. (1994). "Stable Non-Gaussian Random Processes. Stochastic Models with Infinite Variance," Chapman and Hall, New York.
Sapoval, B. (1989). *Physica D* **38,** 296–298.
Sapoval, B., Rosso, M., and Gouyet, J. (1985). *J. Phys. Lett.* **46,** L149–L156.
Sapoval, B., Gobron, T., and Margolina, A. (1991). *Phys. Rev. Lett.* **67**, 2974–2977.
Scholz, C., and Mandelbrot, B. (1989). "Fractals in Geophysics," Birkhäuser, Basel.
Shishikura, M. M. (1994). *Astérisque* **222,** 389–406.
Shlesinger, M., Zaslavsky, G., and Frisch, U. (1995). "Lévy Flights and Related Topics in Physics," Springer-Verlag, New York.
Sinai, Y. (1972). *Russ. Math. Surv.* **27,** 21–70.
Smale, S. (1963). *In* "Differential and Combinatorial Topology" (Cairns, S., ed.), pp. 63–80, Princeton University Press, Princeton, NJ.
Stauffer, D., and Aharony, A. (1992). "Introduction to Percolation Theory," 2nd ed., Taylor and Francis, London.
Strogatz, S. (1994). "Nonlinear Dynamics and Chaos, with Applications to Chemistry, Physics, Biology, Chemistry, and Engineering," Addison-Wesley, Reading, MA.
Sullivan, D. (1985). *Ann. Math.* **122,** 410–418.
Tan, Lei (1984). *In* "Étude dynamique des polynômes complexes" (Douardy, A., and Hubbard, J., eds.), Vol. II, pp. 139–152, Publications Mathematiques d'Orsay, Orsay, France.
Tan, Lei. (2000). "The Mandelbrot Set, Theme and Variations," Cambridge University Press, Cambridge.
Thurston, W. (1997). "Three-Dimensional Geometry and Topology," Princeton University Press, Princeton, NJ.
Tricot, C. (1982). *Math. Proc. Camb. Philos. Soc.* **91,** 54–74.
Vicsek, T. (1992). "Fractal Growth Phenomena," 2nd ed., World Scientific, Singapore.
Voss, R., and Clarke, J. (1975). *Nature* **258,** 317–318.
West, B. (1990). "Fractal Physiology and Chaos in Medicine," World Scientific, Singapore.
Weyl, H. (1912). *Math. Ann.* **71,** 441–479.
Williams, R. (1983). *Ergodic Theory Dynam. Syst.* **4,** 147–163.
Wiggins, S. (1990). "Introduction to Applied Nonlinear Dynamical Systems and Chaos," Springer-Verlag, New York.
Witten, T., and Sander, L. (1981). *Phys. Rev. Lett.* **47,** 1400–1403.
Witten, T., and Sander, L. (1983). *Phys. Rev. B* **27,** 5686–5697.



Functional Analysis

C. W. Groetsch
University of Cincinnati

I. Linear Spaces
II. Linear Operators
III. Contractions
IV. Some Principles and Techniques
V. A Few Applications

GLOSSARY

Banach space A complete normed linear space.
Bounded linear operator A continuous linear operator acting between normed linear spaces.
Compact linear operator A linear operator that maps weakly convergent sequences into convergent sequences.
Decomposition theorem A Hilbert space is the sum of any closed subspace and its orthogonal complement.
Fredholm alternative A theorem that characterizes the solvability of a compact linear operator equation of the second kind.
Hilbert space A complete inner product space.
Inner product A positive definite (conjugate) symmetric bilinear form defined on a linear space.
Linear operator A mapping acting between linear spaces that preserves linear combinations.
Linear functional A linear operator whose range is the field of scalars.
Norm A real-valued function on a linear space having the properties of length.
Orthogonal complement All vectors which are orthogonal to a given set.
Orthogonal vectors Vectors whose inner product is zero; a generalization of perpendicularity.
Riesz representation theorem A theorem that identifies bounded linear functionals on a Hilbert space with vectors in the space.
Spectral theorem Characterizes the range of a compact self-adjoint linear operator as the orthogonal sum of eigenspaces.

FUNCTIONAL ANALYSIS strives to bring the techniques and insights of geometry to bear on the study of functions by treating classes of functions as "spaces." The interplay between such spaces and transformations between them is the central theme of this *geometrization* of mathematical analysis. The spaces in question are *linear* spaces, that is, collections of functions (or more general "vectors") endowed with operations of addition and scalar multiplication that satisfy the well-known axioms of a vector space. The subject is rooted in the vector analysis of J. Willard Gibbs and the investigations of Vito Volterra and David Hilbert in the theory linear integral equations; it enjoyed youthful vigor, spurred by applications in potential theory and quantum mechanics, in

the works of Stefan Banach, John von Neumann, and Marshal Stone on the theory of linear operators, and settled into a dignified maturity with the contributions of M. A. Naimark, S. L. Sobolev, and L. V. Kantorovich to normed rings, partial differential equations, and numerical analysis, respectively. Functional analysis provides a general context, that is rich in geometric and algebraic nuances, for organizing and illuminating the structure of many areas of mathematical analysis and mathematical physics. Further, the subject enables rigorous justification of mathematical principles by use of general techniques of great power and wide applicability. In recent decades functional analysis has become the lingua franca of applied mathematics, mathematical physics and the computational sciences, and it has made inroads into statistics and the engineering sciences. In this survey we will concentrate on functional analysis in normed linear spaces, with special emphasis on Hilbert space.

I. LINEAR SPACES

Functional analysis is a geometrization of analysis accomplished via the notion of a linear space. Linear spaces provide a natural algebraic and geometric framework for the description and study of many problems in analysis.

Definition. A *linear space* (vector space) is a collection $\mathcal{V}$ of elements, called *vectors*, along with an associated field of *scalars*, F, and two operations, vector addition (denoted "+") and multiplication of a vector by a scalar (the scalar product of $\alpha \in F$ with $x \in \mathcal{V}$ is denoted αx). The set $\mathcal{V}$ is assumed to form a commutative group under vector addition (the additive identity is called the *zero vector*, denoted θ, and the additive inverse of a vector x is denoted as $-x$). In addition, the scalar product respects the usual algebraic customs, namely:

$$\begin{aligned} 1x &= x, \\ \alpha(\beta x) &= (\alpha\beta)x, \\ (\alpha+\beta)x &= \alpha x + \beta x, \\ \alpha(x+y) &= \alpha x + \alpha y, \end{aligned}$$

for any $\alpha, \beta \in F$ and any $x, y \in \mathcal{V}$, where 1 denotes the multiplicative identity in F.

The vector $x + (-y)$ is normally written $x - y$. We always take the field F to be the field of real numbers, $\mathbf{R}$, or (less frequently) the field of complex numbers, $\mathbf{C}$. The prototype of all linear spaces is Euclidean n-space, that is, the vector space $\mathbf{R}^n$ consisting of all ordered n-tuples of real numbers with addition and scalar multiplication defined componentwise. That is, for

$$\begin{aligned} x &= [x_1, x_2, \ldots, x_n] \\ y &= [y_1, y_2, \ldots, y_n] \end{aligned}$$

in $\mathbf{R}^n$, and $\alpha \in \mathbf{R}$, the vector operations are defined by

$$\begin{aligned} x + y &= [x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n] \\ \alpha x &= [\alpha x_1, \alpha x_2, \ldots, \alpha x_n] \end{aligned}$$

The most useful linear spaces are *function spaces*. For example, the space, $\mathcal{F}(\Omega)$, of all real-valued functions defined on a set Ω is a linear space when fitted out with the operations

$$\begin{aligned} (f+g)(s) &= f(s) + g(s) \\ (\alpha f)(s) &= \alpha f(s). \end{aligned}$$

Taking $\Omega = \{1, 2, \ldots, n\}$ we see that, with obvious notational conventions, Euclidean n-space is a special case of the function space $\mathcal{F}(\Omega)$. A more useful space is the space $C[a, b]$ consisting of all real-valued functions defined on the interval $[a, b]$ that are continuous at each point of $[a, b]$.

Definition. A subset $\mathcal{S}$ of a linear space $\mathcal{V}$ is called a *subspace* of $\mathcal{V}$ if $\alpha x + \beta y \in \mathcal{S}$ for all $x, y \in \mathcal{S}$ and all $\alpha, \beta \in F$.

A subspace is therefore a subset of a linear space which is a linear space in its own right. Applying the definition iteratively, one sees that a subspace $\mathcal{S}$ contains all *linear combinations* of the form

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

where n is a positive integer, $\alpha_1, \alpha_2, \ldots, \alpha_n$ are arbitrary scalars and $x_1, x_2, \ldots, x_n$ are arbitrary vectors in S. A set of vectors $\mathcal{W}$ is called *linearly independent* if no vector in $\mathcal{W}$ can be written as a linear combination of other vectors in the set $\mathcal{W}$. One might then say that a linearly independent set contains no "linear redundancies." A *basis* for a subspace $\mathcal{S}$ is a set of linearly independent vectors $\mathcal{B} \subset \mathcal{S}$ with the property that every vector in $\mathcal{S}$ can be written as a linear combination of vectors in $\mathcal{B}$. If a subspace has a basis consisting of finitely many vectors, then the subspace is called *finite-dimensional*. For example, the set of all real polynomials of degree not more than 7 is an eight-dimensional subspace of $C[a, b]$; the set of polynomials $\{1, t, t^2, \ldots, t^7\}$ is a basis for this subspace. On the other hand, the set of all real polynomials is an infinite dimensional subspace of $C[a, b]$.

Subspaces allow a layering of sets of increasingly regular vectors while preserving the linear structure of the space. For example, the set $C^1[a, b]$ of all real-valued functions on $[a, b]$ which also have a continuous derivative is a subspace of $C[a, b]$, while $C_0^2[a, b]$, the space of all twice continuously differentiable real-valued functions

on $[a, b]$ that vanish at the endpoints of the interval, is a subspace of $C^1[a, b]$.

The space $AC[a, b]$ of *absolutely continuous* functions is another important subspace of $C[a, b]$. A function f is called absolutely continuous if the sum $\sum |f(b_i) - f(a_i)|$ can be made arbitrarily small for all finite collections of subintervals $\{[a_i, b_i]\}$ whose total length $\sum(b_i - a_i)$ is sufficiently small. The importance of absolutely continuous functions resides in the fact that they are precisely those functions which have integrable (but not necessarily continuous) derivatives on $[a, b]$. Functions in $C^1[a, b]$ are absolutely continuous because such functions have uniformly bounded derivatives. We therefore have the following relationships for these subspaces of $C[a, b]$:

$$C^1[a, b] \subset AC[a, b] \subset C[a, b].$$

A somewhat less specialized notion than a subspace, that of a convex set, is crucial in various applications of functional analysis.

Definition. A subset K of a linear space is called *convex* if $(1 - t)x + ty \in K$ for all $x, y \in K$ and all real numbers t with $0 \le t \le 1$.

Convex sets have a natural geometric interpretation. Given two vectors x and y, the set of vectors of the form $\{(1 - t)x + ty : t \in [0, 1]\}$ is called, in analogy with the corresponding situation in Euclidean space, the *segment* between x and y. A convex set is then a subset of a linear space which contains the segment between any two of its vectors. Of course, every subspace is a convex set.

A. Normed Linear Spaces

Vector spaces become more serviceable when a metric fabric is stretched over the linear framework. Such a metric structure is provided by a function defined on the space which generalizes the concept of length in Euclidean spaces.

Definition. A nonnegative real-valued function $\|\cdot\|$ defined on a linear space $\mathcal{V}$ is called a *norm* if:

(i) $\|x\| = 0$ if and only if $x = \theta$ (the zero vector in $\mathcal{V}$),

(ii) $\|x + y\| \le \|x\| + \|y\|$, for all $x, y \in \mathcal{V}$,

(iii) $\|tx\| = |t|\|x\|$, for all scalars t and all $x \in \mathcal{V}$.

A linear space that comes bundled with a norm is called a *normed linear space*. For specificity, we will sometimes indicate a linear space $\mathcal{V}$ equipped with a norm $\|\cdot\|$ by the pair $(\mathcal{V}, \|\cdot\|)$. For example, $(C[a, b], \|\cdot\|_\infty)$ is a normed linear space under the so-called *uniform norm*

$$\|f\|_\infty = \max\{|f(t)| : t \in [a, b]\}.$$

The space $C[a, b]$ is also a normed linear space when endowed with the norm

$$\|f\|_1 = \int_a^b |f(t)|\, dt$$

and hence $(C[a, b], \|\cdot\|_1)$ is another distinct normed space consisting of the same set of vectors as the normed linear space $(C[a, b], \|\cdot\|_\infty)$. As another example, $C_0^1[a, b]$ is a normed linear space when equipped with the norm $\|f\| = \|f'\|_\infty$, where f' is the derivative of f.

A norm allows the possibility of measuring the distance between two vectors and it also gives meaning to "equality in the limit," that is, to *convergence*.

Definition. A sequence $\{x_n\}$ of vectors is said to *converge* to a vector x, with respect to the norm $\|\cdot\|$, if $\|x_n - x\| \to 0$ as $n \to \infty$.

The notion of convergence is norm-dependent. For example, let l_n denote the "spike" function in $C[0, 1]$ whose graph is obtained by connecting the points $(0, 0)$, $(\frac{1}{2n}, 1)$, $(\frac{1}{n}, 0)$, $(1, 0)$ with straight line segments. The sequence $\{l_n\}$ converges with respect to the $\|\cdot\|_1$ norm to the zero function since $\|l_n\|_1 = \frac{1}{2n}$. However, this sequence does not converge to the zero function in the uniform norm since $\|l_n\|_\infty = 1$. But note that every sequence that converges with respect to the uniform norm also converges with respect to the norm $\|\cdot\|_1$ since $\|f\|_1 \le \|f\|_\infty$ for all $f \in C[0, 1]$. As another example, consider the space $C_0^1[a, b]$ with the norm $\|f\| = \|f'\|_2$. Wirtinger's inequality asserts that if $f \in C_0^1[a, b]$, then

$$\pi^2 \int_a^b |f(t)|^2\, dt \le (b - a)^2 \int_a^b |f'(t)|^2\, dt$$

Hence, if $\|\cdot\|_2$ is the norm on $C_0^1[a, b]$ defined by

$$\|f\|_2 = \sqrt{\int_a^b |f(t)|^2\, dt}$$

then $m\|f\|_2 \le \|f\|$, where $m = \pi/(b - a)$. Therefore, convergence in the norm $\|\cdot\|$ implies convergence in the norm $\|\cdot\|_2$. Wirtinger's inequality can be generalized to functions defined on domains in $\mathbf{R}^n$; the resulting inequality is known as *Poincaré's inequality*.

Two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on a linear space $\mathcal{V}$ are called *equivalent* if there are positive constants c_1 and c_2 such that

$$c_1\|x\|_a \le \|x\|_b \le c_2\|x\|_a$$

for all $x \in \mathcal{V}$. Equivalence therefore means equivalent in the sense of convergence: a sequence converges with respect to the norm $\|\cdot\|_a$ if and only if it converges with respect to the norm $\|\cdot\|_b$. It can be shown that any two norms on a finite-dimensional linear space are equivalent and hence when speaking of convergence in a finite-dimensional space one can dispense with mentioning the norm. However, as we have seen above, convergence is norm-dependent in infinite dimensional spaces.

Definition. The *closure*, $\overline{W}$, of a subset W of a normed linear space $(\mathcal{V}, \|\cdot\|)$ is the set consisting of elements of W along with all limits of convergent sequences in W. A set which is its own closure is called *closed*.

For example, the closure of the set of all polynomials in the space $(C[a, b], \|\cdot\|_\infty)$ is, by virtue of the *Weierstrass approximation theorem*, the space $C[a, b]$ itself. Also, for example, the set of nonnegative functions in $C[a, b]$ is a closed subset of $C[a, b]$.

B. Banach Spaces

Certain sequences of vectors in a normed linear space tend to "bunch up" in a way that imitates convergent sequences. Such sequences are named after the nineteenth century mathematician Augustin Cauchy.

Definition. A sequence of vectors $\{x_n\}$ in a normed linear space $(\mathcal{V}, \|\cdot\|)$ is called a *Cauchy* sequence if $\lim_{n,m\to\infty} \|x_n - x_m\| = 0$.

It is easy to see that every convergent sequence is Cauchy, however, it is not necessarily the case that a Cauchy sequence is convergent. Consider, for example, the "ramp" function h_n in $C[-1, 1]$ whose graph consists of the straight line segments connecting the points $(-1, 0), (-\frac{1}{n}, 0), (0, 1), (1, 1)$. The sequence of ramp functions is Cauchy with respect to the norm $\|\cdot\|_1$ since

$$\|h_n - h_m\|_1 = \frac{1}{2}\left|\frac{1}{n} - \frac{1}{m}\right| \to 0, \quad \text{as} \quad n, m \to \infty.$$

However, $\{h_n\}$ does not converge, with respect to the norm $\|\cdot\|_1$, to a function in $C[-1, 1]$ (the $\|\cdot\|_1$ limit of this sequence of ramp functions is the discontinuous Heaviside function). Spaces, such as $(C[-1, 1], \|\cdot\|_1)$, which do not accommodate limits of Cauchy sequences are in some sense "incomplete." A normed linear space is called *complete* if every Cauchy sequence in the space converges to some vector in the space.

Definition. A *Banach space* is a complete normed linear space.

We have just seen that $C[a, b]$ with the norm $\|\cdot\|_1$ is not a Banach space. However, $C[a, b]$ with the uniform norm is a Banach space. The Lebesgue spaces are particularly important Banach spaces. For $1 \le p < \infty$ the space $L^p[a, b]$ consists of measurable real-valued functions f defined on $[a, b]$ such that $|f|^p$ has a finite Lebesgue integral. The norm $\|\cdot\|_p$ on $L^p[a, b]$ is defined by

$$\|f\|_p = \left\{\int_a^b |f(t)|^p \, dt\right\}^{\frac{1}{p}}.$$

Every normed linear space can be imbedded as a dense subspace of a Banach space by an abstract process of "completion." Essentially, the completion process involves adjoining to the original space, for each Cauchy sequence in the original space, a vector in an extended space (the *completion*) which is the limit in the extended space of the Cauchy sequence. For example, it can be shown that the completion of the space $(C[a, b], \|\cdot\|_1)$ is the space $(L^1[a, b], \|\cdot\|_1)$.

C. Hilbert Spaces

Perpendicularity, the key ingredient in the Pythagorean theorem, is one of the fundamental concepts in geometry. A successful geometrization of analysis requires the incorporation of this concept. The essential aspects of perpendicularity are captured in special normed linear spaces called inner product spaces.

Definition. An *inner product* on a real linear space $\mathcal{V}$ is a function $\langle\cdot, \cdot\rangle : \mathcal{V} \times \mathcal{V} \to \mathbf{R}$ which satisfies:

(i) $\langle x, x\rangle \ge 0$ for all $x \in \mathcal{V}$,

(ii) $\langle x, x\rangle = 0$ if and only if $x = \theta$ (the zero vector),

(iii) $\langle x, y\rangle = \langle y, x\rangle$ for all $x, y \in \mathcal{V}$,

(iv) $\langle tx, y\rangle = t\langle x, y\rangle$ for all $x, y \in \mathcal{V}$ and all $t \in \mathbf{R}$,

(v) $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$, for all $x, y, z \in \mathcal{V}$.

Properties (iii)–(v) are summarized by saying that $\langle\cdot, \cdot\rangle$ is a symmetric bilinear form; properties (i) and (ii) say that $\langle\cdot, \cdot\rangle$ is nonnegative and definite. A linear space endowed with an inner product is called an *inner product space*.

The most familiar inner product space is the Euclidean space $\mathbf{R}^n$ with the inner product

$$\langle x, y\rangle = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

Of course other inner products may be used on the same underlying linear space. For example, in statistics, the Euclidean space is often used with a *weighted* inner product

$$\langle x, y\rangle = w_1x_1y_1 + w_2x_2y_2 + \cdots + w_nx_ny_n$$

where $w_1, w_2, \ldots, w_n$ are fixed positive weights.

Function spaces serve up a particularly rich stew of inner product spaces. For example, the theory of Fourier series can be developed in the Lebesgue space $L^2[a, b]$ with the inner product

$$\langle f, g\rangle = \int_a^b f(t)g(t)\, dt$$

and many variations are possible. For instance, the space $C^1[0, 1]$ with

$$\langle f, g\rangle = \int_0^1 f'(t)g'(t)\,dt + f(0)g(0)$$

is an inner product space.

In the case of a linear space over the field $\mathbf{C}$ of complex scalars the definition of an inner product requires some modification. Here the inner product is a complex valued function that satisfies the properties above, save that (iii) is replaced by

$$\langle x, y\rangle = \overline{\langle y, x\rangle},$$

where the bar indicates complex conjugation, and (iv) is required to hold for all $t \in \mathbf{C}$. Note that this implies that $\langle x, ty\rangle = \bar{t}\langle x, y\rangle$.

Every inner product on a linear space generates a corresponding norm by way of the definition $\|x\| = \sqrt{\langle x, x\rangle}$. The proof that this defines a norm relies on the *Cauchy-Schwarz inequality*:

$$|\langle x, y\rangle|^2 \le \langle x, x\rangle\langle y, y\rangle.$$

Any norm induced by an inner product must satisfy, by virtue of the bilinearity of the inner product, the *Parallelogram Law*:

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

Therefore, any norm which does not satisfy the parallelogram law is not induced by an inner product. For example, the space $(C[-1, 1], \|\cdot\|_\infty)$ is not an inner product space since the parallelogram law for $\|\cdot\|_\infty$ fails for the continuous ramp functions h_1 and h_2 introduced above.

Definition. Two vectors x and y in an inner product space are called *orthogonal*, denoted $x \perp y$, if $\langle x, y\rangle = 0$. A subset S of an inner product space is called an orthogonal set if $x \perp y$ for each distinct pair of vectors $x, y \in S$.

From the properties of the inner product it follows immediately that

$$x \perp y \quad \text{if and only if} \quad \|x + y\|^2 = \|x\|^2 + \|y\|^2.$$

This extension of the classical Pythagorean theorem to inner product spaces is what justifies our association of orthogonality, a purely *algebraic* concept, with the geometrical notion of perpendicularity.

Given a subset S of an inner product space $\mathcal{V}$, the *orthogonal complement* of S is the closed subspace

$$S^\perp = \{y \in \mathcal{V} : \langle x, y\rangle = 0 \quad \text{for all} \quad x \in S\}.$$

For example, consider the space $C^1[0, 1]$ of continuously differentiable real functions on $[0, 1]$ equipped with the inner product

$$\langle f, g\rangle = \int_0^1 f'(t)g'(t)\,dt + f(0)g(0).$$

Let S be the set of linear functions. Then $g \in S^\perp$ if and only if

$$0 = \langle 1, g\rangle = g(0) \quad \text{and} \quad 0 = \langle t, g\rangle = \int_0^1 g'(t)\,dt = g(1) - g(0).$$

Therefore, $S^\perp = C_0^1[0, 1]$, the space of continuously differentiable functions which vanish at the end points of $[0, 1]$. If $\mathcal{W}$ is a subspace, then its second orthogonal complement is its closure: $\mathcal{W}^{\perp\perp} = \overline{W}$.

Definition. A *Hilbert space* is a complete inner product space.

The best known Hilbert space is the space $L^2[a, b]$. The fact that this space is complete is known as the *Riesz-Fischer Theorem*. The *Sobolev spaces* are important Hilbert spaces used in the theory of differential equations. Let Ω be a bounded domain in $\mathbf{R}^n$. $C^m(\Omega)$ denotes the space of all continuous real-valued functions on $\overline{\Omega}$ having bounded continuous partial derivatives through order m. An inner product $\langle\cdot, \cdot\rangle_m$ is defined on $C^m(\Omega)$ by

$$\langle f, g\rangle_m = \sum_{|\alpha| \le m} \langle D^\alpha f, D^\alpha g\rangle$$

where $\langle\cdot, \cdot\rangle$ denotes the usual L^2 inner product, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ is a multi-index of nonnegative integers, $|\alpha| = \alpha_1 + \cdots + \alpha_n$, and D^α is the differential operator

$$D^\alpha f = \frac{\partial^{|\alpha|}}{\partial x_n^{\alpha_n} \cdots \partial x_1^{\alpha_1}} f.$$

The Sobolev space $H^m(\Omega)$ is the completion of $C^m(\Omega)$ with respect to the norm $\|\cdot\|_m$ generated by the inner product $\langle\cdot, \cdot\rangle_m$. A closely associated Sobolev space is the space $H_0^m(\Omega)$ which is formed in the same way, but using as base space the space $C_0^m(\Omega)$ of all functions in $C^m(\Omega)$ which vanish off of some closed bounded subset of Ω (that is, the space of all *compactly supported* functions in $C^m(\Omega)$). The advantage of using such spaces in the study of differential equations lies in the fact that in these spaces ordinary functions are very smooth, convergence is very strict (convergence in $H^m(\Omega)$ implies uniform convergence on compact subsets of all derivatives through order $m - 1$), and Sobolev spaces are *complete*, allowing the deployment of all the weapons of Hilbert space theory. In particular, Hilbert space theory can be used to develop a rich theory of *weak solutions* of partial differential equations and a corresponding theory of finite element approximations to weak solutions.

An orthogonal set of unit vectors (i.e., vectors of norm one) is called an *orthonormal set*. If S is an orthonormal set in a Hilbert space H, then for any $y \in H$ at most countably many of the numbers $\langle y, x\rangle$, where $x \in S$, are nonzero and

$$\sum_{x\in S} |\langle y, x\rangle|^2 \le \|y\|^2$$

(this is called *Bessel's inequality*). The numbers $\{\langle y, x\rangle: x \in S\}$ are called the *Fourier coefficients* of y with respect to the orthonormal set S. If the orthonormal set S has the property that the Fourier coefficients of a vector completely determine the vector in the sense that if, for any vector $y \in H$

$$\langle y, x\rangle = 0 \quad \text{for all } x \in S \quad \text{implies} \quad y = \theta$$

then S is called a *complete orthonormal set*.

Definition. A Hilbert space is called *separable* if it contains a sequence which is a complete orthonormal set. Any such complete orthonormal sequence is called a *basis* for the Hilbert space.

It is not hard to see that an orthonormal sequence $\{\phi_n\}$ in a Hilbert space H is complete if and only if

$$\sum_n |\langle y, \phi_n\rangle|^2 = \|y\|^2$$

for each $y \in H$ (*Parseval's identity*). This is equivalent to the assertion

$$\sum_n \langle y, \phi_n\rangle \phi_n = y$$

where the convergence of the series is understood in the sense of the norm generated by the inner product $\langle \cdot, \cdot \rangle$.

For example, the complex space $L^2[-\pi, \pi]$ with inner product

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(t)\overline{g(t)}\, dt$$

is a separable Hilbert space with complete orthonormal sequence

$$\phi_n(t) = \frac{1}{\sqrt{2\pi}} e^{int}, \qquad , n = 0, \pm 1, \pm 2, \ldots.$$

Note that these orthonormal vectors are integer dilates of a single complex wave, that is,

$$\phi_n(t) = \phi(nt),$$

where

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{it} = \frac{1}{\sqrt{2\pi}} (\cos t + i \sin t).$$

None of the functions ϕ_n just defined lies in the space $L^2(-\infty, \infty)$ because the waves do not "die out" as $|t| \to \infty$. Special bases for $L^2(-\infty, \infty)$ that have particularly attractive decay properties, called *wavelet* bases, have been studied and applied extensively in recent years. We now show how to construct the simplest wavelet basis, the *Haar basis*. As in the previous example, we seek a single function to generate all the basis vectors, but instead of a wave, we choose a function that decays quickly, in fact a function that vanishes off an interval. Our choice is the Haar "wavelet"

$$\psi(t) = \begin{cases} 1 & \text{for} \quad 0 \le t < \frac{1}{2} \\ -1 & \text{for} \quad \frac{1}{2} \le t < 1 \\ 0 & \text{otherwise} \end{cases}$$

Taking dilates of this wavelet will serve the same purpose as the higher frequency waves in the previous example, that is, the dilates of the fundamental wave ϕ. However, this by itself will not serve to represent functions defined over the entire line. To do that we must also shift the dilates about. If this is done properly, then all the needed time and frequency attributes are captured and the resulting wavelets are orthonormal. Specifically, it can be shown that the functions $\{\psi_{j,k}: j, k = 0, \pm 1, \pm 2, \ldots\}$ defined by

$$\psi_{j,k}(t) = 2^{\frac{j}{2}} \psi(2^j t - k)$$

form an orthonormal basis for $L^2(-\infty, \infty)$ called the *Haar basis*.

II. LINEAR OPERATORS

Mappings between linear spaces that preserve the linear structure are called linear operators.

Definition. A mapping T defined on a linear space $\mathcal{V}$ and taking values in a linear space $\mathcal{W}$ is called *linear* if

$$T(x + y) = Tx + Ty, \quad \text{for all} \quad x, y \in \mathcal{V}$$

and

$$T(tx) = tTx, \quad \text{for all} \quad x \in \mathcal{V} \quad \text{and all scalars} \quad t.$$

The space $\mathcal{V}$ on which a linear operator T is defined is called the *domain* of T and denoted $\mathcal{D}(T)$. In finite-dimensional spaces linear operators have matrix representations relative to specific ordered bases. For example, the differentiation operator acting on the space of polynomials of degree not greater that n has, relative to the basis $\{1, t, \ldots, t^n\}$, the $(n+1) \times (n+1)$ matrix representation

$$\begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 2 & \ldots & 0 \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ 0 & 0 & 0 & \ldots & n \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix}.$$

In functional analysis the primary interest is the study of linear operators on infinite dimensional function spaces that are defined intrinsically, that is, without regard to a specific basis. For example, given a real-valued function $k(\cdot, \cdot)$ which is continuous on the square $[0, 1] \times [0, 1]$ one can define the linear integral operator T on $C[0, 1]$, taking values in $C[0, 1]$, by $Tf = g$ where

$$g(s) = \int_0^1 k(s, t) f(t)\, dt.$$

Such integral operators may be viewed as natural generalizations of matrices to the case of continuous, rather than discrete, variables.

Two important subspaces, the *nullspace* and *range*, are associated with each linear operator. The nullspace, $N(T)$, of a linear operator T consists of all vectors that T maps to the zero vector: $N(T) = \{x \in \mathcal{D}(T)\colon Tx = \theta\}$. For example, the differentiation operator acting on the space $C^1[a, b]$ has nullspace consisting of all constant functions. The range, $R(T)$, of a linear operator T is the set of all images of vectors under T, that is, $R(T) = \{Tx\colon x \in \mathcal{D}(T)\}$. For example, the range of the integral operator on $C[0, 1]$ generated by the kernel $k(s, t) = \exp(s + t)$ consists of all scalar multiples of the exponential function $f(s) = \exp(s)$.

A. Bounded Operators

Linear operators acting between normed linear spaces that are continuous with respect to the norms are called bounded.

Definition. A linear operator T defined on a normed linear space $(\mathcal{V}, \|\cdot\|_v)$ and taking values in a normed linear space $(\mathcal{W}, \|\cdot\|_w)$ is called *bounded* if there is a constant L such that $\|Tx\|_w \le L\|x\|_v$ for all $x \in \mathcal{V}$.

The notational specification of norms on various linear spaces is tiresome and annoying; we will therefore often dispense with it and rely on the reader to understand appropriate norms from the context. Hence we will say that a linear operator T is bounded if $\|Tx\| \le L\|x\|$ for all $x \in \mathcal{D}(T)$ and some constant L. The smallest constant L for which this inequality holds is called the *norm* of the operator T and is denoted $\|T\|$. Equivalently,

$$\|T\| = \sup_{x \neq \theta} \frac{\|Tx\|}{\|x\|}$$

where "sup" stands for supremum, or least upper bound. Using linearity of T one sees that

$$\|Tx - Ty\| \le \|T\|\|x - y\|$$

and hence $\|T\|$ gives the smallest universal bound for the relative "spread" that a bounded linear operator T can accomplish. From this it also follows that every bounded linear operator is continuous (relative to the norms in question) and, in fact, every continuous linear operator is bounded. Indeed, if T is linear and continuous, then there is a $\delta > 0$ such that $\|z\| \le \delta$ implies $\|Tz\| \le 1$. For any $x \neq \theta$ one then has $\|T(\delta x/\|x\|)\| \le 1$, that is, $\|Tx\| \le \frac{1}{\delta}\|x\|$ for all x. Therefore, T is bounded and $\|T\| \le 1/\delta$.

The set $L(\mathcal{V}, \mathcal{W})$ of all bounded linear operators from a normed linear space $\mathcal{V}$ to a normed linear space $\mathcal{W}$ is itself a normed linear space under the natural notions of operator sum and scalar multiplication:

$$(T + S)(x) = Tx + Sx$$
$$(tT)(x) = t(Tx)$$

and with the norm on bounded linear operators as defined above. Further, if $\mathcal{W}$ is a Banach space, then so is $L(\mathcal{V}, \mathcal{W})$. If the image space $\mathcal{W}$ is the scalar field (**R**, or **C**), then $L(\mathcal{V}, \mathcal{W})$ is called the *dual space*, $\mathcal{V}^*$, of $\mathcal{V}$. An operator in $\mathcal{V}^*$ is called a *bounded linear functional* on $\mathcal{V}$. For example, given a fixed $t_0 \in [a, b]$, the point evaluation operator E_{t_0}, defined by

$$E_{t_0}(f) = f(t_0)$$

is a bounded linear functional on the space $(C[a, b], \|\cdot\|_\infty)$. The dual space of $C[a, b]$ may be identified with the space of all functions of bounded variation on $[a, b]$ in the sense that any bounded linear functional ℓ on $C[a, b]$ has a unique representation of the form

$$\ell(f) = \int_a^b f(t)\, dg(t)$$

for some function g of bounded variation (the integral is a Riemann-Stieljes integral). In the case of point evaluation the representative function g is the Heaviside function at t_0:

$$g(t) = \begin{cases} 0 & , \quad t < t_0 \\ 1 & , \quad t_0 \le t. \end{cases}$$

It can be shown that for $1 < p < \infty$,

$$(L^p[a, b])^* = L^q[a, b]$$

where $\frac{1}{p} + \frac{1}{q} = 1$, in the sense that every bounded linear functional ℓ on $L^p[a, b]$ has the form

$$\ell(f) = \int_a^b f(t)g(t)\, dt$$

for some $g \in L^q[a, b]$ that is uniquely determined by ℓ. With this understanding the space $L^2[a, b]$ is self-dual.

In fact, by the *Riesz Representation Theorem*, all Hilbert spaces are self-dual for the following reason: a bounded linear functional ℓ on a Hilbert space H has the form $\ell(x) = \langle x, z\rangle$, for some vector $z \in H$ which is uniquely

determined by ℓ. Further, $\|\ell\| = \|z\|$, where the first norm is an operator norm and the other is the underlying Hilbert space norm. The association of the bounded linear functional ℓ with its Riesz representor z provides the identification of H^* with H.

Bounded linear functionals may be thought of as linear measurements on a Hilbert space in that a bounded linear functional gives a numerical measure on vectors in the space which is continuous with respect to the norm. Such measures distinguish vectors in the space because $x \neq y$ if and only if there is a vector z with $\langle x, z\rangle \neq \langle y, z\rangle$. However, it may happen that no such linear measure can ultimately distinguish the vectors in a sequence from a certain fixed vector. This gives rise to the notion of weak convergence.

Definition. A sequence $\{x_n\}$ in a Hilbert space H is said to converge *weakly* to a vector x, denoted $x_n \rightharpoonup x$, if $\langle x_n, z\rangle \to \langle x, z\rangle$ for all $z \in H$.

It is a consequence of the Cauchy-Schwarz inequality that every convergent sequence is weakly convergent to the same limit. Also, Bessel's inequality shows that every orthonormal sequence of vectors converges weakly to the zero vector.

B. Adjoint Operators

Adjoint operators mimic the behavior of the transpose matrix on real Euclidean space. Recall that the transpose A^T of a real $m \times n$ matrix A satisfies

$$\langle Ax, y\rangle = \langle x, A^T y\rangle$$

for all $x \in \mathbf{R}^n$ and $y \in \mathbf{R}^m$, where $\langle \cdot, \cdot\rangle$ is the Euclidean inner product. If T is a bounded linear operator from a Hilbert space H_1 into a Hilbert space H_2, i.e., $T : H_1 \to H_2$, then for fixed $y \in H_2$ the linear functional ℓ defined on H_1 by

$$\ell(x) = \langle Tx, y\rangle$$

is bounded and hence by the Riesz Representation Theorem

$$\langle Tx, y\rangle = \langle x, z\rangle$$

for some $z \in H_1$. This z is uniquely determined by y, via T, and we denote it by T^*y, that is,

$$\langle Tx, y\rangle = \langle x, T^*y\rangle.$$

The operator $T^* : H_2 \to H_1$ is a bounded linear operator called the *adjoint* of T. If T is a bounded linear operator, then $\|T\| = \|T^*\|$ and $T^{**} = T$.

Suppose, for example, the linear operator $T : L^2[a, b] \to L^2[c, d]$ is generated by the kernel $k(\cdot, \cdot) \in C([c, d] \times [a, b])$, that is,

$$(Tf)(s) = \int_a^b k(s, t) f(t)\, dt, \qquad s \in [c, d],$$

then

$$\begin{aligned}\langle Tf, g\rangle &= \int_c^d \int_a^b k(s, t) f(t)\, dt\, g(s)\, ds \\ &= \int_a^b \int_c^d k(s, t) g(s)\, ds\, f(t)\, dt = \langle f, T^* g\rangle\end{aligned}$$

and hence T^* is the integral operator generated by the kernel $k^*(\cdot, \cdot)$ defined by $k^*(t, s) = k(s, t)$. In particular, if the kernel k is symmetric and $[a, b] = [c, d]$, then the operator T is self-adjoint.

C. Compact Operators

A linear operator $T : H_1 \to H_2$ is called an operator of *finite rank* if its range is spanned by finitely many vectors in H_2. In other words, T has finite rank if there are linearly independent vectors $\{v_1, \ldots, v_m\}$ in H_2 such that

$$Tx = \sum_{k=1}^{m} a_k(x) v_k$$

where the coefficients a_k are bounded linear functionals on H_1. Therefore, by the Riesz Representation Theorem,

$$Tx = \sum_{k=1}^{m} \langle x, u_k\rangle v_k$$

for certain vectors $\{u_1, \ldots, u_m\} \subset H_1$. Note that every finite rank operator is continuous, but more is true. If $\{x_n\}$ is a weakly convergent sequence in H_1, say $x_n \rightharpoonup w$, then

$$Tx_n = \sum_{k=1}^{m} \langle x_n, u_k\rangle v_k \to \sum_{k=1}^{m} \langle w, u_k\rangle v_k = Tw,$$

and hence a finite rank operator T maps weakly convergent sequences into strongly convergent sequences. Finite rank operators are a special case of an important class of linear operators that enjoy this weak-to-strong continuity property.

Definition. A linear operator $T : H_1 \to H_2$ is called *compact* (also called *completely continuous*) if $x_n \rightharpoonup w$ implies $Tx_n \to Tw$.

Note that every finite rank operator is completely continuous and every completely continuous operator is *a fortiori* continuous. Also, limits of finite rank operators are compact. More precisely, if $\{T_n\}$ is a sequence of finite rank operators converging in operator norm to an operator T, then T is compact.

Completely continuous operators acting on a Hilbert space have a particularly simple structure expressed in terms of certain characteristic subpaces known as

eigenspaces. The *eigenspace* of a linear operator $T: H \to H$ associated with a scalar λ is the subspace

$$N(T - \lambda I) = \{x \in H : Tx = \lambda x\}.$$

In general, an eigenspace may be trivial, that is, it may consist only of the zero vector. If $N(T - \lambda I) \neq \{\theta\}$, we say that λ is an *eigenvalue* of T. If T is self-adjoint, then the eigenvalues of T are real numbers and vectors in distinct eigenspaces are orthogonal to each other. If T is self-adjoint, compact, and of infinite rank, then the eigenvalues of T form a sequence of real numbers $\{\lambda_n\}$. This sequence converges to zero, for taking an orthonormal sequence $\{x_n\}$ with $x_n \in N(T - \lambda_n I)$ we have $\lambda_n x_n = Tx_n \to \theta$ since $x_n \rightharpoonup \theta$ (a consequence of Bessel's inequality). Since $\|x_n\| = 1$, it follows that $\lambda_n \to 0$. The fact that $Tx_n \to 0$ for a sequence of unit vectors $\{x_n\}$ is an abstract version of the *Riemann-Lebesgue Theorem*.

The prime exemplar of a compact self-adjoint operator is the integral operator on the real space $L^2[a, b]$ generated by a symmetric kernel $k(\cdot, \cdot) \in L^2([a, b] \times [a, b])$:

$$(Tf)(s) = \int_a^b k(s, t) f(t)\, dt.$$

This operator is compact because it is the limit in operator norm of the finite rank operators

$$T_N f = \sum_{n,m \leq N} c_{n,m} \int_a^b f(t)\phi_m(t)\, dt \phi_n$$

where $\{\phi_n\}_1^\infty$ is a complete orthonormal sequence in $L^2[a, b]$ and

$$k(s, t) = \sum_{n,m=1}^{\infty} c_{n,m}\phi_n(s)\phi_m(t)$$

is the Fourier expansion of $k(\cdot, \cdot)$ relative to the orthonormal basis $\{\phi_n(s)\phi_m(t)\}$ for $L^2([a, b] \times [a, b])$.

D. Unbounded Operators

It is not the case that every interesting linear operator is bounded. For example, the differentiation operator acting on the space $C^1[0, \pi]$ and taking values in the space $C[0, \pi]$, both with the uniform norm, is unbounded. Indeed, for the functions $\phi_n(t) = \sin(nt)$ we find that the quotient

$$\frac{\|\phi_n'\|_\infty}{\|\phi_n\|_\infty} = n$$

is unbounded.

Multiplication by the variable in the space $L^2(-\infty, \infty)$ is another famous example of an unbounded operator. Let

$$\mathcal{D}(M) = \left\{ f \in L^2(-\infty, \infty) : \int_{-\infty}^{\infty} t^2 |f(t)|^2\, dt < \infty \right\}.$$

Then $\mathcal{D}(M)$ is a dense subspace of $L^2(-\infty, \infty)$ and one can define the linear operator $M : \mathcal{D}(M) \to L^2(-\infty, \infty)$ by

$$Mf = g \quad \text{where} \quad g(t) = tf(t).$$

Let $\phi_n(t) = 1$ for $t \in [n, n + 1]$ and $\phi_n(t) = 0$ otherwise. Then $\|\phi_n\|_2 = 1$ and $\|M\phi_n\|_2 > n$, and hence M is unbounded. The multiplication operator has an important interpretation in quantum mechanics.

The adjoint operator may also be defined for unbounded linear operators with dense domains. Given a linear operator $T : \mathcal{D}(T) \subset H_1 \to H_2$ with dense domain $\mathcal{D}(T)$, let $\mathcal{D}(T^*)$ be the subspace of all vectors $y \in H_2$ satisfying

$$\langle Tx, y\rangle = \langle x, y^*\rangle$$

for some vector $y^* \in H_1$ and all $x \in \mathcal{D}(T)$. The vector y^* is then uniquely defined and we set $T^* y = y^*$. Then the operator $T^* : \mathcal{D}(T^*) \subset H_2 \to H_1$ is densely defined and linear.

As an example, let $\mathcal{D}(T)$ be the space of absolutely continuous complex-valued functions f defined on $[0, 1]$ with $f' \in L^2[0, 1]$ satisfying the periodic boundary condition $f(0) = f(1)$. Then $\mathcal{D}(T)$ is dense in $L^2[0, 1]$. Define $T : \mathcal{D}(T) \to L^2[0, 1]$ by $Tf = if'$. For $g \in \mathcal{D}(T)$ we have

$$\begin{aligned}\langle Tf, g\rangle &= \int_0^1 if'(t)\overline{g(t)}\, dt \\ &= i\overline{g(t)}f(t)\Big|_0^1 - i\int_0^1 f(t)\overline{g'(t)}\, dt \\ &= \int_0^1 f(t)\overline{ig'(t)}\, dt = \langle f, ig'\rangle\end{aligned}$$

for all $f \in \mathcal{D}(T)$. Therefore, $\mathcal{D}(T) \subset \mathcal{D}(T^*)$ and, in fact, it can be shown that $\mathcal{D}(T^*) = \mathcal{D}(T)$. This calculation shows that $T^* g = Tg$, that is, T is self-adjoint.

A linear operator $T : \mathcal{D}(T) \subseteq H_1 \to H_2$ is called *closed* if its graph $\mathcal{G}(T) = \{(x, Tx) : x \in \mathcal{D}(T)\}$ is a closed subspace of the product Hilbert space $H_1 \times H_2$. This means that if $\{x_n\} \subset \mathcal{D}(T)$, $x_n \to x \in H_1$, and $Tx_n \to y \in H_2$, then $(x, y) \in \mathcal{G}(T)$, that is, $x \in \mathcal{D}(T)$ and $Tx = y$. For example, the differentiation operator defined in the previous paragraph is closed. In fact, the adjoint of any densely defined linear operator is closed.

A densely defined linear operator $T : \mathcal{D}(T) \subseteq H \to H$ is called *symmetric* if

$$\langle Tx, y\rangle = \langle x, Ty\rangle \quad \text{for all} \quad x, y \in \mathcal{D}(T).$$

Every self-adjoint transformation is, of course, symmetric; however, a symmetric transformation is not necessarily self-adjoint. Consider, for instance, a slight modification of the previous example. Let $\mathcal{D}(T)$ be the space of absolutely continuous complex-valued functions on $[0, 1]$ which vanish at the end points, and let $Tf = if'$. For

$f, g \in \mathcal{D}(T)$, integration by parts gives $\langle Tf, g\rangle = \langle f, Tg\rangle$, and hence T is symmetric, and the adjoint of T satisfies $T^*g = if'$. However, $\mathcal{D}(T^*)$ is a proper extension of $\mathcal{D}(T)$, in that no boundary conditions are imposed on functions in $\mathcal{D}(T^*)$, and hence T is not self-adjoint. The examples just given show that a symmetric linear operator is not necessarily bounded. The *Hellinger-Toeplitz* Theorem gives sufficient conditions for a symmetric operator to be bounded: a symmetric linear operator whose domain is the entire space is bounded.

If a linear operator $T : \mathcal{D}(T) \subseteq H_1 \to H_2$ is closed, then $\mathcal{D}(T)$ is a Hilbert space when endowed with the *graph inner product*:

$$\langle (x, Tx), (y, Ty)\rangle = \langle x, y\rangle + \langle Tx, Ty\rangle.$$

If T is closed and everywhere defined, i.e., $\mathcal{D}(T) = H_1$, then since the graph norm dominates the norm on H_1, we find, by the corollary to Banach's theorem (see the *inversion* section), that the norm in H_1 is equivalent to the graph norm. In particular, the operator T is then bounded. This is the *closed graph theorem*: a closed everywhere defined linear operator is bounded.

III. CONTRACTIONS

Suppose X is a Banach space and $D \subseteq X$. A vector $x \in D$ is called a *fixed point* of the mapping $T : D \to X$ if $Tx = x$. Every linear operator has a fixed point, namely the zero vector. But a nonlinear mapping may be free of fixed points. However, a mapping that draws points together in a uniform relative sense (a condition called contractivity) is guaranteed to have a fixed point.

Definition. A mapping $T : D \subseteq X \to X$ is called a *contraction* (relative to the norm $\|\cdot\|$ on X) if, $\|Tx - Ty\| \le \alpha\|x - y\|$, for all $x, y \in D$ and some positive constant $\alpha < 1$.

For example, if $L : X \to X$ is a bounded linear operator with $\|L\| < 1$, and $g \in X$, then the (*affine*) mapping $T : X \to X$ defined by $Tx = Lx + g$ is a contraction with contraction constant $\alpha = \|L\|$. As another example, consider the nonlinear integral operator $T : C[0, 1] \to C[0, 1]$ defined by

$$(Tu)(s) = \int_0^1 k(s, u(t))\, dt$$

where $k(\cdot, \cdot) \in C^1([0, 1] \times \mathbf{R})$ is a given kernel. If the kernel satisfies

$$\left|\frac{\partial}{\partial t}k(s, t)\right| \le \alpha, \quad \text{for all} \quad s, t$$

then

$$|(Tw)(s) - (Tu)(s)| = \left|\int_0^1 k(s, w(t)) - k(s, u(t))\, dt\right|$$
$$\le \alpha \int_0^1 |w(t) - u(t)|\, dt \le \alpha\|w - u\|.$$

Therefore, $\|Tw - Tu\| \le \alpha\|w - u\|$ and hence T is a contraction for the uniform norm if $\alpha < 1$.

The *Contraction Mapping Theorem* (elucidated by Banach in 1922) is a constructive existence and uniqueness theorem for fixed points. If D is a closed subset of a Banach space and $T : D \to D$ is a contraction, then the theorem guarantees the existence of a unique fixed point $x \in D$. This fixed point is the limit of any sequence constructed iteratively by $x_{n+1} = Tx_n$, where x_0 is an arbitrary vector in D. If α is a contraction constant for T, then there is an a priori error bound

$$\|x_n - x\| \le \frac{\alpha^n}{1 - \alpha}\|x_1 - x_0\|$$

and an *a posteriori* error bound

$$\|x_n - x\| \le \frac{1}{1 - \alpha}\|x_{n+1} - x_n\|$$

for x_n as an approximation to the unique fixed point x.

The contraction mapping theorem is often used to establish the existence and uniqueness of solutions of problems in function spaces. One such application is a simple *implicit function theorem*. Suppose $f \in C([a, b] \times \mathbf{R})$ satisfies

$$0 < m \le \left|\frac{\partial f}{\partial s}(t, s)\right| \le M$$

for some constants m and M. Then one can show that the equation $f(t, x) = 0$ implicitly defines a continuous function x on $[a, b]$. Indeed, the nonlinear operator $T : C[a, b] \to C[a, b]$ defined by

$$(Tx)(t) = x(t) - \frac{2}{m + M} f(t, x(t))$$

is, under the stated conditions, a contraction mapping on $C[a, b]$ with contraction constant $\alpha = (M - m)/(M + m)$. Therefore, T has a unique fixed point $x \in C[a, b]$. That is, there is a unique function $x \in C[a, b]$ satisfying

$$x(t) = x(t) - \frac{2}{m + M} f(t, x(t))$$

or, equivalently $f(t, x(t)) = 0$.

IV. SOME PRINCIPLES AND TECHNIQUES

A. Projection and Decomposition

The *projection property* is a key feature of the geometry of Hilbert space: a closed convex subset of a Hilbert space

contains a unique vector of smallest norm. By shifting the origin to a vector x, one sees that this is equivalent to saying that given a closed convex subset S of a Hilbert space H and a vector $x \in H$, there is a unique vector $Px \in S$ satisfying

$$\|x - Px\| = \min_{y \in S} \|x - y\|.$$

This purely geometric projection property has important applications in optimization theory. For example, consider the following simple example of optimal control of a one-dimensional dynamical system. Suppose a unit point mass is steered from the origin with initial velocity 1 by a control (external force) u. We are interested in a control that will return the particle to a "soft landing" at the origin in unit time while expending minimal effort, where the measure of effort is

$$\int_0^1 |u(t)|^2 \, dt.$$

We may formulate this problem in the Hilbert space $L^2[0, 1]$. The dynamics of the system are governed by the equations

$$\ddot{x} = u, \quad x(0) = 0, \quad \dot{x}(0) = 1, \quad x(1) = 0, \quad \dot{x}(1) = 0.$$

Suppose C is the set of all vectors u in $L^2[0, 1]$ for which the equations above are satisfied for some vector $x \in H^2[0, 1]$. It may be routinely verified that C is a closed convex subset of $L^2[0, 1]$ and hence C contains a unique vector of smallest L^2-norm, i.e., there is a unique minimal effort control that steers the system in the specified manner.

The (generally nonlinear) operator P defined above is called the (metric) *projection* of H onto S. If S is a closed subspace of H, then P is a bounded self-adjoint linear operator and $I - P$, where I is the identity operator, is the projection of H onto $S^\perp$. Since $x = Px + (I - P)x$, this provides a Cartesian decomposition of H, written $H = S \oplus S^\perp$, meaning that each vector in H can be written uniquely as a sum of a vector in S and a vector in $S^\perp$.

For example, suppose H is the completion of the space $C^1[0, 1]$ with respect to the inner product

$$\langle f, g \rangle = \int_0^1 f'(t)g'(t) \, dt + f(0)g(0),$$

and let S be the subspace of linear functions. Then $H = S \oplus S^\perp$ and we have seen that $S^\perp$ consists of those functions in H which vanish at 0 and 1. So in this instance the decomposition theorem expresses the fact that each function in H can be uniquely decomposed into the sum of a linear function and a function in H that vanishes at both end points of [0, 1].

If S is a separable closed subspace of a Hilbert space H, then a representation of the projection operator P of H onto the subspace S can be given in terms of a complete orthonormal sequence $\{\phi_n\}$ for S. Indeed, if $x \in H$, then

$$\sum_n \langle x, \phi_n \rangle \phi_n - x \in S^\perp$$

since this vector is orthogonal to each member of the basis $\{\phi_n\}$ for S. Therefore,

$$Px = Px + P\left(\sum_n \langle x, \phi_n \rangle \phi_n - x \right) = \sum_n \langle x, \phi_n \rangle \phi_n.$$

There is an important relationship involving the nullspace, range, and adjoint of a bounded linear operator acting between Hilbert spaces. If $T : H_1 \to H_2$ is a bounded linear operator, then $N(T^*) = R(T)^\perp$. To see this, note that $w \in R(T)^\perp$ if and only if

$$0 = \langle Tx, w \rangle = \langle x, T^* w \rangle$$

for all $x \in H_1$, that is, if and only if $w \in N(T^*)$. By a previously discussed result on the second orthogonal complement we get the related result that $\overline{R(T)} = N(T^*)^\perp$. In particular, if T is a bounded linear operator with closed range, then the equation $Tf = g$ has a solution if and only if g is orthogonal to all solutions x of the homogeneous adjoint equation $T^* x = \theta$.

Replacing T with T^* and noting that $T^{**} = T$, we obtain two additional relationships between the nullspace, range, and adjoint. Taken together these relationships, namely

$$N(T^*) = R(T)^\perp, \qquad N(T^*)^\perp = \overline{R(T)},$$
$$N(T) = R(T^*)^\perp, \qquad N(T)^\perp = \overline{R(T^*)}$$

are sometimes collectively called the theorem on the *four fundamental subspaces*.

The Riesz Representation Theorem is a simple consequence of the decomposition theorem. If ℓ is a nonzero bounded linear functional on a Hilbert space H, then $N(\ell)^\perp = R(\ell^*)$ is one-dimensional and $H = N(\ell) \oplus N(\ell)^\perp$. Let $y \in N(\ell)^\perp$ be a unit vector. Then $x = Px + \langle x, y \rangle y$, where P is the projection operator of H onto $N(\ell)$. Therefore,

$$\ell(x) = \ell(\langle x, y \rangle y) = \langle x, y \rangle \ell(y) = \langle x, z \rangle$$

where $z = \overline{\ell(y)} y$.

B. The Spectral Theorem

We limit our discussion of the spectral theorem to the case of a compact self-adjoint operator. The spectral theorem gives a particularly simple characterization of the range

of such an operator. We have seen that the nonzero eigenvalues of a compact self-adjoint operator T of infinite rank form a sequence of real numbers $\{\lambda_n\}$ with $\lambda_n \to 0$ as $n \to \infty$. The corresponding eigenspaces $N(T - \lambda_n I)$ are all finite-dimensional and there is a sequence $\{v_n\}$ of orthonormal eigenvectors, that is, vectors satisfying

$$\|v_j\| = 1, \quad \langle v_i, v_j \rangle = 0, \quad \text{for} \quad i \neq j,$$
$$\text{and,} \quad Tv_j = \lambda_j v_j.$$

The closure of the range of T is the closure of the span of this sequence of eigenvectors. Since T is self-adjoint, $N(T)^\perp = \overline{R(T)}$; the decomposition theorem then gives the representation

$$w = Pw + \sum_{j=1}^{\infty} \langle w, v_j \rangle v_j$$

for all $w \in H$, where P is the projection operator from H onto $N(T)$. The range of T then has the form

$$Tw = \sum_{j=1}^{\infty} \lambda_j \langle w, v_j \rangle v_j.$$

This result is known as the spectral theorem. It can be extended (in terms of a Stieljes integral with respect to a projection valued measure on the real line) to bounded self-adjoint operators (and beyond).

C. The Singular Value Decomposition

The decomposition of a Hilbert space into the nullspace and eigenspaces of a compact self-adjoint operator can be simply extended to obtain a similar decomposition, called the *singular value decomposition* (SVD), for compact operators which are not necessarily self-adjoint. If $T : H_1 \to H_2$ is a compact linear operator from a Hilbert space H_1 into a Hilbert space H_2, then the operators T^*T and TT^* are both self-adjoint compact linear operators with nonnegative eigenvalues. By the spectral theorem, T^*T has an orthonormal sequence of eigenvectors, $\{u_j\}$, associated with its positive eigenvalues $\{\lambda_j\}$, that is complete in the subspace

$$\overline{R(T^*T)} = N(T^*T)^\perp = N(T)^\perp.$$

The numbers $\mu_j = \sqrt{\lambda_j}$ are called the *singular values* of T. If the vectors $\{v_j\}$ are defined by $v_j = \mu_j^{-1} T u_j$, then $\{v_j\}$ is a complete orthonormal set for $N(T^*)^\perp$ and the following relations hold:

$$Tu_j = \mu_j v_j \qquad \text{and} \qquad T^* v_j = \mu_j u_j.$$

The system $\{u_j, v_j; \mu_j\}$ is called a *singular system* for the operator T, and any $f \in H_1$ has, by the decomposition theorem, a representation in the form

$$f = Pf + \sum_{j=1}^{\infty} \langle f, u_j \rangle u_j$$

where P is the projection of H_1 onto $N(T)$ (the sum is finite if T has finite rank). It then follows that

$$Tf = \sum_{j=1}^{\infty} \mu_j \langle f, u_j \rangle v_j.$$

This is called the SVD of the compact linear operator T. The SVD may be viewed as a nonsymmetric spectral representation.

D. Operator Equations

Suppose $T : H_1 \to H_2$ is a compact linear operator and $\lambda \in \mathbf{C}$. If $\lambda \neq 0$, then it can be shown that $R(T - \lambda I)$ is closed. The *Fredholm Alternative* Theorem completely characterizes the solubility of equations of the type

$$Tf - \lambda f = g$$

where $\lambda \neq 0$ and $g \in H_2$. Such equations are called linear operator equations of the *second kind*. The "alternative" is this: either the operator $(T - \lambda I)$ has a bounded inverse, or λ is an eigenvalue of T. If the first alternative holds, then the equation has the unique solution $f = (T - \lambda I)^{-1} g$, and this solution depends continuously on g. In this case we say that the equation of the second kind is *well-posed*. On the other hand, if λ is an eigenvalue of T, then $\bar{\lambda}$ is an eigenvalue of T^* and the equation has a solution only if

$$g \in R(T - \lambda I) = N(T^* - \lambda I)^\perp.$$

That is, the equation has a solution only if g is orthogonal to the finite-dimensional eigenspace $N(T^* - \bar{\lambda} I)$.

The Fredholm Alternative is particularly simple if $T : H \to H$ is self-adjoint. In this case the eigenvalues are real and $N(T^* - \bar{\lambda} I) = N(T - \lambda I)$. Therefore, if λ is not an eigenvalue of T, that is, if the equation $Tf - \lambda f = g$ has no more than one solution, then

$$H = \{\theta\}^\perp = N(T - \lambda I)^\perp = R(T - \lambda I)$$

and hence the equation has a solution for any $g \in H$. Simply put, the Fredholm Alternative for a self-adjoint compact operator says that uniqueness of solutions $(N(T - \lambda I) = \{\theta\})$ implies existence of solutions for any right hand side $(R(T - \lambda I) = H)$.

The contraction mapping theorem can be applied to establish the existence and uniqueness of the solutions of certain *nonlinear* operator equations of the second kind, that is, equations of the form $Tf - \lambda f = g$. For example,

suppose $T: X \to Y$ is a mapping from a Banach space X to a Banach space Y, satisfying $\|Tu - Tv\| \le \mu\|u - v\|$. For nonzero $\lambda \in \mathbf{C}$ the equation $Tf - \lambda f = g$ then has a unique solution $f \in X$ for each $g \in Y$, if $\mu < |\lambda|$. Indeed, a solution of this equation is the unique fixed point of the contractive mapping $Af = \frac{1}{\lambda}(Tf - g)$. Further, the inverse mapping $g \mapsto f$ is continuous, that is, the original nonlinear equation of the second kind is well-posed.

Monotone operators form another general class of nonlinear operators for which unique solutions of certain operator equations of the second kind can be assured. Suppose H is a real Hilbert space. An operator $T: H \to H$ is called monotone if $\langle Tx - Ty, x - y\rangle \ge 0$ for all $x, y \in H$. If T is a continuous monotone operator and $\lambda < 0$, then a theorem of Minty insures the existence of a unique solution f of the operator equation of the second kind $Tf - \lambda f = g$, for each $g \in H$. Further, the inverse operator $J = (T - \lambda I)^{-1}$ is not just continuous, but *nonexpansive*, i.e., $\|Jh - Jg\| \le \|h - g\|$.

When $\lambda = 0$, the linear operator equation equation treated above becomes an operator equation of the *first kind*:

$$Tf = g.$$

In this case, the role of the Fredholm Alternative Theorem is to some extent played by *Picard's Theorem*. Let $\{u_j, v_j; \mu_j\}$ be a singular system for the compact linear operator T. Then $\{v_j\}$ is a complete orthonormal system for $\overline{R(T)}$. Picard's Theorem gives a necessary and sufficient condition for a vector $g \in \overline{R(T)} = N(T^*)^\perp$ to lie in $R(T)$, that is, for the equation of the first kind to have a solution. The condition, *Picard's criterion*, is that the singular coefficients of g decay sufficiently quickly, specifically, that

$$\sum_{j=1}^{\infty} \mu_j^{-2} |\langle g, v_j\rangle|^2 < \infty.$$

If Picard's criterion is satisfied, then the series

$$\sum_{j=1}^{\infty} \frac{\langle g, v_j\rangle}{\mu_j} u_j$$

converges to some vector $f \in H_1$ and

$$Tf = \sum_{j=1}^{\infty} \langle g, v_j\rangle v_j = g,$$

and hence the equation has a solution, namely f. The solution is unique only if $N(T) = \{\theta\}$.

E. Inversion

An invertible bounded linear operator need not have a bounded inverse. For example, the operator $T: L^2[0, 1] \to C[0, 1]$ defined by $(Tf)(s) = \int_0^s f(t)\,dt$ has an inverse defined on the subspace of absolutely continuous functions which vanish at 0. But the inverse operator $T^{-1}g = g'$ is an unbounded operator.

If $T \in L(X, Y)$, where X and Y are Banach spaces, then the existence of a bounded inverse for T is equivalent to the condition $\|Tx\| \ge m\|x\|$ for some $m > 0$. Indeed, if this condition is satisfied, then $N(T) = \{\theta\}$, $R(T)$ is closed in Y, and $T^{-1}: R(T) \to X$ satisfies $\|T^{-1}\| \le m^{-1}$. On the other hand, if T^{-1} is bounded, then the condition holds with $m = \|T^{-1}\|^{-1}$. If T is bijective, i.e., if $N(T) = \{\theta\}$ and $R(T) = Y$, then *Banach's Theorem* insures that the inverse of T is also bounded, i.e., $T^{-1} \in L(Y, X)$. One consequence of this fact is that if a normed linear space X is a Banach space under two norms $\|\cdot\|_1$ and $\|\cdot\|_2$, and if these norms satisfy $\|x\|_2 \le M\|x\|_1$ for some $M > 0$, then the norms are equivalent (apply Banach's Theorem to the identity operator $I: (X, \|\cdot\|_1) \to (X, \|\cdot\|_2)$).

The familiar geometric series has an operator analog. If X is a Banach space and $T: X \to X$ is a bounded linear operator with $\|T\| < 1$, then $(I - T)^{-1} \in L(X, X)$ and, in fact,

$$(I - T)^{-1} = I + T + T^2 + \cdots$$

where the series, called the *Neumann series*, converges in the operator norm. This result can be used immediately to guarantee the existence of a unique solution of the linear operator equation of the second kind $Tf - \lambda f = g$, if $|\lambda|$ is sufficiently large ($|\lambda| > \|T\|$ does the trick).

The existence of an inverse of $T \in L(X, Y)$ is equivalent to the unique solvability of the equation $Tf = g$ for each $g \in Y$. This, in turn, is equivalent to $R(T) = Y$ and $N(T) = \{\theta\}$. If either of these conditions is violated, then there is no inverse, but all is not lost. In certain circumstances a *generalized inverse* can be defined; we limit our discussion of the generalized inverse to the Hilbert space context. Suppose $T: H_1 \to H_2$ is a bounded linear operator from a Hilbert space H_1 to a Hilbert space H_2. Suppose we wish to solve the equation $Tf = g$. If $g \notin R(T)$, then the equation has no solution and we might settle for finding an $f \in H_1$ whose image under T is a close to g as possible, that is, we seek $f \in H_1$ satisfying

$$\|Tf - g\| = \inf\{\|Tx - g\| : x \in H_1\}.$$

Such an f is called a *least-squares solution* of $Tf = g$. A least-squares solution exists if and only if the projection of g onto $\overline{R(T)}$ actually lies in $R(T)$. This is equivalent to requiring g to lie in the dense subspace $R(T) + R(T)^\perp$ of H_2. The condition for a least-squares solution may also be phrased as requiring that $Tf - g \in R(T)^\perp = N(T^*)$ (see the theorem on the four fundamental subspaces). This gives another characterization of least-squares solutions: u is a least-squares solution of $Tf = g$ if and only if u

satisfies the *normal equation* $T^*Tu = T^*g$. Since T is bounded, the set of solutions of the normal equation is closed and convex, and hence, by the projection theorem, if there is a least-squares solution, then there is a least-squares solution of smallest norm. These ideas enable us to define a generalized inverse (the *Moore-Penrose generalized inverse*) for T. Let $\mathcal{D}(T^\dagger) = R(T) + R(T)^\perp$ and for $g \in \mathcal{D}(T^\dagger)$, define $T^\dagger g$ to be the least-squares solution having smallest norm. The Moore-Penrose generalized inverse, $T^\dagger : \mathcal{D}(T^\dagger) \subseteq H_2 \to H_1$, is a closed densely defined linear operator. However, $T^\dagger$ is bounded if and only if $R(T)$ is closed. If T is compact, then $R(T)$ is closed if and only if T has finite rank. In particular, a linear integral operator generated by a square integrable kernel has a bounded Moore-Penrose generalized inverse if and only if the kernel is degenerate. If T is compact with singular system $\{u_j, v_j; \mu_j\}$, the the Moore-Penrose generalized inverse has the explicit representation

$$T^\dagger g = \sum_j \frac{\langle g, v_j \rangle}{\mu_j} u_j.$$

F. Variational Inequalities

Suppose H is a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\|\cdot\|$. A bilinear form $a(\cdot, \cdot): H \times H \to \mathbf{R}$ is called *bounded* if there is a constant C such that

$$|a(u, v)| \le C\|u\|\|v\| \quad \text{for all} \quad u, v \in H$$

and *coercive* if there is a constant $m > 0$ such that

$$m\|u\|^2 \le a(u, u)$$

for all $u \in H$. A fundamental result of Stampacchia asserts that if $a(\cdot, \cdot)$ is a bounded, coercive bilinear form (which *need not be symmetric*), and if $f \in H$ and K is a closed convex subset of H, then there is a unique $u \in K$ satisfying

$$a(u, v - u) \ge \langle f, v - u \rangle \quad \text{for all} \quad v \in K.$$

This is called a *variational inequality* for the form $a(\cdot, \cdot)$, the closed convex set K, and the vector $f \in H$.

The fundamental nature of this result becomes apparent when one notices that the projection property for Hilbert space is a special case of this result on variational inequalities. Indeed, if $a(x, y) = \langle x, y \rangle$, then the theorem insures the existence of a unique $u \in K$ satisfying

$$\langle u, v - u \rangle \ge \langle f, v - u \rangle$$

or, equivalently

$$\langle f - u, v - u \rangle \le 0 \quad \text{for all} \quad v \in K.$$

Geometrically, this says that the angle between the vectors $f - u$ and $v - u$ is obtuse for all vectors $v \in K$. From this it follows that $u \in K$ is the vector in K that is nearest to f, that is, $u = Pf$, the metric projection of f onto K.

A nonsymmetric version of the Riesz Representation Theorem also follows from the theorem on variational inequalities. Suppose ℓ is a bounded linear functional on H. Then, by the Riesz Representation Theorem, there is a $f \in H$ such that $\ell(w) = \langle f, w \rangle$ for all $w \in H$. Let the closed convex set K be the entire Hilbert space H, then there is a unique $u \in H$ satisfying

$$a(u, v - u) \ge \langle f, v - u \rangle \quad \text{for all} \quad v \in H$$

and hence $a(u, w) \ge \langle f, w \rangle$ for all $w \in H$. Replacing w by $-w$, we also get $a(u, -w) \ge \langle f, -w \rangle$ for all $w \in H$. Therefore, $a(u, w) = \langle f, w \rangle$ for all $w \in H$. That is, the functional ℓ has the representation $\ell(w) = a(u, w)$ for a unique $u \in H$. This representation of bounded linear functional in terms of a possibly nonsymmetric bilinear form is known as the *Lax-Milgram lemma*. The Lax-Milgram lemma can be used to establish the existence of a unique weak solution for certain nonsymmetric elliptic boundary value problems in the same way that the Riesz Representation Theorem is used to prove the existence of a unique weak solution of the Poisson problem.

As a simple application of the Lax-Milgram lemma, consider the two-point boundary value problem

$$-u'' + u' + u = f, \qquad u'(0) = u'(1) = 0.$$

Integration by parts yields $a(u, v) = \langle f, v \rangle$, where $\langle \cdot, \cdot \rangle$ is the $L^2[0, 1]$ inner product and $a(u, v)$ is the nonsymmetric, bounded, coercive, bilinear form defined on $H^1[0, 1]$ by

$$a(u, v) = \int_0^1 (u'v' + u'v + uv)(s)\, ds.$$

The Lax-Milgram lemma then ensures the existence of a unique weak solution $u \in H^1[0, 1]$ of the boundary value problem, that is, a unique vector $u \in H^1[0, 1]$ satisfying

$$a(u, v) = \langle f, v \rangle \quad \text{for all} \quad v \in H^1[0, 1].$$

V. A FEW APPLICATIONS

A. Weak Solutions of Poisson's Equation

Suppose Ω is a bounded domain in $\mathbf{R}^2$ with smooth boundary $\partial\Omega$. Given $f \in C(\Omega)$ a classical solution of Poisson's equation is a function $u \in C^2(\Omega)$ satisfying

$$\begin{aligned} -\Delta u &= f \quad \text{in} \quad \Omega \\ u &= 0 \quad \text{on} \quad \partial\Omega \end{aligned}$$

where Δ is the Laplacian operator. If the Poisson equation is multiplied by $v \in C_0^2(\Omega)$ and integrated over Ω, then Green's identity gives

$$\langle \nabla u, \nabla v \rangle = \langle f, v \rangle$$

where ∇ is the gradient operator and $\langle \cdot, \cdot \rangle$ is the $L^2(\Omega)$ inner product. There are two things to notice about this equation: $\langle f, v \rangle$ is defined for $f \in L^2(\Omega)$, allowing consideration of "rougher" data f, and on the left-hand side only first derivatives are required rather than second derivatives, allowing less smooth "solutions" u. These observations permit us to propose a *weaker* formulation of the Poisson problem.

The bilinear form $a(\cdot, \cdot)$ defined by $a(u, v) = \langle \nabla u, \nabla v \rangle$ is an inner product (sometimes called the *energy* inner product) on the space $C_0^1(\Omega)$ (this is a consequence of Poincaré's inequality: $a(u, u) \geq C\|u\|_2^2$, where $\|\cdot\|_2$ is the L^2 norm). The norm generated by this inner product is equivalent to the Sobolev $H_0^1(\Omega)$ norm. Therefore, the completion of $C_0^1(\Omega)$ relative to this inner product is the Sobolev space $H_0^1(\Omega)$, and we define a *weak solution* of the Poisson problem to be a vector $u \in H_0^1(\Omega)$ satisfying

$$a(u, v) = \langle f, v \rangle$$

for all $v \in H_0^1(\Omega)$. The linear functional $\ell : H_0^1(\Omega) \to \mathbf{R}$ defined by $\ell(v) = \langle f, v \rangle$ is (again, as a consequence of Poincaré's inequality) a bounded linear functional on $H_0^1(\Omega)$, and hence, by the Riesz Representation Theorem, there is a unique $u \in H_0^1(\Omega)$ satisfying

$$a(u, v) = \langle f, v \rangle \quad \text{for all} \quad v \in H_0^1(\Omega).$$

In other words, Poisson's problem has a unique weak solution for each $f \in L^2(\Omega)$.

B. A Finite Element Method

A finite element method is a constructive precedure for approximating a weak solution by a linear combination of "basis" functions. As a simple illustration we treat a piecewise linear finite element method for the Poisson problem in the plane. A finite dimensional subspace U_N of $H_0^1(\Omega)$ chosen and a basis (whose members are called *finite elements*) is selected for U_N. The finite element approximation to the weak solution will be a certain linear combination of these basis functions, i.e., a member of U_N. First the region Ω is triangulated; the vertices of the resulting triangles are called *nodes* of the triangulation. The functions in U_N will be continuous on Ω, linear on each triangle, and zero on the boundary of Ω. With each interior node of the triangulation we associate a basis function which is 1 at the node and zero at all other nodes of the triangulation (these basis functions are called the linear *Lagrange* elements).The dimension of the finite element subspace is therefore equal to the number of interior nodes in the triangulation and each basis function has a pyramidal shape with a peak at the associated nodal point of the triangulation.

A finite element solution of the Poisson problem is defined by restricting the conditions for a weak solution to the subspace of finite elements, that is, $u_N \in U_N$ is a finite element solution of the Poisson problem if

$$a(u_N, v) = \langle f, v \rangle \quad \text{for all} \quad v \in U_N.$$

When this condition is expressed in terms of the finite element basis, the resulting coefficient matrix is positive definite and hence there is a unique finite element solution.

If u is the weak solution of the Poisson problem, then

$$a(u - u_N, v) = \langle f, v \rangle - \langle f, v \rangle = 0$$

for all $v \in U_N$. Geometrically, this says that the finite element solution is the projection (relative to the energy inner product) of the weak solution onto the finite element subspace and hence,

$$\|u - u_N\| \leq \|u - v\|$$

for all $v \in U_N$, where $\|\cdot\|$ is the energy norm. That is, the finite element solution is the best approximation to the weak solution, with respect to the energy norm, in the finite element subspace.

C. Two-Point Boundary Value Problems

We briefly treat a simple class of Sturm-Liouville problems. The goal is to find $x \in C^2[a, b]$ satisfying the differential equation

$$\frac{d}{ds}\left[p(s)\frac{dx}{ds}\right] - [\mu + q(s)]\,x(s) = f(s)$$

where q, $f \in C[a, b]$ and $p \in C^1[a, b]$ are given functions and $\mu \neq 0$ is a given scalar. Define $T : C_0^2[a, b] \to C[a, b]$ by

$$Tx = [px']' - qx.$$

We suppose the this differential operator is *nonsingular*, that is, $N(T) = \{\theta\}$ (such is the case, for example, if $p(s) < 0$ and $q(s) \leq 0$). Then there is a symmetric *Green's function* $k(\cdot, \cdot) \in C[a, b] \times C[a, b]$ for T. That is, T^{-1} is the integral operator generated by the kernel $k(\cdot, \cdot)$. In other words, $Tx = h$ if and only if

$$x(s) = \int_a^b k(s, t)h(t)\,dt.$$

The original problem may then be expressed as

$$Tx = \mu x + f$$

or, in terms of the Green's function $k(\cdot, \cdot)$:

$$x(s) = \int_a^b k(s, t)[\mu x(t) + f(t)]\,dt.$$

Equivalently,

$$Kx - \lambda x = g$$

where $\lambda = 1/\mu$ and $g = -\lambda K f$, where K is the compact self-adjoint operator on $L^2[a, b]$ generated by the kernel $k(\cdot, \cdot)$. If λ is not an eigenvalue of K, then this integral equation of the second kind has, by Fredholm's Alternative, a solution which is expressable, via the spectral theorem, as a series of eigenfunctions of K. Specifically, $\overline{R(K)}$ has an orthonormal basis of eigenfunctions $\{v_j\}$ of K and therefore,

$$x = (K - \lambda I)^{-1} g = -\lambda (K - \lambda I)^{-1} K f = \sum_j \frac{\lambda \lambda_j}{\lambda - \lambda_j} \langle f, v_j \rangle v_j.$$

D. Inverse Problems

Many inverse problems in mathematical physics can be modeled as operator equations of the first kind, $Kx = y$, where K is a compact linear operator acting between Hilbert spaces. Consider, for example, the simple model problem in gravimetry in which the vertical component of gravity, $y(s)$, along a horizontal segment $0 \le s \le 1$ is engendered by a mass distribution $x(p), 0 \le p \le 1$ on a parallel segment one unit distant from the first. The relationship between y and x is given by

$$y(s) = \gamma \int_0^1 ((s - p)^2 + 1)^{-\frac{3}{2}} x(p)\, dp$$

where γ is a constant. The inverse problem consists of determining the mass distribution from observations of the vertical force y. The model may be phrased abstractly as $y = Kx$, where $K : H \to H$ is a compact linear operator acting on the Hilbert space $H = L^2[0, 1]$.

Formally, the inverse problem may be solved by using the Moore-Penrose generalized inverse: $x = K^\dagger y$. However, the representation of $K^\dagger$ in terms of the SVD of K

$$x = K^\dagger y = \sum_{j=1}^{\infty} \frac{\langle y, v_j \rangle}{\mu_j} u_j$$

points to a serious problem. The singular values $\{\mu_j\}$ converge to zero leading to instability in the solution process. The vector y is a measured entity and hence is subject to error. Suppose, for example, the measured data consists of a vector $y^\delta \in H$ satisfying $\|y - y^\delta\| \le \delta$, where δ is a known bound for the measurement error. While $y^\delta \to y$ as $\delta \to 0$, generally $\|K^\dagger y - K^\dagger y^\delta\| \to \infty$ as $\delta \to 0$. For example, if $\delta = \sqrt{\mu_n}$, where $n = 1, 2, \ldots$, and $y^\delta = y + \sqrt{\mu_n} v_n$, then $\|y - y^\delta\| = \sqrt{\mu_n} \to 0$, while $\|K^\dagger y - K^\dagger y^\delta\| = 1/\sqrt{\mu_n} \to \infty$ as $n \to \infty$. The lesson is that even very accurate measurements can lead to wildly unstable computations.

Stability can be restored (at the expense of accuracy) by the method of *regularization*. In its simplest form this method replaces the normal equation with an augmented (or *regularized*) normal equation

$$K^* K x_\alpha + \alpha x_\alpha = K * y,$$

where $\alpha > 0$ is a *regularization parameter*. That is, an ill-posed equation of the first kind is replaced by an approximating well-posed equation of the second kind. The solution x_α of the regularized equation is stable with respect to perturbations in the data y since $-\alpha < 0$ is not an eigenvalue of $K^* K$ and hence, by the Fredholm Alternative, $(K^* K + \alpha I)^{-1}$ is bounded.

The regularized normal equation must be solved using the available data y^δ and hence the choice of the regularization parameter, in terms of the available data, is a matter of considerable importance. One general method for choosing the regularization parameter is known as Morozov's discrepancy principle. According to this principle, if the signal-to-noise ratio is greater than one, i.e., $\|y^\delta\| > \delta \ge \|y - y^\delta\|$ and $x_\alpha^\delta = (K^* K + \alpha I)^{-1} K^* y^\delta$, then the equation

$$\left\| K x_\alpha^\delta - y^\delta \right\| = \delta$$

has a unique positive solution $\alpha = \alpha(\delta)$ and $x_{\alpha(\delta)}^\delta \to K^\dagger y$ as $\delta \to 0$.

E. Heisenberg's Principle

Heisenberg's uncertainty principle in quantum mechanics is a consequence of an inequality involving unbounded self-adjoint operators. We will consider only a very simple case. Suppose a particle moves in one dimension, its position at a given time t denoted by $x \in (-\infty, \infty)$. In the quantum mechanical formalism the position of the particle is understood to be a random variable and it is the *state*, rather than the position, that is at issue. The state is a function ψ which is a unit vector in the complex Hilbert space $H = L^2(-\infty, \infty)$; the interpretation being that

$$\int_a^b \psi(s) \overline{\psi(s)}\, ds$$

represents the probability that at the time in question the particle is positioned between a and b. In other words, the integrable real-valued function $\psi \bar{\psi}$ is a probability density for the position. In general, a unit vector in H is called a *state*. A self-adjoint linear operator $T : \mathcal{D}(T) \subseteq H \to H$ is called an *observable*. The *expected value* of an observable T when the system is in state ψ is

$$E(T) = \langle T\psi, \psi \rangle = \int_{-\infty}^{\infty} T\psi(s) \overline{\psi(s)}\, ds$$

For example, the expected value of the multiplication by the independent variable operator, $(M\psi)(x) = x\psi(x)$,

$$E(M) = \int_{-\infty}^{\infty} x\psi(x)\overline{\psi(x)}\, dx$$

is the mean position of the particle (slight modifications of the arguments given in the section on unbounded operators show that M is unbounded and self-adjoint). For this reason the observable M is called the *position* operator. The *variance* of an observable T when the system is in state ψ is defined by

$$Var(T) = \|(T - E(T)I)\psi\|^2.$$

That is, the variance gives a measure of the dispersion of an observable from its expected value.

Definition. The *commutator* of two observables S and T is the observable $[S, T] : \mathcal{D}(ST) \cap \mathcal{D}(TS) \to H$ defined by $[S, T] = ST - TS$.

S and T are said to *commute* if $ST = TS$, i.e., $\mathcal{D}(ST) = \mathcal{D}(TS)$ and $[S, T] = 0$. In abstract form the Heisenberg principle says that for any state $\psi \in \mathcal{D}([S, T])$

$$\frac{1}{4}|E([S, T])|^2 \leq Var(S) \times Var(T).$$

Suppose $\mathcal{D}(\mathcal{P})$ is the subspace of all absolutely continuous functions in H whose first derivative is also in H. Define the operator $\mathcal{P} : \mathcal{D}(\mathcal{P}) \to H$ by

$$\mathcal{P}\psi(x) = \frac{h}{2\pi i}\psi'(x)$$

where h is *Planck's constant*. Then $\mathcal{P}$ is self-adjoint, that is, an observable. A physical argument shows that $E(\mathcal{P})$ is the expected value of the momentum of the system and hence $\mathcal{P}$ is called the *momentum operator*. The commutator of the momentum and position operators can be found from the relation

$$(\mathcal{P}M\psi)(x) = \frac{h}{2\pi i}\frac{d}{dx}[x\psi(x)] = \frac{h}{2\pi i}[\psi(x) + x\psi'(x)]$$
$$= \frac{h}{2\pi i}\psi(x) + (M\mathcal{P}\psi)(x)$$

i.e., for all $\psi \in \mathcal{D}([\mathcal{P}, M])$:

$$[\mathcal{P}, M]\psi = \frac{h}{2\pi i}\psi.$$

The general Heisenberg principle then gives

$$Var(\mathcal{P}) \times Var(M) \geq \left(\frac{h}{4\pi}\right)^2.$$

This is an expression of the physical uncertainty principle: no matter the state ψ, the position and momentum can not both be determined with arbitrary certainty.

SEE ALSO THE FOLLOWING ARTICLES

CONVEX SETS • DATA MINING AND KNOWLEDGE DISCOVERY • DIFFERENTIAL EQUATIONS, ORDINARY • FOURIER SERIES • GENERALIZED FUNCTIONS • TOPOLOGY, GENERAL

BIBLIOGRAPHY

Brenner, S. C., and Scott, L. R. (1994). "The Mathematical Theory of Finite Element Methods," Springer-Verlag, New York.

Groetsch, C. W. (1980). "Elements of Applicable Functional Analysis," Dekker, New York.

Kantorovich, L. V., and Akilov, G. P. (1964). "Functional Analysis in Normed Spaces," Pergamon, New York.

Kirsch, A. (1996). "An Introduction to the Mathematical Theory of Inverse Problems," Springer-Verlag, New York.

Kreyszig, E. (1978). "Introductory Functional Analysis with Applications," Wiley, New York.

Lebedev, L. P., Vorovich, I. I., and Gladwell, G. M. L. (1996). "Functional Analysis: Applications in Mechanics and Inverse Problems," Kluwer, Dordrecht.

Naylor, A. W., and Sell, G. R. (1971). "Linear Operator Theory in Engineering and Science," Holt, Rinehart and Winston, New York.

Riesz, F., and Sz.-Nagy, B. (1955). "Functional Analysis," Ungar, New York.



Generalized Functions

Ram P. Kanwal
Pennsylvania State University

I. Introduction
II. Distributions
III. Algebraic Operations on Distributions
IV. Analytic Operations on Distributions
V. Pseudo-Function, Hadamard Finite Part and Regularization
VI. Distributional Derivatives OF DiscontinuousFunctions
VII. Convergence of Distributions and Fourier Series
VIII. Direct Product and Convolution of Distributions
IX. Fourier Transform
X. Poisson Summation Formula
XI. Asymptotic Evaluation of Integrals: A Distributional Approach

GLOSSARY

Convolution The convolution $f * g$ of two functions $f(x)$ and $g(x)$ is defined as $\int_{-\infty}^{\infty} f(y)\, g(x-y)\, dy$. This concept carries over to the generalized functions also.

Dirac delta function $\delta(x)$ This function is intuitively defined to be zero when $x \neq 0$ and infinite at $x = 0$, in such a way that the area under it is unity.

Distribution A linear continuous functional.

Distributional (generalized) derivatives Nonclassical derivatives of the generalized functions which do not have classical derivatives at their singularities.

Functional A rule which assigns a number $\langle t, \phi \rangle$ to a test function $\phi(x)$ through a generalized function $t(x)$.

Hadamard finite part Finite difference of two infinite terms for defining divergent integrals.

Impulse response The particular solution of a differential equation with $\delta(x)$ (impulse) as its forcing term.

Pseudo-function Singular functions such as $1/x^k$ which are regularized at their singularity with the help of concepts such as the Hadamard finite part.

Regular distributions Distributions based on locally integrable functions.

Singular distributions (generalized functions) Distributions which are not regular.

Test function space A space of suitable smooth functions $\phi(x)$ which is instrumental in defining a generalized function such as $\delta(x)$ as a distribution.

GENERALIZED FUNCTIONS are objects, such as the Dirac delta function, with such inherent singularities that they cannot be integrated or differentiated in the classical sense. By defining them as functionals (distributions) which carry smooth functions to numbers, we can overcome the difficulty. Then they possess remarkable properties that extend the capabilities of the classical mathematics. Indeed, the techniques based on the generalized functions not only solve the classical problems in a simple fashion but also produce many new concepts. Accordingly, they have influenced many topics in mathematical and physical sciences.

I. INTRODUCTION

Functions such as Dirac's delta function and Heaviside function have been used by scientists even though the former is not a function and the latter is not a differentiable function in the classical sense. The theory of distributions has provided us not only with mathematical foundations for these functions but also for various other nonclassical functions. The rudiments of the theory of distributions can be found in the concepts of the Hadamard finite part of the divergent integrals and sequences which in the limit behave like delta function. However, it was Sobolev and Schwartz who established the modern theory of distributions.

The Dirac delta function $\delta(x-\xi)$, also called the impulse function, is usually defined as a function which is zero everywhere except at $x=\xi$, where it has a spike such that $\int_{-\infty}^{\infty}\delta(x-\xi)\,dx=1$. More generally, it is defined by its sifting property,

$$\int_{-\infty}^{\infty} f(x)\delta(x-\xi)\,dx = f(\xi), \tag{1}$$

for all continuous functions $f(x)$. However, for any acceptable definition of integration there can be no such function. This difficulty was subsequently overcome by two approaches. The first is to define the delta function as the limit of delta sequences, while the second is to define it as a distribution. The reasoning behind a delta sequence is that although the delta function cannot be justified mathematically, there are sequences $\{s_m(x)\}$ which in the limit $m\to\infty$ satisfy the relation (1). An interesting sequence is

$$s_m(x) = \frac{1}{\pi}\frac{m}{1+m^2x^2}. \tag{2}$$

It is instructive to imagine (2) as a continuous charge distribution on a line, so that the total charge $s_m(x)$ to the left of x is $r_m(x)=\int_{-\infty}^{x}s_m(u)\,du=\frac{1}{2}+(1/\pi)\tan^{-1}mx$. Because $\int_{-\infty}^{\infty}s_m(x)\,dx=1$, the total charge on the line is equal to unity. Furthermore, it can be readily proved that

$$\lim_{m\to\infty}\int_{-\infty}^{\infty} f(x)s_m(x)\,dx = f(0).$$

Accordingly, it satisfies the sifting property (1).

Defining $\delta(x)$ as a distribution is much more convenient and useful. Moreover, the theory of distributions helps us define many more functions which are singular in nature.

II. DISTRIBUTIONS

To make $\delta(x)$ meaningful, we appeal to various spaces of smooth functions. We start with the real-valued functions $\phi(\boldsymbol{x})$, where $\boldsymbol{x}=(x_1,\ldots,x_n)\in\mathbb{R}^n$, the n-dimensional Euclidean space. To describe various properties of these functions we introduce the multi-index notation. Let $\boldsymbol{k}$ be an n-tuple of non-negative integers: $\boldsymbol{k}=(k_1\ldots,k_n)$. Then we define

$$\begin{aligned}
&|\boldsymbol{k}| = k_1+\cdots+k_n; \qquad \boldsymbol{x}^{\boldsymbol{k}} = x_1^{k_1}\cdots x_n^{k_n};\\
&\boldsymbol{k}! = k_1!\cdots k_n!;\\
&\binom{\boldsymbol{k}}{\boldsymbol{m}} = \frac{\boldsymbol{k}!}{\boldsymbol{m}!(\boldsymbol{k}-\boldsymbol{m})!}; \qquad D_j = \frac{\partial}{\partial x_j};\\
&\boldsymbol{D}^{\boldsymbol{k}} = \frac{\partial^{|\boldsymbol{k}|}}{\partial x_1^{k_1}\cdots\partial x_n^{k_n}} = \frac{\partial^{k_1+\cdots+k_n}}{\partial x_1^{k_1}\cdots\partial x_n^{k_n}} = D_1^{k_1}\ldots D_n^{k_n}.
\end{aligned} \tag{3}$$

We are now ready to state the properties of the functions $\phi(\boldsymbol{x})$. They are (1) $D^{\boldsymbol{k}}\phi(\boldsymbol{x})$ exists for all multi-indices $\boldsymbol{k}$. (2) There exists a number A such that $\phi(\boldsymbol{x})$ vanishes for $r>A$, where r is the radial distance $r=(x_1^2+\cdots+x_n^2)^{1/2}$. This means that $\phi(\boldsymbol{x})$ has a compact support. These two properties are written symbolically as $\phi(\boldsymbol{x})\in C_0^{\infty}$, where the superscript stands for infinite differentiability, the subscript for the compact support. This space of functions is denoted by $\mathcal{D}$, and $\phi(\boldsymbol{x})$ are called the test functions. The prototype example in $\mathbb{R}$ is

$$\phi(x) = \begin{cases} \exp\left(-\dfrac{a^2}{a^2-x^2}\right), & |x|<a,\\ 0, & |x|>a. \end{cases} \tag{4}$$

Note that the definition of $\mathcal{D}$ does not demand that all $\phi(\boldsymbol{x})$ have the same support. Furthermore, we observe that (a) $\mathcal{D}$ is a linear (vector) space, because if ϕ_1 and ϕ_2 are in $\mathcal{D}$, then so is $c_1\phi_1+c_2\phi_2$ for arbitrary real numbers. (b) If $\phi\in\mathcal{D}$, then so is $D^{\boldsymbol{k}}\phi$. (c) For a C^{∞} function of $f(\boldsymbol{x})$ and a $\phi(\boldsymbol{x})\in\mathcal{D}$, $f\phi\in\mathcal{D}$. (d) If $\phi(x_1,\ldots,x_m)$ is an m-dimensional test function and $\psi(x_{m+1,\ldots,x_n})$ is an $(n-m)$-dimensional test function, then $\phi\psi$ is an n-dimensional test function in the variables $x_1,\ldots,x_n$.

With the following four definitions we are able to grasp the concept of distributions:

1. A sequence $\{\phi_m(\boldsymbol{x})\}$, $m = 1, 2, \ldots$, for all $\phi_m \in \mathcal{D}$, converges to ϕ_0 if the following two conditions are satisfied. (a) All ϕ_m as well as ϕ_0 vanish outside a common region. (b) $D^{\boldsymbol{k}}\phi_m \to \mathcal{D}^{\boldsymbol{k}}\phi_0$ uniformly over $\mathbb{R}^n$ as $m \to \infty$, for all multi-indices $\boldsymbol{k}$. For the special case $\phi_0 = 0$, the sequence $\{\phi_m\}$ is called a null sequence.
2. A linear functional $t(\boldsymbol{x})$ on the space $\mathcal{D}$ is an operation by which we assign to every test function $\phi(\boldsymbol{x})$ a real number $\langle t(\boldsymbol{x}), \phi(\boldsymbol{x})\rangle$ such that $\langle t, c_1\phi_1 + c_2\phi_2\rangle = c_1\langle t, \phi_1\rangle + c_2\langle t, \phi_2\rangle$ for arbitrary test functions $\phi_1(\boldsymbol{x})$ and $\phi_2(\boldsymbol{x})$ and real numbers c_1 and c_2. Then it follows that $\langle t, \sum_{j=1}^m c_j\langle\phi_j\rangle = \sum_{j=1}^m c_j\langle t, \phi_j\rangle$, where c_j are arbitrary real numbers.
3. A linear functional $t(\boldsymbol{x})$ on $\mathcal{D}$ is called *continuous* if and only if the sequence of numbers $\langle t, \phi_m\rangle \to \langle t, \phi\rangle$, as $m\to\infty$, when the sequence $\{\phi_m\}$ of test functions converges to $\phi \in \mathcal{D}$, that is,
$$\lim_{m\to\infty}\langle t, \phi_m\rangle = \langle t, \lim_{m\to\infty}\phi_m\rangle.$$
4. A continuous linear functional on the space $\mathcal{D}$ is called a distribution.

The set of distributions that is most useful is the ones generated by locally integrable functions $f(x)$ (that is, $\int_\Omega [f(\boldsymbol{x})]\,d\boldsymbol{x}$ exists for every bounded region Ω of $\mathbb{R}^n$). Indeed, every locally integrable function $f(\boldsymbol{x})$ generates a distribution through the relation $\langle f, \phi\rangle = \int_\Omega f(\boldsymbol{x})\phi(\boldsymbol{x})\,d\boldsymbol{x}$. The linearity and continuity of this functional can be readily proved. The distributions produced by locally integrable functions are called *regular distributions* or distributions of order zero. All the other ones are called *singular distributions*. We present two examples in the one-dimensional space $\mathbb{R}$.

1. Heaviside distribution:
$$H(x) = \begin{cases} 0, & x < 0, \\ 1, & x > 0. \end{cases}$$
The Heaviside distribution defines the regular distribution $\langle H, \phi\rangle = \int_0^\infty \phi(x)\,dx$. It is clearly a linear functional. It is also continuous because $\lim_{m\to\infty}\langle H(x), \phi_m(x)\rangle = \lim_{m\to\infty}\int_0^\infty \phi_m(x) = \int_0^\infty \phi(x)\,dx$, where $\phi(x)$ is the limit of the sequence $\{\phi_m(x)\}$ as $m \to \infty$. Thus,
$$\lim_{m\to\infty}\langle H(x), \phi_m(x)\rangle = \langle H(x), \lim_{m\to\infty}\phi_m(x)\rangle.$$

2. The Dirac delta function $\delta(x-\xi)$. By the sifting property (1) we have $\langle\delta(x-\xi), \phi(x)\rangle = \int_{-\infty}^\infty \delta(x-\xi)\times\phi(x)\,dx = \phi(\xi)$, a number. The linearity also follows from the sifting property. This functional is continuous because $\lim_{m\to\infty}\langle\delta(x-\xi), \phi_m(x)\rangle = \lim_{m\to\infty}\phi_m(\xi) = \langle\delta(x-\xi), \lim_{m\to\infty}\phi_m(x)\rangle$. Since $\delta(x-\xi)$ is not a locally integrable function, it produces a singular distribution. These results hold in $\mathbb{R}^n$ as well. The functions which generate singular distributions are called generalized functions. We shall use these words interchangeably. The definition of a distribution can be extended to include complex-valued functions.

The dual space $\mathcal{D}'$. The space of all distributions on $\mathcal{D}$ is called the *dual space* of $\mathcal{D}$ and is denoted as $\mathcal{D}'$. It is also a linear space.

There are many other interesting spaces. For example, the test function space $\mathcal{E}$ consists of all the functions $\phi(\boldsymbol{x})$, so there is no limit on their growth at infinity. Accordingly, $\mathcal{E} \supset \mathcal{D}$. The corresponding space $\mathcal{E}'$ consists of the distributions which have compact support, so that $\mathcal{E}' \supset \mathcal{D}'$. The distributions of slow growth are defined in Sect. IX.

III. ALGEBRAIC OPERATIONS ON DISTRIBUTIONS

Let $\langle t(x), \phi(x)\rangle$ be a regular distribution generated by a locally integrable function $t(x) \times \in \mathbb{R}$. Let $x = ay - b$, where a and b are constants. Then we have
$$\begin{aligned}\langle t(ay-b), \phi(y)\rangle &= \int_{-\infty}^{\infty} t(ay-b)\phi(y)\,dy \\ &= \frac{1}{|a|}\int_{-\infty}^{\infty} t(x)\phi\left(\frac{x+b}{a}\right)dx \\ &= \frac{1}{|a|}\left\langle t(x), \phi\left(\frac{x+b}{a}\right)\right\rangle. \qquad (5)\end{aligned}$$
For the special case $a = 1$, relation (5) becomes $\langle t(y-b), \phi(y)\rangle = \langle t(x), \phi(x+b)\rangle$. As another special case, (5) yields $\langle\delta(-y), \phi(y)\rangle\langle\delta(x), \phi(-x)\rangle = \phi(0)$. Thus $\delta(x)$ is an even function.

A distribution is called homogeneous of degree λ if $t(ax) = a^\lambda t(x)$, $a > 0$. In view of relation (5) with $b = 0$, we find that for a homogeneous distribution we have $\langle t(ay), \phi(y)\rangle = (1/a)\langle t(x), \phi(x/a)\rangle = a^\lambda\langle t(x), \phi(y)\rangle$, so that $\langle t(x), \phi(x/a)\rangle = a^{\lambda+1}\langle t(x), \phi(x)\rangle$.

All these relations hold also for distributions $t(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$. In that case we set $\boldsymbol{x} = \boldsymbol{A}y - \boldsymbol{B}$, where $\boldsymbol{A}$ is a nonsingular $n \times n$ matrix and $\boldsymbol{B}$ is a constant vector. Then we have $\langle t(\boldsymbol{A}y - \boldsymbol{B}), \phi(y)\rangle = (1/\det \boldsymbol{A})\langle t(\boldsymbol{x}), \phi[\boldsymbol{A}^{-1}(\boldsymbol{x} + \boldsymbol{B})]\rangle$, where $\boldsymbol{A}^{-1}$ is the inverse of the matrix $\boldsymbol{A}$.

Product of a distribution and a fucntion: In general it is difficult to define the product of two distributions,

even for two regular distributions. Indeed, if we take $f(x)=g(x)=1/\sqrt{x}$, which is a regular distribution, then their product $1/x$ is not a locally integrable function in the neighborhood of $x=0$ and as such does not define a regular distribution (we shall discuss the function $1/x$ in Sect. V). However, we can multiply a distribution $t(\boldsymbol{x})$ with a C^∞ function $\psi(x)$ so that we have $\langle \psi t, \phi\rangle = \langle t, \psi\phi\rangle$. Because $\psi\phi \in \mathcal{D}$, it follows that ψt is a distribution.

IV. ANALYTIC OPERATIONS ON DISTRIBUTIONS

Let us start with a regular distribution generated by C^1 function $t(x)$, $x\in\mathbb{R}$, so that $\langle t(x), \phi(x)\rangle = \int_{-\infty}^{\infty} t(x)\phi(x)\,dx$. When we integrate the quantity $\int_{-\infty}^{\infty} t'(x),\ \phi(x)\,dx$ by parts, we get $\int_{-\infty}^{\infty} t'(x) \times \phi(x)\,dx = -\int_{-\infty}^{\infty} t(x)\phi'(x)\,dx$, where we have used the fact that $\phi(x)$ has a compact support. Thus, $\langle t'(x), \phi(x)\rangle = -\langle t(x),\ \phi'(x)\rangle$. Because $\phi'(x)$ is also in $\mathcal{D}$, this relation helps us in defining the distributional derivative $t'(x)$ of a distribution $t(x)$ (regular or singular) as above. Continuing this process, we have $\langle t^{(n)}(x), \phi(x)\rangle = (-1)^n\langle t(x), \phi^{(n)}(x)\rangle$, where the superscript (n) stands for nth-order differentiation. The corresponding formula in $\mathbb{R}^n$ is

$$\langle D^{\boldsymbol{k}} t(\boldsymbol{x}), \phi(\boldsymbol{x})\rangle = (-1)^{\boldsymbol{k}}\langle t(\boldsymbol{x}), \boldsymbol{D}^{\boldsymbol{k}}\phi(\boldsymbol{x})\rangle. \tag{6}$$

Thus a generalized function is infinitely differentiable. This result has tremendous ramifications, as we shall soon discover.

The primitive of a distribution $t(x)$ is a solution of $ds(x)/dx = t(x)$. This means that we seek $s(x)\in\mathcal{D}'$ such that $\langle s, \phi'\rangle = -\langle t, \phi\rangle$, $\phi\in\mathcal{D}$.

Let us illustrate the concept of distributional differentiation with the help of a few applications.

1. Recall that the Heaviside function is defined as a distribution by the relation (6) $\langle H(x), \phi(x)\rangle = \int_0^\infty \phi(x)\,dx$. According to definition (6), we have $\langle H'(x), \phi(x)\rangle = -\langle H(x), \phi'(x)\rangle = -\int_0^\infty \phi'(x) = \phi(0) = \langle \delta(x), \phi(x)\rangle$, where we have use the fact that $\phi(x)$ vanishes at ∞. Thus,

$$\frac{dH}{dx} = \delta(x). \tag{7}$$

Because $H(x)$ is a distribution of order zero, $\delta(x)$ is a distribution of order 1. We can continue this process and find that $\langle \delta^{(n)}(x), \phi(x)\rangle = (-1)^n\langle \delta(x), \phi^{(n)}(x)\rangle$. Just as $\delta(x)$ stands for an impulse or a pole at $x=0$, $\delta'(x)$ stands for a dipole and $\delta^{(n)}(x)$ is a multiple of order $(n+1)$ as well as a distribution of order $(n+1)$.

2. *Impulse response*. With the help of the formula (7) we can find the derivative of the signum function sgn x, which is 1 for $x>0$ and -1 for $x<0$. Thus, sgn $x = 2H(x)-1$. Then it follows from (7) that $(\text{sgn } x)' = 2\delta(x)$. This result, in turn, helps us in finding the derivative of the function $|x|$. Indeed, $\langle |x|, \phi(x)\rangle = \int_{-\infty}^{\infty} |x|\phi(x)\,dx = \int_0^\infty x\phi(x)\,dx - \int_{-\infty}^0 x\phi(x)\,dx$. Thus, $\langle |x|', \phi(x)\rangle = -\langle |x|, \phi'(x)\rangle = -\int_0^\infty x\phi'(x)\,dx + \int_\infty^0 x\phi'\times(x)\,dx$, which, when integrated by parts, yields the formula $\langle |x|', \phi(x)\rangle = \langle \text{sgn } x, \phi(x)\rangle$. Thus, $|x|' = \text{sgn } x$. The second differentiation yields $(d^2/dx^2)(|x|) = 2\delta(x)$.

3. A distribution $E(\boldsymbol{x})$ is said to be a fundamental solution or a free-space Green's function or an impulse response if it satisfies the relation $L\,E(\boldsymbol{x}) = \delta(\boldsymbol{x})$, where L is a differential operator. Then from the previous paragraph we find that $\frac{1}{2}|x|$ is the impulse response for the differential operator $L = (d^2/dx^2)$.

Relation (6) for differentiation helps us in deriving many interesting formulas such as

$$\begin{aligned} t(x)\delta^{(n)}(x) = {} & (-1)^n t^{(n)}(0)\delta(x) \\ & + (-1)^{(n-1)} t^{(n-1)}(0)\delta'(x) \\ & + (-1)^{n-2}\frac{n(n-1)}{2!} t^{(n-2)}(0)\delta''(x) + \cdots \\ & + f(0)\delta^{(n)}(x) \end{aligned} \tag{8}$$

for a distribution $t(x)$. The proof follows by evaluating the quantity $\langle t(x)\delta^{(n)}(x), \phi(x)\rangle = \langle \delta^{(n)}(x), t(x)\phi(x)\rangle = (-1)^n\,\langle\, \delta(x), (t(x)\phi(x))^{(n)}\rangle$. For $n=0$ and 1, formula (8) becomes

$$\begin{aligned} t(x)\delta(x) &= t(0)\delta(x); \\ t(x)\delta'(x) &= -t'(0)\delta(x) + t(0)\delta'(x). \end{aligned} \tag{9}$$

One of the most important consequences of (8) is

$$x^m\delta^{(n)}(x) = \begin{cases} (-1)^m\dfrac{n!}{(n-m)!}\delta^{(n-m)}(x), & n \ge m, \\ 0, & n < m. \end{cases} \tag{10}$$

4. Next, we attempt to evaluate $\delta[f(x)]$. Let us first assume that $f(x)$ has a simple zero at x_1 such that $f(x_1)=0$ but $f'(x_1)>0$ [the case $f'(x)_1<0$ follows in a similar fashion]. Thus $f(x)$ increases monotonically in the neighborhood of x_1 so that $H[f(x)] = H(x-x_1)$, where $H(x)$ is the Heaviside function. Then we use (7) and find that $(d/dx)H[f(x)] = \delta(x-x_1)$ or $\delta[f(x)] = |(f'(x_1))|^{-1}\delta(x-x_1)$. If there are n zeros of $f(x)$, then the above result yields

$$\delta[f(x)] = \sum_{m=1}^{n}\frac{\delta(x-x_m)}{f'(x_m)}. \tag{11}$$

This result has many applications.

5. The formula $(\operatorname{sgn} x)' = 2\delta(x)$, derived in Example 2, is also instrumental in deriving an integral representation for the delta function. Indeed, if we appeal to the relations

$$\int_{-\infty}^{\infty} \frac{\sin tx}{x}\,dx = \begin{cases} \pi, & t > 0 \\ -\pi, & t < 0; \end{cases} \qquad \int_{-\infty}^{\infty} \frac{\cos tx}{x} = 0,$$

from calculus, we find that $(1/2\pi)\int_{-\infty}^{\infty}(e^{itx}/ix)\,dx = \frac{1}{2}\operatorname{sgn} t$. When we differentiate this relation with respect to t we get the important formula

$$\delta(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{itx}\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\,dx, \tag{12}$$

where we have used the fact that $\delta(t)$ is an even function. This formula can be generalized to n dimensions if we observe that $\delta(x_1, \ldots, x_n) = \delta(x_1)\cdots\delta(x_n)$. Accordingly, in the four-dimensional space (x_1, x_2, x_3, t), relation (12) yields the planewave expansion of the delta function,

$$\delta(\boldsymbol{x}, t) = \frac{1}{(2\pi)^4}\int_{-\infty}^{\infty} e^{-i(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)} d^3\boldsymbol{k}\,d\omega, \tag{13}$$

where $\boldsymbol{k} = (k_1, k_2, k_3)$ and we have a fourfold integral.

V. PSEUDO-FUNCTION, HADAMARD FINITE PART AND REGULARIZATION

The functions $1/x^m$ and $H(x)/x^m$, where m is an integer, are not locally integrable at $x = 0$ and, as such, do not define distributions. However, we can appeal to the concepts of Cauchy principal value and the Hadamard finite part and define these functions as distributions. The simplest example is the function $1/x$. For a test function $\phi(x)$ the integral $\int[\phi(x)/x]\,dx$ is not absolutely convergent unless $\phi(0) = 0$. However, it has an interpretation as a principal value integral, namely,

$$\left\langle Pv\left(\frac{1}{x}\right), \phi(x)\right\rangle = \lim_{\epsilon\to 0}\int_{|x|>\epsilon} \frac{\phi(x)}{x}\,dx. \tag{14}$$

Writing $\phi(x) = \phi(0) + [\phi(x) - \phi(0)]$, relation (14) takes the form

$$\left\langle Pv\left(\frac{1}{x}\right), \phi(x)\right\rangle = \lim_{\epsilon\to 0}\int_{|x|>\epsilon} \frac{\phi(0)}{x}\,dx + \int_{|x|>\epsilon} \frac{[\phi(x) - \phi(0)]}{x}\,dx. \tag{15}$$

Since the function $1/x$ is odd, the first term on the right side of (15) vanishes. The integrand in the second integral on the right side of (15) approaches $\phi'(0)$ as $x \to 0$. Accordingly, we have

$$\left\langle Pv\left(\frac{1}{x}\right), \phi(x)\right\rangle = \lim_{\epsilon\to 0}\int_{|x|>\epsilon} \frac{\phi(x) - \phi(0)}{x}\,dx.$$

Because $\epsilon > 0$ is arbitrary, the previous relation is also written as

$$\left\langle Pv\left(\frac{1}{x}\right), \phi(x)\right\rangle = \lim_{\epsilon\to 0}\int_{|x|>1} \frac{\phi(x) - \phi(0)}{x}\,dx. \tag{16}$$

In this form the function $Pv(1/x)$ is easily proved to be a linear continuous function. As such, it is a distribution and is written as $Pf(1/x)$, where Pf stands for pseudo-function.

We come across many divergent integrals whose principal values do not exist. Consider, for instance, the integral $\int_a^b dx/x^2$, $a < 0 < b$. Because

$$\int_{[a,b]/[-\epsilon,\epsilon]} \frac{dx}{x^2} = \lim_{\epsilon\to 0}\left(-\frac{1}{b} + \frac{1}{a} + \frac{2}{\epsilon}\right), \tag{17}$$

we cannot get the principal value. In these situations the concept of the Hadamard finite part becomes helpful. Indeed, in relation (17), the value $(-1/b + 1/a)$ is the finite part and $(2/\epsilon)$ is the infinite part. Thus, we write

$$Fp\int_a^b \frac{dx}{x^2} = -\frac{1}{b} + \frac{1}{a}. \tag{18}$$

As another example, we consider the integral $\int_0^1 dx/x^\alpha$. It is divergent if $\alpha \geq 1$. Indeed, for $\alpha > 1$, we obtain

$$\int_\epsilon^1 \frac{dx}{x^\alpha} = \frac{1}{1-\alpha} - \frac{\epsilon^{1-\alpha}}{1-\alpha} \tag{19}$$

so that we have

$$Fp\int_0^1 \frac{dx}{x^\alpha} = \frac{1}{1-\alpha}, \qquad \alpha > 1. \tag{20}$$

When $\alpha = 1$, $\int_\epsilon^1 dx/x = -\ln\epsilon$, which is infinite, so that

$$Fp\int_0^1 \frac{dx}{x} = 0. \tag{21}$$

The process of finding the principal value and the finite part of divergent integrals is called the regularization of these integrals. When both of them exist, they are equal.

Let us use the definition of the Hadamard finite part to regularize the function $H(x)/x$. The action of $H(x)/x$ on a test function $\phi(x)$ is $\int_0^\infty(\phi(x)/x)\,dx$. Thus, we consider

$$\int_\epsilon^\infty \frac{\phi(x)}{x}\,dx = \int_\epsilon^1 \frac{\phi(x)}{x}\,dx + \int_1^\infty \phi(x)\,dx. \tag{22}$$

In the numerator of the first term on the right side of this equation we add and subtract $\phi(0)$ to get

$$\int_\epsilon^\infty \frac{\phi(x)}{x}\,dx = \int_\epsilon^1 \frac{\phi(0)}{x}\,dx + \int_\epsilon^1 \frac{\phi(x)-\phi(x)}{x} + \int_1^\infty \phi(x)\,dx$$

$$= -\phi(0)\ln\epsilon + \int_\epsilon^1 \frac{\phi(x)-\phi(0)}{x}\,dx + \int_1^\infty \frac{\phi(x)}{x}\,dx.$$

The first term on the right side of this equation is infinite as $\epsilon \to 0$, while the other two terms are finite. Accordingly, we define

$$\int \frac{H(x)\phi(x)}{x}\,dx = \int_\epsilon^1 \frac{\phi(x)-\phi(0)}{x}\,dx + \int_1^\infty \frac{\phi(x)}{x}\,dx. \tag{23}$$

The function $Pf(1/|x|)$ can be regularized in the same way. Indeed,

$$\left\langle Pf\left(\frac{1}{|x|}\right), \phi(x)\right\rangle = \int_{|x|\le 1} \frac{\phi(x)-\phi(0)}{|x|}\,dx + \int_{|x|>1} \frac{\phi(x)}{|x|}\,dx. \tag{24}$$

These concepts can be used to regularize the functions $1/x^m$ and $H(x)/x^m$ to yield the generalized functions $Pf(1/x^m)$ and $Pf(H(x)/x^m)$. The combination of $Pf(1/x^m)$ and $\delta^{(m)}(x)$ yield the Heisenberg distributions

$$\delta^{\pm(m)}(x) = \frac{1}{2}\delta^{(m)}(x) \mp \frac{1}{2\pi i}\,Pf\left(\frac{1}{x^m}\right).$$

They arise in quantum mechanics.

Let us end this section by solving the equation $xt(x) = g(x)$. The homogeneous part $xt(x) = 0$ has the solution $t(x) = \delta(x)$ because $x\delta(x) = 0$. Thus, the complete solution is

$$t(x) = \delta(x) + g(x)Pf\left(\frac{1}{x}\right). \tag{25}$$

VI. DISTRIBUTIONAL DERIVATIVES OF DISCONTINUOUSFUNCTIONS

Let us start with $x \in \mathbb{R}$ and consider a function $F(x)$ that has a jump discontinuity at $x = \xi$ of magnitude a but that has the drivative $F'(x)$ in the intervals $x < \xi$ and $x > \xi$. The derivative is undefined at $x = \xi$. To find the distributional derivative of $F(x)$ in the entire interval we define the function $f(x) = F(x) - aH(x-\xi)$, where H is the Heaviside function. This function is continuous at $x = \xi$ and has a derivative which coincides with $F'(x)$ on both sides of ξ. Accordingly, we differentiate both sides of this equation and get $F'(x) = \bar{F}'(x) - a\delta(x-\xi)$, where the bar over F stands for the generalized (distributional) derivative of F. Thus,

$$\bar{F}'(x) = F'(x) + [F]\delta(x-\xi), \tag{26}$$

where $[F] = a$ is the value of the jump of F at $x = \xi$. Before we present the corresponding n-dimensional theory, we give an interesting application of (26) to the Sturm-Liouville differential equation,

$$\frac{d}{dx}\left[p(x)\frac{dE(x;\xi)}{dx}\right] = q(x)E(x,\xi) - \delta(x-\xi), \tag{27}$$

where $E(x,\xi)$ stands for the impulse response and $p(x)$ and $q(x)$ are continuous at $x = \xi$. When we compare (22) and (23) and use the continuity of $p(x)$ at $x=\xi$, we find that the jump of $[dE/dx]_{x=\xi}$ is

$$[dE/dx]_{x=\xi} = -1/p(\xi).$$

These concepts easily generalize to higher-order derivatives and to the surfaces of discontinuity and, therefore, have applications in the theory of wave fronts. Accordingly, we include time t in our discussion and consider a function $F(\boldsymbol{x}, t)$, $\boldsymbol{x} \in \mathbb{R}^n$, which has a jump discontinuity across a moving surface $\Sigma(\boldsymbol{x}, t)$. Such a surface can be represented locally either as an implicit equation of the form $u(x_1, \ldots, x_n, t) = 0$, or in terms of the curvilinear Gaussian coordinates $v_1, \ldots, v_{n-1}$ on the surface: $x_i = x_i(v_1, \ldots, v_{n-1}, t)$. The surface Σ is regular, so the above-mentioned functions have derivatives of all orders with respect to each of their arguments, and for all values of t, the corresponding Jacobian matrices of transformation have appropriate ranks, that is, grad $u \neq 0$, and the rank of the matrix $(\partial x_i/\partial v_j) = n-1$. Furthermore, Σ divides the space into two parts, which we shall call positive and negative.

The basic distribution concentrated on a moving and a deforming surface $\Sigma(\boldsymbol{x}, t)$ is the delta function $\delta[\Sigma(\boldsymbol{x}, t)]$ whose action on a test function $\phi(\boldsymbol{x}, t) \in \mathcal{D}$ is

$$\langle \delta(\Sigma), \phi\rangle = \int_{-\infty}^{\infty} \int_\Sigma \phi(\boldsymbol{x}, t)\,dS(\boldsymbol{x})\,dt, \tag{28}$$

where dS is the surface element on Σ. This is a simple layer. The second surface distribution is the normal derivative operator, given as

$$\langle d_n\delta(\Sigma),\phi\rangle = -\int_{-\infty}^{\infty}\int_{\Sigma}\frac{d\phi}{dn}\,dS(\boldsymbol{x})\,dt, \tag{29}$$

where $d\phi/dn$ stands for the differentiation along the normal to the surface. This is a dipole layer. Another surface distribution that we need in our discussion is $\delta'(\Sigma)$, defined as

$$\delta'(\Sigma) = n_i\frac{\bar\partial}{\partial x_i}[\delta(\Sigma)], \tag{30}$$

where n_i are the components of the unit normal vector $\boldsymbol{n}$ to the surface. These three distributions are connected by the relation

$$d_n\delta(\Sigma) = \delta'(\Sigma) - 2\Omega\delta(\Sigma), \tag{31}$$

where Ω is the mean curvature of Σ.

Let a function $F(\boldsymbol{x},t)$ have a jump discontinuity across $\Sigma(\boldsymbol{x},t)$. Then the formulas for the time derivative, gradient, and curl of a discontinuous function $F(\boldsymbol{x},t)$ can be derived in a manner similar to (26). Indeed,

$$\frac{\bar\partial \boldsymbol{F}}{\partial t} = \frac{\partial \boldsymbol{F}}{\partial t} - G[\boldsymbol{F}]\delta(\Sigma), \tag{32}$$

$$\overline{\text{grad}}\,\boldsymbol{F} = \text{grad}\,\boldsymbol{F} + \boldsymbol{n}[\boldsymbol{F}]\delta(\Sigma), \tag{33}$$

$$\overline{\text{div}}\,\boldsymbol{F} = \text{div}\,\boldsymbol{F} + \boldsymbol{n}[\boldsymbol{F}]\delta(\Sigma), \tag{34}$$

$$\overline{\text{curl}}\,\boldsymbol{F} = \text{curl}\,\boldsymbol{F} + \boldsymbol{n}\times[\boldsymbol{F}]\delta(\Sigma), \tag{35}$$

where G is the normal speed of the front Σ and $[\boldsymbol{F}] = \boldsymbol{F}_+ - \boldsymbol{F}_-$ is the jump of $\boldsymbol{F}$ across Σ.

Before we extend these concepts, we apply relation (33) to a scalar function F and derive the distributional derivative of $1/r$, where r is the radial distance. In mathematical physics the function $1/r$ is important because it describes the gravitational and the Coulomb potentials. In the modern theories of small particles the singularity of $1/r$ at $r=0$ makes a significant contribution. Accordingly, its distributional derivatives are needed to get the complete picture of the singularity at $r=0$ and we present it as follows. For the sake of computational simplicity, we restrict ourselves to $\mathbb{R}^3$, so that $r=(x_1^2+x_2^2+x_3^3)^{1/2}$. The function $1/r$ corresponds to the $Pf[H(x)/x]$ that we considered in the previous section. Accordingly, we introduce the function $F(\boldsymbol{x}) = H(r-\epsilon)/r$, where $H(r-\epsilon)$ is the Heaviside function, which is unity for $r>\epsilon$ and is zero for $r<\epsilon$. This, in turn, helps us in defining the distribution $1/r$ as

$$\left\langle\frac{1}{r},\phi(\boldsymbol{x})\right\rangle = \lim_{\epsilon\to 0}\int_{-\infty}^{\infty}\frac{1}{r}H(r-\epsilon)\phi(\boldsymbol{x})\,dx. \tag{36}$$

Our aim is to differentiate $1/r$ by taking into account the singularity at $r=0$. For this we appeal to formula (33) and observe that $F(\boldsymbol{x}) = H(r-\epsilon)/r$ so that the jump of this function at the spherical surface of discontinuity $\Sigma: r=\epsilon$, is $1/\epsilon$. Thus, formula (33) becomes

$$\frac{\bar\partial}{\partial x_j}\left(\frac{H(r-\epsilon)}{r}\right) = -\frac{x_j}{r^3}H(r-\epsilon) + \hat r_j\frac{1}{\epsilon}\delta(\Sigma), \tag{37}$$

where $n_j = \hat r_j - x_j/r$. To evaluate the limit of the second term on the right side of (37) we evaluate $\lim_{\epsilon\to 0}\int_\Sigma \phi(\boldsymbol{x})(1/\epsilon)\hat r\,dS = \lim_{\epsilon\to 0}(1/\epsilon)\int_{\Sigma(1)}\phi(\boldsymbol{x})\hat r\,\epsilon^2\,d\omega = 0$, where $\Sigma(1)$ is the unit sphere and ω is the solid angle. Thus $\bar\partial/\partial x_j(1/r) = -x_j/r^3$, which is the same as the classical derivative. In order to compute the second-order distributional derivatives we apply formula (33) to the function $(\partial/\partial x_j)[H(r-\epsilon)/r]$ and get

$$\begin{aligned}\frac{\bar\partial^2}{\partial x_i\,\partial x_j}\left[\frac{H(r-\epsilon)}{r}\right] &= \left[\frac{\partial^2}{\partial x_i\,\partial x_j}\left(\frac{1}{r}\right)\right]H(r-\epsilon)\\ &\quad+\left(\frac{x_i}{r}\right)\left(\frac{-x_j}{r^3}\right)\delta(\Sigma)\\ &= \left(\frac{3x_ix_j-r^2\delta_{ij}}{r^5}\right)H(r-\epsilon)\\ &\quad-\frac{x_ix_j}{r^4}\delta(\Sigma),\end{aligned} \tag{38}$$

where δ_{ij} is the Kronecker delta, which is 1 when $i=j$ and 0 when $i\neq j$. Because $\lim_{\epsilon\to 0}\int_\Sigma(x_ix_j/r^4)\phi(\boldsymbol{x})\,dS = (4\pi/3)\delta_{ij}\phi(\boldsymbol{0})$, relation (38) becomes

$$\frac{\bar\partial^2}{\partial x_i\partial x_j}\left(\frac{1}{r}\right) = \frac{3x_ix_j-r^2\delta_{ij}}{r^5} - \frac{4\pi}{3}\delta_{ij}\delta(\boldsymbol{x}). \tag{39}$$

From this formula we can derive the impulse response for the Laplace operator. Indeed, if we set $i=j$ and sum on j, we get $\nabla^2(1/r) = -4\pi\delta(\boldsymbol{x})$. Thus, the impulse response of $-\nabla^2$ is $1/4\pi r$. We can continue this process and drive the nth-order distributional derivative of the function $1/r^k$.

With the help of the foregoing analysis we can obtain the results that correspond to (32)–(35) for singular surfaces which carry infinite singularities, such as charge sheets and vortex sheets. Suppose that the surface of discontinuity $\Sigma(\boldsymbol{x},t)$ carries a single layer of strength $f(\boldsymbol{x},t)\delta(\Sigma)$. Then

$$\frac{\bar\partial}{\partial t}[f\delta(\Sigma)] = \frac{\bar\delta f}{\delta t}\delta(\Sigma) - Gf\delta'(\Sigma), \tag{40}$$

where $\bar\delta/\delta t = \bar\partial f/\partial t + G\,df/dn$ is the distributional time derivative as apparent to an observer moving with the front Σ. Similarly, $\bar\delta f/\delta x_i = \bar\partial f/\partial x_i - n_i\,df/dn$ is the distributional surface derivative with respect to the Cartesian coordinates of the surrounding space. Thereby,

$$\frac{\bar\partial}{\partial x_i}[f\delta(\Sigma)] = \frac{\bar\delta f}{\delta x_i}\delta(\Sigma) + n_i f\delta'(\Sigma). \tag{41}$$

If we use relation (31) between $d_n\delta(\Sigma)$ and $\delta'(\Sigma)$, we can rewrite formulas (40) and (41) which contain the mean curvature Ω.

VII. CONVERGENCE OF DISTRIBUTIONS AND FOURIER SERIES

A sequence $\{t_m(\boldsymbol{x})\}$ of distributions with $t_m(\boldsymbol{x}) \in \mathcal{D}'$, $m = 1, 2, \ldots$, is said to converge to a distribution $t(\boldsymbol{x}) \in \mathcal{D}'$ if $\lim_{m\to\infty}\langle t_m, \phi\rangle = \langle t, \phi\rangle$ for all $\phi \in \mathcal{D}$. This is called distributional (or weak) convergence. An important consequence is that if the sequence $\{t_m(\boldsymbol{x})\}$ converges to $t(\boldsymbol{x})$, then the sequence $\{D^k t_m\}$ converges to $\{D^k t\}$ because $\lim_{m\to\infty}\langle D^k t_m, \phi\rangle = \lim_{m\to\infty}(-1)^{|k|}\langle t_m, D^k\phi\rangle = (-1)^{|k|}\langle t, D^k\phi\rangle = \langle D^k t, \phi\rangle$. As an example, consider the sequence $\{\cos mx/m\}$, $x \in \mathbb{R}$, which is a sequence of regular distributions and converges to zero pointwise. Then the sequence $\{-\sin mx\}$ which arises by differentiating $\{\cos mx/m\}$, also converges to zero. It is remarkable because the sequence $\{\sin mx\}$ does not have a pointwise limit, as $m \to \infty$, in the classical sense.

A series of distributions $\sum_{p=1}^{\infty} s_p(x)$ converges distributionally to $t(x)$ if the sequence of the partial sums $\{t_m(x)\} = \{\sum_{p=1}^{m} s_p(x)\}$ converges to t distributionally. Thus, we observe from the above analysis that term-by-term differentiation of a convergent series is always possible, provided the resultant series is interpreted in the sense of distributions.

It is known that the Fourier series $\sum_{m=-\infty}^{\infty} c_m e^{imx}$ converges uniformly if for a large m, $|c_m| \le M/m^k$ where m is a constant and k is an integer greater than 2. The series may diverge for other values of k. However, the above analysis assures us that the series converges distributionally for any integer because it can be obtained from the uniformly convergent series $\sum_{m=-\infty}^{\infty}(im)^{k-2}c_m e^{imx}$ by $(k+2)$ successive differentiations. For example, the series $\sum_{m=-\infty}^{\infty} e^{imx}$ has no meaning in the classical sense, but it can be written as $1 + (d^2/dx^2)\{\sum_{m=-\infty}^{\infty}(1/m^2)e^{imx}\}$. Accordingly, the series $\sum_{m=-\infty}^{\infty} e^{imx}$ is distributionally convergent.

VIII. DIRECT PRODUCT AND CONVOLUTION OF DISTRIBUTIONS

The direct product of distributions is defined by using the space of test functions in two variables x and y. Let us denote the direct product of the distributions $s(x) \in \mathcal{D}'(x)$ and $t(y) \in \mathcal{D}'(y)$ as $s(x) \otimes t(y)$. Then

$$\langle s(x) \otimes t(y), \phi(x, y)\rangle = \langle s(x), \langle (t(y), \phi(x, y)\rangle\rangle \quad (42)$$

where $\phi(x, y)$ is a test function in $\mathcal{D}(x, y)$. This makes sense because the function $\psi(x) = \langle t(y), \phi(x, y)\rangle$ is a test function in $\mathcal{D}(x)$ and $D^k\psi(x) = \langle t(y), D_x^k\phi(x, y)\rangle$, where D_x^k implies kth-order derivative with respect to x. Thus, $s(x) \otimes t(y)$ is a functional in $\mathcal{D}'(x, y)$. Indeed, it is a linear and continuous functional. Let us mention some of the properties of the direct product:

1. Linearity: $s \otimes (\alpha t + \beta u) = \alpha s \otimes t + \beta s \otimes u$.
2. Commutativity: $s \otimes t = t \otimes s$.
3. Continuity: if $s_m(x) \to s(x)$, in $\mathcal{D}'(x)$ as $m \to \infty$, then $s_m(x) \otimes t(y) \to s(x) \otimes t(y)$ in $\mathcal{D}'(x, y)$ as $m \to \infty$.
4. Associativity: when $s(x) \in \mathcal{D}'(x)$, $t(y) \in \mathcal{D}'(y)$ and $u(z) \in \mathcal{D}'(z)$, then $s(x) \otimes [t(y) \otimes u(z) = s(x) \otimes t(y)] \otimes u(z)$.
5. Support: $\text{supp}\,(s \times t) = \text{supp}\, s \times \text{supp}\, t$, where $\times$ stands for the Cartesian product.
6. Differentiation: $D_x^k[s(x) \otimes t(y)] = D^k[s(x)] \otimes t(y)$.
7. Translation: $(s \otimes t)(x + h, y) = s(x + h) \otimes t(y)$.

An interesting example of the direct product is $\delta(x) \otimes \delta(y) = \delta(x, y)$, where $\delta(x, y)$ is the two-dimensional delta function. This follows by observing that $\langle \delta(x) \otimes \delta(y), \phi(x, y)\rangle = \langle \delta(x), \langle \delta(y), \phi(x, y)\rangle\rangle = \langle \delta(y), \phi(0, y)\rangle = \phi(0, 0) = \langle \delta(x, y), \phi(x, y)\rangle$.

The *convolution* $s * t$ of two functions $s(x)$ and $t(x)$ is $(s * t)(x) = \int_{-\infty}^{\infty} s(y)t(x - y)\,dy = \int_{-\infty}^{\infty} t(y)s(x - y)\,dy = t * s(x)$. When the functions $s(x)$ and $t(x)$ are locally integrable, then $s(x) * t(x)$ is also locally integrable, and as such, defines a regular distribution.

Next, let us examine $\langle s * t, \phi\rangle$:

$$\begin{aligned}
\langle s * t, \phi\rangle &= \int_{-\infty}^{\infty} (s * t)\phi(z)\,dz \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} s(z - y)t(y)\,dy\right] \phi(z)\,dz \\
&= \int_{-\infty}^{\infty} t(y) \left[\int_{-\infty}^{\infty} s(z - y)\phi(z)\right] dy \\
&= \int_{-\infty}^{\infty} t(y) \int_{-\infty}^{\infty} [s(x)\phi(x + y)\,dx]\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x)t(y)\phi(x + y)\,dx\,dy.
\end{aligned}$$

Thus,

$$\langle s(x) * t(y), \phi(x, y)\rangle = \langle s(x) \otimes t(y), \phi(x + y)\rangle. \quad (43)$$

There is, however, one difficulty with this definition. Even though the function $\phi(x + y)$ is infinitely differentiable, its support is not bounded in the (x, y) plane because x and y may be unboundly large while $x + y$ remains finite. This is remedied by ensuring that the intersection of the supports of $s(x) \otimes t(y)$ and $\phi(x + y)$ is a bounded set. This happens if (1) either s or t has a bounded support or (2) both s and t have support bounded on the same side.

As an example, we find that $(\delta * t)(x), = \int \delta(y)\, t(x-y)\, dx = t(x)$, that is, $\delta(x)$ is the unit element in the convolution algebra. The operation of the convolution has the following properties:

1. Linearity: $s * (\alpha t + \beta u) = \alpha s * t + \beta s * u$.
2. Commutivity: $s * t = t * s$.
3. Associativity: $(s * t) * u = s * (t * u)$.
4. Differentiation: $(D^k s) * t = D^k(s * t) = s * D^k t$.
5. For instance, $[H(x) * t(x)]' = \delta(x) * t(x) = H(x) * t'(x)$.

IX. FOURIER TRANSFORM

Let us write formula (12) as

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{iu(x-\xi)}\, du = \delta(x-\xi),$$

and multiply both sides of this formula by $f(x)$, integrate with respect to x from $-\infty$ to ∞, use the sifting property, and interchange the order of integration. The result is the celebrated Fourier integral theorem,

$$f(\xi) = \frac{1}{2\pi} \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} f(x) e^{iu(x-\xi)}\, dx, \qquad (44)$$

which splits into the pair

$$\hat{f}(u) = \int_{-\infty}^{\infty} f(x) e^{iux}\, dx, \qquad (45)$$

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(u) e^{-iux}\, du, \qquad (46)$$

where we have relabeled ξ with x in (46). The quantity $\hat{f}(u)$ is the Fourier transform of $f(x)$, while formula (46) gives the inverse Fourier transform $F^{-1}[f(u)]$. We shall now examine formula (45) in the context of test functions and distributions. If we attempt to define the Fourier transform of a distribution $t(x)$ as in (45), then we get $\hat{t}(u) = \langle t(x),\ e^{iux} \rangle$, but we are in trouble because e^{iux} is not in $\mathcal{D}$. We could try Parseval's theorem, $\int_{-\infty}^{\infty} \hat{f}(x) g(x)\, dx = \int_{-\infty}^{\infty} f(x) \hat{g}(x)\, dx$, which follows from (45) and (46) and connects the Fourier transforms of two functions. Then we have $\langle \hat{t}, \phi \rangle = \langle t, \hat{\phi} \rangle$, $\phi \in \mathcal{D}$. We again run into trouble because $\hat{\phi}$ may not be in $\mathcal{D}$ even if ϕ is in it. These difficulties are overcome by enlarging the class of test functions.

Test functions of rapid decay. The space $\mathcal{S}$ of test functions of rapid decay contains the complex-valued functions $\phi(\boldsymbol{x})$ with the properties: (1) $\phi(\boldsymbol{x}) \in C^\infty$, (2) $\phi(\boldsymbol{x})$ and its derivatives of all orders vanish at infinity faster than the reciprocal of any polynomial, i.e., $|\boldsymbol{x}^{\boldsymbol{p}} D^{\boldsymbol{k}} \phi(\boldsymbol{x})| < C_{pq}$, where $\boldsymbol{p} = (p_1, \ldots, p_n), \boldsymbol{k} = (k_1, \ldots, k_n)$ are n-tuples and C_{pq} is a constant depending on $\boldsymbol{p}, \boldsymbol{k}$ and $\phi(\boldsymbol{x})$. Another way of saying this is that after multiplication by any polynomial $P(\boldsymbol{x})$, the function $P(\boldsymbol{x})\phi(\boldsymbol{x})$ still tends to zero as $\boldsymbol{x} \to \infty$. Clearly, $S \supset \mathcal{D}$. Convergence in this space is defined as follows. The sequence of functions $\{\phi_m(\boldsymbol{x})\} \to \phi(\boldsymbol{x})$ if and only if $|\boldsymbol{x}^k (D^{\boldsymbol{k}} \phi_m - D^{\boldsymbol{k}} \phi)| < C_{pq}$ for all multi-indices $\boldsymbol{p}$ and $\boldsymbol{k}$ and all m.

Functions of slow growth. A function $f(\boldsymbol{x})$ in $\mathbb{R}^n$ is of slow growth if $f(\boldsymbol{x})$, together with all its derivatives, grows at infinity more slowly than some polynomial, i.e., there exist constants C, m, and A such that $|D^{\boldsymbol{k}} f(\boldsymbol{x})| \langle C|\boldsymbol{x}|^m, |\boldsymbol{x}| \rangle A$.

Tempered distributions. A linear continuous functional $t(\boldsymbol{x})$ over the space S of test functions is called a *distribution of slow growth* or *tempered distribution*. According to each $\phi \in S$, there is assigned a complex number $\langle t, \phi \rangle$ with the properties: (1) $\langle t, c_1\phi_1 + c_2\phi_2 \rangle = c_1\langle t, \phi_1 \rangle + c_2 \langle t, \phi_2 \rangle$; (2) $\lim_{m \to 0} \langle t, \phi_m \rangle = 0$, for every null sequence $\{\phi_m(\boldsymbol{x})\} \in S$. This yields us the dual space S'. If follows from the definitions of convergence in $\mathcal{D}$ and S that a sequence $\{\phi_m(\xi)\}$ that converges in the sense of $\mathcal{D}$ also converges in the sense of S, so that $S' \subset \mathcal{D}'$. Fortunately, most of the distributions in $\mathcal{D}'$ encountered previously are also in S'. However, some locally integrable functions in D' are not in S'. The regular distributions in S' are those generated by functions $f(x)$ of slow growth through the formula $\langle f, \phi \rangle = \int_{\infty}^{\infty} f(x)\phi(x)\, dx$, $\phi \in S$. It is a linear continuous functional.

Fourier transform of the test functions. We shall first give the analysis in $\mathbb{R}$ and then state the corresponding results in $\mathbb{R}^n$. For $\phi(x) \in S$, we can use the definition (45) so that $\hat{\phi}(u) = \int_{\infty}^{\infty} \phi(x) e^{iux}\, dx$. The inverse follows from (46). Since $\hat{\phi}(u)$ is also in S, as can be easily proved, the difficulty encountered earlier has disappeared. Moreover, it follows by the inversion formula (46) that

$$[\hat{\phi}]\hat{}(x) = 2\pi\phi(-x), \qquad (47)$$

which shows that every function $\phi(x) \in S$ is a Fourier transform of some function in S. Indeed, the Fourier transform and its inverse are linear, continuous, and one-to-one mapping of S on to itself.

In order to obtain the transform of tempered distributions, we need some specific formulas of the transform of the test functions $\phi(x)$. Let us list them and prove some of them. The transform of $d^k\phi/dx^k$ is equal to $\int_{-\infty}^{\infty} (d^k\phi/dx^k) e^{iux}\, dx = (-iu)^k \int_{-\infty}^{\infty} \phi(x) e^{iux}\, dx$ so that $[d^k\phi/dx^k]\hat{}(u) = (-iu)^k \hat{\phi}(u)$. Thus, if $P(\lambda)$ is an arbitrary polynomial with constant coefficients, we find that

$[P(d/dx)\phi]\hat{} = P(-iu)\hat{\phi}(u)$. Similarly, from the relation $\int_{-\infty}^{\infty}(ix)^k\phi(x)e^{iux}\,dx = (d^k/du^k)\int_{-\infty}^{\infty}\phi(x)e^{iux}\,dx$, we derive the relation $[x^k\phi]\hat{}(u) = P(-d/du)\hat{\phi}(u)$. Because $\int_{-\infty}^{\infty}\phi(x-a)e^{iux}\,dx = e^{iua}\int_{-\infty}^{\infty}\phi(y)e^{iuy}\,dy$, we have $[\phi(x-a)]\hat{}(u) = e^{iau}\hat{\phi}(u)$. Similarly, $[\phi(x)]\hat{}(u+a) = [e^{ax}(x)]\hat{}(u)$ and $[\phi(ax)\hat{}(u) = (1/|a|\hat{\phi}(u/a))$. The corresponding n-dimensional formulas are listed below.

$$\hat{\phi}(\boldsymbol{u}) = \int_{-\infty}^{\infty} e^{i\boldsymbol{u}\cdot\boldsymbol{x}}\phi(\boldsymbol{x})\,d\boldsymbol{x}, \tag{48}$$

$$\phi(x) = \frac{1}{(2\pi)^n}\int_{-\infty}^{\infty} e^{i\boldsymbol{u}\cdot\boldsymbol{x}}\hat{\phi}(\boldsymbol{u})\,d\boldsymbol{u}, \tag{49}$$

$$[\hat{\phi}]\hat{}(\boldsymbol{u}) = (2\pi)^n\phi(-\boldsymbol{u}), \tag{50}$$

$$[\boldsymbol{D}^{\boldsymbol{k}}\phi]\hat{}(\boldsymbol{u}) = (-i\boldsymbol{u})^{\boldsymbol{k}}\hat{\phi}(\boldsymbol{u}), \tag{51}$$

$$\left[P\left(\frac{\partial}{\partial x_1},\ldots,\frac{\partial}{\partial x_n}\right)\phi(\boldsymbol{x})\right]\hat{}(\boldsymbol{u}) = P(-iu_1,.,-u_n)\hat{\phi}(\boldsymbol{x}), \tag{52}$$

$$[\boldsymbol{x}^k\phi]\hat{}(\boldsymbol{x}) = (-i\boldsymbol{D})^k x\hat{\phi}(u), \tag{53}$$

$$[P(x_1,\ldots,x_n)\phi]\hat{}(\boldsymbol{u}) = P\left(-i\frac{\partial}{\partial u_1},\ldots,-i\frac{\partial}{\partial u_n}\right)\hat{\phi}(\boldsymbol{u}), \tag{54}$$

$$[\phi(\boldsymbol{x}-a)]\hat{}(\boldsymbol{u}) = e^{i\boldsymbol{a}\cdot\boldsymbol{u}}\hat{\phi}(\boldsymbol{u}), \tag{55}$$

$$[\hat{\phi}(\boldsymbol{x})](\boldsymbol{u}+\boldsymbol{a}) = [e^{i\boldsymbol{a}\cdot\boldsymbol{x}}]\hat{}\phi(\boldsymbol{u}), \tag{56}$$

$$[\phi(\boldsymbol{A}\boldsymbol{x})]\hat{}(u) = |\det\boldsymbol{A}|^{-1}\hat{\phi}(\boldsymbol{A}^T)\boldsymbol{u}), \tag{57}$$

where $\boldsymbol{u}\cdot\boldsymbol{x} = u_1x_1+\cdots+u_nx_n$, $\boldsymbol{A}$ is a nonsingular matrix, and $\boldsymbol{A}^T$ is its transpose. We shall refer to the numbered formulas above for $n=1$ also.

As a simple example we consider the function $\exp(-x^2/2)$, which is clearly a member of S. To find its transform we first observe that it satisfies the equation $\phi'(x)+x\phi(x)=0$. Taking Fourier transforms of both sides of this equation and using (52) and (53), we find that $(d/du)[\hat{\phi}(u)e^{u^2/2}]=0$. Thus, $\hat{\phi}(u) = Ce^{-u^2/2}$, where C is a constant. To evaluate C, we observe that $\hat{\phi}(u) = \int_{-\infty}^{\infty}\exp(-x^2/2)\,dx = \sqrt{2\pi}$, so that $C=\sqrt{2\pi}$ and we have $\hat{\phi}(u) = \sqrt{2\pi}\phi(u)$. Thus we have found a function which is its own inverse. [The multiplicative $\sqrt{2\pi}$ disappear if we use the factors $1/\sqrt{2\pi}$ in the definition of the transform pairs (48) and (49)].

Fourier transform of tempered distributions. Having discovered that $\hat{\phi}\in S$ when ϕ is, we can apply the relation $\langle\hat{t},\phi\rangle = \langle t,\hat{\phi}\rangle$ to define the Fourier transform of the tempered distributions $t(x)$. Then all the formulas given above for $\hat{\phi}$ carry over for $\hat{t}$. For instance, $\langle|(d^k/dx^k)t(x)]\hat{}u, \phi(u)]\rangle = \langle d^kt/dx^k,\ \hat{\phi}(x)\rangle = (-1)^k\langle t(x), [(d^k/dx^k)\phi(x)]\hat{}\rangle = \langle t(x), (-iu)^k\phi(u)]\hat{}(x)\rangle = \langle\hat{t}(u), (-iu)^k\phi(u)\rangle = \langle(-iu)^k\hat{t}(u), \phi(u)\rangle$, which shows that $[d^kt/dx^k]\hat{}(u) = (-iu)^k\hat{t}(u)$, which agrees with (51). Instead of writing the formulas (50)–(57) all over again for a distribution $t(x)$, we shall merely refer to them with ϕ replaced by t. Let us now give some important applications of these formulas.

1. *Delta function.* $\langle[\delta(x)]\hat{}, \phi(u)\rangle = \langle\delta(x), \int_{-\infty}^{\infty}\phi(u)e^{iux}\,du\rangle = \int_{-\infty}^{\infty}\phi(u)du = \langle 1, \phi(u)\rangle$. Thus $\hat{\delta}(x)=1$ so that $\hat{1} = 2\pi\delta(x)$. In the n-dimensional case, the corresponding formulas are $\hat{\delta}(x)=1$ and $\hat{1} = (2\pi)^n\delta(\boldsymbol{x})$. Incidentally, we recover formula (12) from this relation.

2. *Heaviside function.* We use formula (51), so that $[t'(x)]\hat{}(u) = -iu\hat{t}(u)$. Because $t(x)=H(x)$, we have $[\delta(x)]\hat{}(u) = -iu[H(x)]\hat{}(u)$ or $u(H)(x)]\hat{}(u)=i$, whose solution follows from (25) as $[H(x)]\hat{}(u) = c\delta(u) + iPf(1/u)$. Similarly, from (57) and the above result we have $[H(-x)]\hat{}(u) = c\delta(u) - iPf(1/u)$. To find the constant c we observe that $H(x)+H(-x)=1$, whose Fourier transform is $2c\delta(u) = 2\pi\delta(u)$, so that $c=1$. Thus $[H(\pm x)]\hat{}(u) = \pi\delta(u) \pm iPf(1/u)$.

3. *Signum function.* Because $\operatorname{sgn}x = H(x)-H(-x)$, we take Fourier transforms of both sides and use Example 2 above to get $[\operatorname{sgn}x]\hat{}(u) = 2iPf(1/u)$.

4. $Pf(1/x)$. We use formulas (47) for $t(x)$ and the example above. The result is $[(\operatorname{sgn}x)\hat{}(u)](x) = [2iPf(-1/u)]\hat{}(x)$, so that

$$\left[Pf\left(\frac{1}{x}\right)\right]\hat{}(u) = i\pi\operatorname{sgn}u. \tag{58}$$

5. *The function* $|x|$. We write it as $|x| = xH(x) - xH(-x)$. Taking the Fourier transforms of both sides of this equation, we get

$$\begin{aligned}[|x|]\hat{}(u) &= [xH(x)]\hat{}(u) - [xH(-x)]\hat{}(u)\\ &= -i\frac{d}{du}[H(x)]\hat{}(u) + i\frac{d}{du}[H(-x)]\hat{}(u)\\ &= -i\frac{d}{du}\left[\pi\delta(u) + iPf\left(\frac{1}{u}\right)\right]\\ &\quad \pm i\frac{d}{du}\left[\pi\delta(u) - iPf\left(\frac{1}{u}\right)\right]\\ &= \frac{2}{u^2}.\end{aligned}$$

Thus $[|x|]\hat{}(u) = 2/u^2$.

6. *The function* $1/x^2$. To find its transform we appeal to the previous example and relation (47) so that $[(|x|)\hat{}(u)]\hat{}(x) = (2/u^2)\hat{}(x)$ and we get $(1/x^2)\hat{}(u) = \frac{1}{2}|x|$.

7. *Polynomial* $P(x) = a_0 + a_1 + \cdots + a_nx^n$. Formula (54) gives $[P(x)t(x)]\hat{}(u) = P[-i(d/du)\hat{t}(u)]$. When we

substitute $t(x)=1$ in this formula and use Example 1, we get $[P(x)]\hat{}(u)=2\pi P(-i\,d/du)\delta(u)$. Thus,

$$\left[a_0+a_1+\cdots+a_nx^n\right]\hat{}(u)=2\pi\,(a_0\delta-ia_1\delta'+\cdots+(-i)^na_n\delta^{(n)}(u)\Big]. \quad (59)$$

In particular, $\hat{x}(u)=-2\pi i\delta'(u)$, $[x^2]\hat{}(u)=-2\pi\delta''(u)$, $\ldots$, $[x^n]\hat{}(u)=(-i)^n2\pi\delta^{(n)}(u)$.

8. $Pf(1/|x|)$. With the help of the definition (24) of this function we have

$$\left\langle Pf\left(\frac{1}{|x|}\right)\right]\hat{}(u),\phi(u)\Big\rangle=\left\langle Pf\left(\frac{1}{|x|}\right),\hat{\phi}(x)\right\rangle$$

$$=\int_{-1}^{1}\frac{\hat{\phi}(x)-\hat{\phi}(0)}{|x|}\,dx+\int_{|x|>1}\frac{\hat{\phi}(x)}{|x|}\,dx,$$

which, after some algebraic manipulation, yields

$$\left[Pf\left(\frac{1}{|x|}\right)\right]\hat{}(u)=-2(\gamma+\ln|u|), \quad (60)$$

where γ is Euler's constant.

9. $\ln|x|$. For evaluation of the Fourier transform of this important function, we take the Fourier transform of both sides of (60), use formula (47), and obtain

$$Pf\left(\frac{1}{|x|}\right)=2[2\pi\gamma\delta(x)+[\ln(u)]\hat{}(x)]$$

and relabel. The result is

$$[\ln|x|]\hat{}(u)=-\left[\pi\,Pf\frac{1}{2|u|}+2\pi\gamma\delta(u)\right]. \quad (61)$$

In the theories of wavelets, sampling, and interpolation, we need the Fourier transforms of the square and triangular functions. We derive them in the next two examples.

10. *The square function.* It is defined as $f(x)=H(x+\frac{1}{2})-H(x-\frac{1}{2})$. Thus

$$\hat{f}(u)=\int_{-1/2}^{1/2}e^{iux}\,dx=\left[\frac{1}{iu}e^{iux}\right]_{-1/2}^{1/2}$$

$$=\frac{2}{u}\sin\left(\frac{u}{2}\right)=\operatorname{sinc}\left(\frac{u}{2}\right), \quad (62)$$

where $\operatorname{sinc}t=\sin t/t$.

11. *The triangular function.* It is defined as

$$f(x)=\begin{cases}1-|x|, & |x|<1,\\ 0, & |x|>1,\end{cases}$$

so that

$$\hat{f}(u)=\int_{-1}^{0}(1+x)\,e^{iux}dx+\int_{0}^{1}(1-x)e^{iux}\,dx.$$

By a routine computation it yields

$$\hat{f}(u)=\left[\frac{\sin(u/2)}{(u/2)}\right]^2=\left[\operatorname{sinc}\left(\frac{u}{2}\right)\right]^2. \quad (63)$$

12. *Fourier transform of an integral.* We have found previously that taking the Fourier transform of the diferential of a distribution $t(x)$ has the effect of multiplying $\hat{t}(u)$ by $(-iu)$. In the case of taking the Fourier transform of the integral of $t(x)$, it amounts to dividing $\hat{t}(u)$ by $(-iu)$ so that

$$\left\{\int_a^x[t(s)\,ds]\right\}\hat{}\,(u)=\frac{i\hat{t}(u)}{u}. \quad (64)$$

Similarly,

$$\left\{\int_{-\infty}^x[t(s)\,ds]\right\}\hat{}\,(u)=\frac{i\hat{t}(u)}{u}+\frac{1}{2}\hat{t}(0)\delta(u). \quad (65)$$

13. *Fourier transform of the convolution.* The Fourier transform of the convolution $f*g$ of two locally integrable functions f and g is $[f*g]\hat{}(u)=\int_{-\infty}^{\infty}f*ge^{iux}=\int_{-\infty}^{\infty}e^{iux}\int_{-\infty}^{\infty}f(x-y)g(y)\,dy=\int_{-\infty}^{\infty}g(y)e^{iuy}\,dy\times\int_{-\infty}^{\infty}f(x-y)\,e^{iu(x-y)}\,d(x-y)=\hat{g}(u)\hat{f}(u)=\hat{f}(u)\hat{g}(u)$. Thus, the Fourier transform of the convolution of two regular distributions is the product of their transforms. This relation also holds for singular distributions with slight restrictions. For instance, if at least one of these distributions has compact support, then the relation holds.

X. POISSON SUMMATION FORMULA

To derive the Poisson summation formula we first find the Fourier series of the delta function in the period $[0,2\pi]$: $\delta(x)=\sum_{m=0}^{\infty}(a_m\cos mx+b_m\sin mx)$, where the coefficients a_m, b_m are given by $a_0=(1/2\pi)\int_0^{2\pi}\delta(x)\,dx=(1/2\pi)$, $a_m=(1\pi)\int_0^{2\pi}\delta(x)\cos mx\times dx=(1/\pi)$, and $b_m=(1/\pi)\int_0^{\pi}\delta(x)\sin mx\,dx=0$. Thus, we have

$$\delta(x)=\frac{1}{2\pi}\left(1+2\sum_{m=1}^{\infty}\cos mx\right)=\frac{1}{2\pi}\sum_{m=-\infty}^{\infty}e^{imx}. \quad (66)$$

Now we periodize $\delta(x)$ by putting the row of deltas at the points $2\pi m$ so that relation (66) can be written as

$$\sum_{m=-\infty}^{\infty}\delta(x-2\pi m)=\frac{1}{2\pi}\left(1+2\sum_{m=1}^{\infty}\cos mx\right)$$

$$=\frac{1}{2\pi}\sum_{m=-\infty}^{\infty}e^{imx}.$$

When we set $x = 2\pi y$ in this relation and use the formula $\delta[2\pi(y-m)] = (1/2\pi)\delta(y-m)$ and relabel, we obtain

$$\sum_{m=-\infty}^{\infty} \delta(x-m) = 1 + 2\sum_{m=1}^{\infty} \cos 2\pi m x = \sum_{m=-\infty}^{\infty} e^{i2\pi m x}. \tag{67}$$

The action of this formula on a function $\phi(x) \in \mathcal{S}$ yields

$$\sum_{m=-\infty}^{\infty} \phi(m) = \hat{\phi}(2\pi m), \tag{68}$$

which relates the sum of functions ϕ and their Fourier transforms $\hat{\phi}$ and is a very useful formula.

In the classical theory, it is necessary that both sides of relation (68) converge. Moreover, they must converge in the same interval. With the help of the theory of distributions we can obtain many variants of relation (68) which are applicable even when one or both of the series (68) are divergent. As an example, we set $x = x/\lambda$, where λ is a real number, and use the relation $\delta[(x/\lambda) - m] = |\lambda|\delta(x - m\lambda)$. Then relation (67) becomes

$$(\lambda) \sum_{m=-\infty}^{\infty} \delta(x - m\lambda) = \sum_{m=-\infty}^{\infty} e^{2i\pi x m/\lambda}. \tag{69}$$

When we multiply both sides of this relation by a test function $\phi(x)$ and integrate with respect to x, we obtain a variant of (68) as

$$\sum_{m=-\infty}^{\infty} \phi(m\lambda) = \frac{1}{|\lambda|} \sum_{m=-\infty}^{\infty} \hat{\phi}\left(\frac{2\pi m}{\lambda}\right). \tag{70}$$

This is called the distributional Poisson summation formula. Among other things, the Poisson summation formula (70) transforms a slowly converging series to a rapidly converging series. For instance, if we take $\phi(x) = e^{-x^2}$, then $\hat{\phi}(u) = \sqrt{\pi} e^{-u^2/2}$, so that (70) becomes

$$\sum_{m=-\infty}^{\infty} e^{-m^2\lambda^2} = (\pi/\lambda)^{1/2} \sum_{m=-\infty}^{\infty} e^{-m^2\pi^2/\lambda^2}. \tag{71}$$

The series on the left side of (71) converges rapidly for large λ, that on the right side for small λ.

XI. ASYMPTOTIC EVALUATION OF INTEGRALS: A DISTRIBUTIONAL APPROACH

Let $g(x)$ and $h(x)$ be sufficiently smooth functions on the interval $[a, b]$, then the main contribution to the Laplace integral,

$$I(\lambda) = \int_a^b e^{-\lambda h(x)} g(x)\, dx, \qquad \lambda \to \infty \tag{72}$$

arises from the points where the function $h(x)$ has a minimum. If x_0 is the only global minumum of $h(x)$, then we have the Laplace formula

$$I(\lambda) \sim \left[\frac{2\pi}{\lambda h''(x_0)}\right]^{1/2} g(x_0) e^{-\lambda h(x_0)}. \tag{73}$$

If we could somehow prove that $e^{-\lambda h(x)}$ has an asymptotic series of delta functions such that

$$e^{-\lambda h(x)} \sim \left[\frac{2\pi}{\lambda h'(x_0)}\right]^{1/2} e^{-\lambda h(x_0)} \delta(x - x_0), \tag{74}$$

then all that we have to do is to substitute (74) in (72) and use the sifting property of the delta function and formula (73) follows immediately. To achieve the expansion (74) we first define the moments of a function $f(x)$. They are

$$\mu_n = \langle f(x), x^n \rangle = \int_{-\infty}^{\infty} f(x) x^n\, dx. \tag{75}$$

The Taylor expansion of a test function $\phi(x)$ at $x = 0$ is

$$\phi(x) = \sum_{n=0}^{\infty} \phi^{(n)}(0) \frac{x^n}{n!}. \tag{76}$$

Then it follows from (75) that

$$\begin{aligned} \langle f(x), \phi(x) \rangle &= \left\langle f(x), \sum_{n=0}^{\infty} \phi^{(n)}(0) \frac{x^n}{n!} \right\rangle \\ &= \sum_{n=0}^{\infty} \phi^n(0) \frac{\mu_n}{n!}. \end{aligned} \tag{77}$$

But $\phi^n(0) = (-1)^n \langle \delta^{(n)}(x), \phi(x) \rangle$, so that (77) becomes

$$\langle f(x), \phi(x) \rangle = \left\langle \sum_{n=0}^{\infty} \frac{(-1)^n \mu_n \delta^{(n)}(x)}{n!}, \phi(x) \right\rangle. \tag{78}$$

Thus

$$f(x) = \sum_{n=0}^{\infty} \frac{(-1)^n \mu_n \delta^{(n)}(x)}{n!}. \tag{79}$$

Finally, we use the formula $\delta^{(n)}(\lambda x) = (1/\lambda^{n+1})\delta^n(x)$ in (79) and obtain the complete asymptotic expansion

$$f(\lambda x) = \sum_{n=-\infty}^{\infty} \frac{(-1)^n \mu_n \delta^{(n)}(x)}{n!\, \lambda^{n+1}}, \qquad \lambda \to \infty. \tag{80}$$

In the case of the Laplace integral (72) we have $f(x) = e^{-\lambda h(x)}$. At the minimum point x_0, $h(x) \sim h(x_0) + [h''(x_0)(x - x_0)^2]/2$. Accordingly, we can find an increasing smooth function $\psi(x)$ with $\psi(x_0) = 0$, $\psi'(x_0) > 0$, so that $h(x) = h(x_0) + [\psi(x)]^2$ in the support of $h(x)$. Then $h'(x_0) = 2\psi(x_0)\psi'(x_0) = 0$, which yields $h'(x_0) = 0$ as required for $h(x)$ to have minimum at x_0. Also,

$h''(x) = 2[\psi'(x)]^2 + 2\psi(x)\psi''(x)$, which, for $x = x_0$, becomes $h''(x_0) = 2[\psi'(x_0)]^2$ or

$$\psi'(x_0) = \frac{1}{2}[h''(x_0)]^{1/2}. \tag{81}$$

Substituting this information about $h(x)$ in the Laplace integral (72), we obtain

$$I(\lambda) = e^{-\lambda h(x_0)} \int_{-\infty}^{\infty} e^{-\lambda[\psi(x)]^2} g(x)\, dx. \tag{82}$$

The next step is to set $u = \psi(x)$, which gives $dx = du/\psi'(x)$ so that (82) becomes

$$I(\lambda) = e^{-\lambda h(x_0)} \int_{-\infty}^{\infty} e^{-\lambda u^2} \frac{g(u)}{\psi'(x_0)}\, du, \tag{83}$$

so we need the asymptotic expansion of $e^{-\lambda u^2}$ from formula (80). This is obtained by finding the moments of e^{-u^2}, and they are

$$\mu_n = \int_{-\infty}^{\infty} e^{-u^2} u^n\, du = \begin{cases} \Gamma\left(\dfrac{n+1}{2}\right), & n \text{ even}, \\ 0, & n \text{ odd}. \end{cases} \tag{84}$$

Then formula (80) yields

$$e^{-\lambda u^2} = \sum_{n=0}^{\infty} \frac{\Gamma[(2n+1)/2]\delta^{(2n)}(u)}{(2n)!\lambda[(2n+1)/2]} \tag{85}$$

so that formula (83) becomes

$$I(\lambda) = e^{-\lambda h(x_0)} \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{\Gamma[(2n+1)/2]}{2n!} \frac{\delta^{(2n)}(u)}{\lambda^{(2n+1/2)}} \times \frac{g(u)}{\psi'(x_0)}\, du. \tag{86}$$

The interesting feature of the distributional approach is that we have obtained the complete asymptotic expansion. Thereafter, we evaluate as many terms as are needed to get the best approximation.

The first term of formula (86) is

$$I(\lambda) = e^{-\lambda h(\alpha_0)} \int_{-\infty}^{\infty} \Gamma\left(\frac{1}{2}\right) \frac{\delta(u)}{\sqrt{\lambda}} \frac{g(u)}{\psi'(x_0)}\, du.$$

When we substitute the value of $\psi'(x_0)$ from relation (81) in the above relation and use the sifting property of $\delta(u)$, we obtain

$$I(\lambda) \sim \left[\frac{2\pi}{\lambda h''(x_0)}\right]^{1/2} g(x_0) e^{-\lambda h(x_0)},$$

which agrees with (73).

The *oscillatory integral*

$$I(\lambda) = \int_{-\infty}^{\infty} e^{i\lambda h(x)} g(x)\, dx, \qquad \lambda \to \infty \tag{87}$$

can also be processed in the same manner as the Laplace integral. Indeed, the steps leading (87) to relation

$$I(\lambda) = e^{-i\lambda h(x_0)} \int_{-\infty}^{\infty} e^{i\lambda u^2} \frac{g(u)}{\psi'(x_0)}\, du \tag{88}$$

are almost the same as these from (72) to (83). The difference arises in the value of the moments μ_n which now are

$$\mu_n = \int_{-\infty}^{\infty} e^{i\lambda u^2} u^n\, du = \begin{cases} \Gamma\left(\dfrac{n\pi}{2}\right) e^{\pi i[(2n+1)/4]}, & n \text{ even}, \\ 0, & n \text{ odd}. \end{cases} \tag{89}$$

This yields the moment expansion for $e^{i\lambda u^2}$ as

$$e^{i\lambda u^2} = \sum_{n=0}^{\infty} \frac{\Gamma[(2n+1)/2]\, e^{-\pi i[(2n+1)/4]}\delta^{(2n)}(u)}{(2n)!\lambda^{(2n+1)/2}}\, du. \tag{90}$$

When we substitute this value for $e^{i\lambda u^2}$ in integral (88) we obtain

$$I(\lambda) \sim e^{-ih(x_0)} \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{\Gamma[(2n+1)/2]\, e^{-\pi i[(2n+1)/4]}}{(2n)!} \times \frac{\delta^{(2n)}(n)}{\lambda^{(2n+1/2)}} \frac{g(u)}{\psi'(x_0)}\, du. \tag{91}$$

The first term of this formula yields

$$I(\lambda) \sim e^{i(\lambda h(x_0) + \pi/4)} \left[\frac{2\pi}{h''(x_0)}\right]^{1/2} \frac{\phi(x_0)}{\sqrt{\lambda}}, \tag{92}$$

where we have used relation (81).

Let us observe an essential difference between the Laplace integral and the oscillatory integral. In the case of the Laplace integral we collect the contributions only from the points where $h(x)$ has the minimum such that $h'(x) = 0$ and $h''(x) > 0$. For the oscillatory integral we collect the contributions from all the points where $h(x)$ is stationary such that $h'(x) = 0, h''(x) \neq 0$. For this reason the classical method for the oscillatory integral is called the method of stationary phase.

SEE ALSO THE FOLLOWING ARTICLES

FOURIER SERIES • FUNCTIONAL ANALYSIS

BIBLIOGRAPHY

Benedetto, J. J. (1996). "Harmonic Analysis and Applications," CRC Press, Boca Raton, FL.

Duran, A. L., Estrada, E., and Kanwal, R. P. (1998). "Extensions of the Poisson Summation Formula," *J. Math. Anal. Appl*. **218**, 581–606.

Estrada, R., and Kanwal, R. P. (1985). "Regularization and distributional derivatives of $1/r^n$ in $\mathbb{R}^p$," *Proc. Roy. Soc. Lond*. A **401**, 281–297.

Estrada, R., and Kanwal, R. P. (1993). "Asymptotic Analysis: A Distributional Approach," Birkhäuser, Boston.

Estrada, R., and Kanwal, R. P. (1999). "Singular Integral Equations," Birkhäuser, Boston.

Gel'fand, I. M., and Shilov, G. E. (1965). "Generalized Functions, Vol. I," Academic Press, New York.

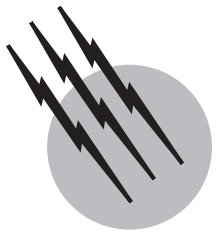
Jones, D. S. (1982). "Generalized Functions," Cambridge University Press, Cambridge, UK.

Kanwal, R. P. (1983). "Generalized Functions," 2nd ed., Birkhauser, Boston.

Lighthill, M. J. (1958). "Introduction to Fourier Analysis and Generalised Functions," Cambridge University Press.

Saichev, A. I., and Woyczynsky, W. A. (1997). "Distributions in the Physical and Engineering Sciences," Vol. 1, Birkhäuser, Boston.

Schwartz, L. (1966). "Théorie des distributions," Nouvell èdition, Hermann, Paris.



Graph Theory

Ralph Faudree
University of Memphis

I. Introduction
II. Connectedness
III. Trees
IV. Eulerian and Hamiltonian Graphs
V. Colorings of Graphs
VI. Planar Graphs
VII. Factorization Theory
VIII. Graph Reconstruction
IX. Extremal Theory
X. Directed graphs
XI. Networks
XII. Random Graphs

GLOSSARY

Bipartite graph A graph whose vertices can be partitioned into two sets of independent vertices.
Connectivity (edge) The minimum number of vertices (edges) which when deleted leaves a disconnected graph.
Coloring (edge) An assignment of colors to the vertices (edges) of a graph so that adjacent vertices (edges) have different colors.
Component A maximal connected subgraph of a graph.
Digraph A pair consisting of a finite nonempty set called the vertex set and a collection of ordered pairs of the vertices called the arc set.
Factorization A collection of subgraphs of a graph which partition the edges of the graph.
Good algorithm An algorithm with a polynomial upper bound on the time required to terminate in the worst case.
Graph A pair consisting of a finite nonempty set called the vertex set and a collection of unordered pairs of the vertices called the edge set.
Independent set A set of vertices (edges) such that no pair is adjacent.
Isomorphism A one-to-one map from the vertices of one graph onto the vertices of another graph which preserves edges.
Matching A collection of independent edges in a graph.

Network A directed graph in which each arc is given a label.

Planar graph A graph which can be embedded in the plane or on a sphere.

Reconstructible graph (edge) A graph which is determined by its subgraphs obtained by deleting a single vertex (edge) of the graph.

Tournament A directed graph obtained by orienting the edges of a complete graph.

Tree A connected graph which has no cycles.

A GRAPH is a finite collection of elements, which are called vertices, and a finite collection of lines or curves, which are called edges, that join certain pairs of these vertices. This is an abstract mathematical structure which is of interest on its own, but it is also of interest as a structure which can be used to model a wide variety of real-life situations and problems. Street maps, communications systems, electrical networks, organizational structures, and chemical molecules all can be viewed as graphs. Some historical roots of the subject, important results about the structure and theory of graphs, outstanding open problems in graph theory, and graphs as mathematical models will be presented in this article. Procedures (called algorithms) for solving graphical problems will be described, and the application of the structure of graphs and graphical algorithms to problems modeled by graphs will also be discussed.

I. INTRODUCTION

A *graph* G is generally thought of as a collection of points together with curves or lines which join certain pairs of these points. Thus, this simple structure is a natural model for many real-life situations. For example, a road map uses a graph as a model. The cities are represented by points and the roads between cities are represented by curves or lines. Organizational charts can be considered as graphs, with the points representing the individuals in the organization and lines indicating that one person is an immediate supervisor of another. In this case it may be more appropriate to think of the lines as being directed from the supervisor to the person being supervised. Communication systems, electrical networks, family trees, and molecules in chemistry can all be represented as graphs. Our initial objectives will be to formalize this idea of graph, introduce some standard notation and terminology, develop the basic concepts, and illustrate these with examples.

A *graph* G is a nonempty finite set $V(G)$ together with a finite set $E(G)$ of distinct unordered pairs of elements of $V(G)$. Each element in $V(G)$ is called a vertex and each element in $E(G)$ is called an edge. When it is clear just what graph is being considered, the vertex set $V(G)$ and the edge set $E(G)$ will be written as just V and E, respectively. The *order* of G, usually denoted by $|V|$, is the number of elements in V, while the *size* of G is the number of elements in E, and is denoted by $|E|$.

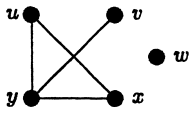


FIGURE 1 Graph G.

For vertices u and v of V the edge $e = \{u, v\}$ will be expressed more compactly as just uv. We say that u is adjacent to v and e is incident to u and v. The edge e is said to join the vertices u and v. Also, two edges incident to a common vertex are said to be adjacent to each other. The neighborhood $N_G(v)$ [or just $N(v)$] of a vertex v is the set of vertices of G which are adjacent to v, and the number of elements in $N(v)$ is called the *degree* of v and is denoted by $d_G(v)$ [or just $d(v)$]. Of course, the degree $d_G(v)$ is also the number of edges of G incident to v.

A graph has a useful geometric representation in the plane, with points representing vertices and curves or lines representing edges. The graph G pictured in Fig. 1 can be used to illustrate the notation that has been presented. The graph G has order 5 and size 4 with $V = \{u, v, w, x, y\}$ and $E = \{ux, uy, vy, xy\}$. The degrees of the vertices are $d(u) = d(x) = 2, d(v) = 1, d(y) = 3$, and $d(w) = 0$. We say that w is an *isolated* vertex.

With this minimal background we can now make an elementary but useful observation. If we sum the degrees of vertices of a graph, then each edge of the graph will be counted exactly twice, once at each end. This gives the following result.

Theorem 1.1: *If G is a graph of order p and size q with $V = \{v_1, v_2, \ldots, v_p\}$, then $\sum_{i=1}^{p} d(v_i) = 2q$.*

The concept of substructure is of basic importance in graph theory. A graph H is a *subgraph* of a graph G if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. Also, the graph G is said to be a *supergraph* of H. If the edges of H are precisely the edges of G which join pairs of vertices of H, then H is an *induced* subgraph of G. If H is a subgraph with $V(H) = V(G)$, then H is called a *spanning subgraph* of G. There are some special subgraphs and supergraphs of a graph G that occur so frequently that special notation is given to them. If $e \in E$, then $G - e$ is the subgraph of G with the same vertices and edges as G except the edge e is deleted. Similarly, if $v \in V$, then $G - v$ is the subgraph of G with the vertex v deleted from the vertex set and all the edges incident to v deleted from the edge set. There are

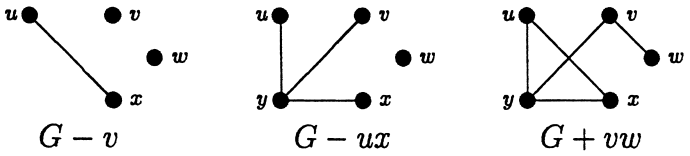


FIGURE 2

obvious generalizations to sets of vertices or edges. Also, if e is an edge not in E, then $G+e$ is the supergraph of G with the same vertex set as G but with e added to the edge set. The concepts of subgraph and supergraph are illustrated in Fig. 2, with G being the graph pictured in Fig. 1.

Two graphs G and H are identical (written as $G=H$) if $V(G)=V(H)$ and $E(G)=E(H)$. However, it is possible for graphs to have the same appearance if they are appropriately labeled but yet not be identical. Two graphs G and H are said to be *isomorphic* (written as $G\cong H$) if there is a one-to-one mapping (called an isomorphism) θ of $V(G)$ onto $V(H)$ such that edge uv is in $E(G)$ if and only if the edge $\theta(u)\theta(v)$ is in $E(H)$. The graph G pictured in Fig. 3 is isomorphic to the graph H of Fig. 3; in fact, the map θ defined by $\theta(u)=a, \theta(v)=d, \theta(w)=f, \theta(x)=b$ and $\theta(y)=c$ is an isomorphism from G onto H.

Some classes of graphs occur so frequently that special names are given to them. A graph of order n in which every pair of its vertices are adjacent is called a *complete graph* and is denoted by K_n. Each vertex in K_n has degree $n-1$ and the size of the graph is $\binom{n}{2}=n(n-1)/2$. The *complement* $\bar{G}$ of a graph G is the graph with vertex set $V(G)$ such that two vertices are adjacent in $\bar{G}$ if and only if they are not adjacent in G. Thus $\bar{K}_n$ is a graph with no edges. A *bipartite* graph is one in which the vertices can be partitioned into two sets, say A and B, such that every edge of the graph joins a vertex in A and a vertex in B. The sets A and B are called the parts of the bipartite graph. If every vertex in A is joined to every vertex in B, then the bipartite graph is called a *complete bipartite graph*, and is denoted by $K_{m,n}$ if $|A|=m$ and $|B|=n$. Thus, $K_{m,n}$ has order $m+n$ and size mn. The special complete bipartite graph $K_{1,n}$ is called a *star* of size n (and order $n+1$). A collection of vertices of a graph are said to be *independent* if there are no edges joining them. Thus, the vertices of the complement of a complete graph and the vertices in the parts of a bipartite graph are independent.

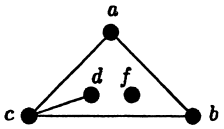


FIGURE 3 Isomorphic graph H.

If all of the vertices of a graph have the same degree r, then the grahph is said to be *r-regular* or just *regular*. The complete graph K_{r+1} and the complete bipartite graph $K_{r,r}$ are examples of r-regular graphs. There are many other special graphs, but we conclude with the graph C_n, a *cycle* of length n. This is a connected graph of order n which is 2-regular. The graph C_5 is pictured in Fig. 4 along with other examples of complete graphs, bipartite graphs, and their complements.

Up to this point only one form of a "graph" has been discussed, and that is what is sometimes called a simple graph. No edges of the form vv, which are called *loops*, were allowed. Also, any pair of vertices in a graph were joined by at most one edge, not *multiple edges*. If loops and multiple edges are allowed, the structure is called a *multigraph*. This definition of a multigraph is not completely standard because loops are sometimes not allowed. Figure 5a shows an example of a multigraph which has multiple edges joining vertices u and x and loops at the vertex v. All edges have to this point been assumed to have no direction, so that the edge uv is the same as the edge vu. Sometimes it is appropriate for each edge to have a direction, and such directed edges will be called *arcs*. Therefore each arc is an ordered pair of distinct vertices, and the arc uv is not the same as the arc vu. Such structures are called *directed graphs* or *digraphs*. An example of a digraph is given in Fig. 5b. All of these structures and some additional variations will be considered later.

There are many ways to describe or determine a particular graph. So far, we have generally described graphs by using the definition and listing the vertices and edges of the graph. In a few cases, a picture or drawing of the

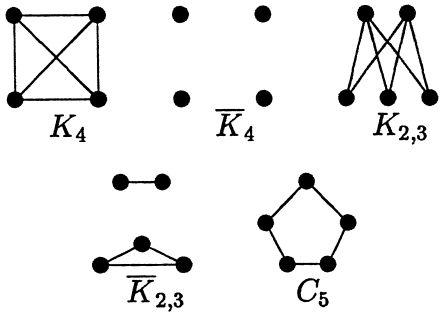


FIGURE 4 Special graphs and complements.

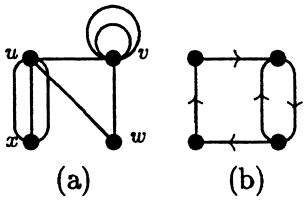


FIGURE 5 (a) Multigraph. (b) Digraph.

graph was used to determine the graph. The *adjacency matrix* $A(G)$ of a graph G is a common and useful structure to describe a graph. If G has order n, then $A(G)$ is an n by n matrix with n rows and columns and 0 or 1 entries in which the ith vertex corresponds to the ith row and column. The (i, j) entry is 1 if and only if the ith vertex is adjacent to the jth vertex. The matrix $A(G)$ depends on an ordering of the vertices, and a different ordering of the vertices will cause a permutation of the rows and columns of the adjacency matrix. Also, $A(G)$ is a symmetric matrix with only 0's on the diagonal, and the number of 1's in the ith row (column) is the degree of ith vertex. Similar adjacency matrices can be defined for digraphs and multigraphs.

The study of procedures or *algorithms* for solving graphical problems is central in the theory of graphs. After the existence of an algorithm has been verified, the question of efficiency must be answered. An algorithm is said to be *good* if it is a polynomial-time algorithm. This means that if n is some parameter which describes the magnitude of the graph (such as order or size), then there exists positive constants c and k such that in the worst case the time (the number of operations such as additions or comparisons) required for the algorithm to terminate is bounded above by cn^k. An n^k algorithm is one in which the time required is at most cn^k for some positive constant c. If $k = 1$, then the algorithm is said to be linear.

The reason for searching for good algorithms as opposed to those with no polynomial upper bound is that those with exponential growth can be extremely inefficient. The following is a dramatic example of this. If $n = 20$, which is a relatively small number, then $n^3 = 8000$. On the other hand $n!$ is a number with 19 digits. If a computer could do 1,000,000,000 calculations a second, it would take less than 1/(100,000)th of a second to do n^3 calculations, but it would take over 75 years to do $n!$ calculations. However, some care must be taken in interpreting the relative efficiencies of algorithms for small values of n. Although an n^3 algorithm will be quicker than an n^5 or an $n!$ algorithm for large values of n, for small values of n the constant associated with each of the algorithms may be the dominant factor. Either an n^5 or $n!$ algorithm could be faster than an n^3 algorithm for small values of n.

Consider two graphs G and H of the same order (say, n) and size, which are each described by a vertex and edge set. There is a good algorithm to determine if the graphs are identical. First determine if $V(G) = V(H)$. If this is true, order the vertex sets in the same way and form the adjacency matrices $A(G)$ and $A(H)$. If $A(G) = A(H)$, then G and H are identical. Each of these steps takes at most n^2 comparisons, so we have a good algorithm; in fact, this is an n^2 algorithm for determining if two graphs are identical. Next consider the problem of determining if G and H are isomorphic. A reasonable approach would be to order the sets $V(G)$ and $V(H)$ (they need not be identical) and construct the adjacency matrices $A(G)$ and $A(H)$. However, $G \cong H$ if and only if there is some ordering of the sets $V(G)$ and $V(H)$ such that $A(G) = A(H)$. Unfortunately, there are $n!$ possible permutations of an ordered set with n elements. Since $n!$ is not polynomial in n, this approach will not give a good algorithm. It is not known if a good algorithm for isomorphism exists. In fact, this problem, the *isomorphism problem*, is one of the outstanding problems in graph theory.

In the remaining sections, selected but representative topics from the theory of graphs will be discussed. The topics are by no means exhaustive; for example, graphical enumeration, algebraic graph theory, and matroids are not discussed, and all are important and interesting topics.

II. CONNECTEDNESS

Connectedness is a fundamental concept in graph theory, and most readers will have an intuitive feel for it. However, we first need to introduce some elementary related ideas to present the concept carefully. In a graph G, a *walk* from a vertex u to a distinct vertex v is a finite sequence of alternating vertices and edges $W = v_0 e_1 v_1 \cdots v_{k-1} e_k v_k$ with $v_0 = u$, $v_k = v$ and with each edge of the sequence incident in G to each of the vertices on either side of it in the sequence. The length of the walk W is k, the number of edges in W. If all of the edges of the sequence are distinct, then W is called a *trail*. If, in addition, all of the vertices of W are distinct, it is called a *path*. If the initial vertex and the final vertex are allowed to be the same ($u = v$), the walk, trail or path is called *closed*. A closed path of length at least three determines a graph which is called a *cycle*. A path (or cycle) with n vertices is usually denoted by P_n (or C_n).

Two vertices u and v in a graph G are connected if there is a path from u to v. A graph G is *connected* if each pair of its vertices are connected. Even when G is not connected, all of the vertices of G connected to a fixed vertex form a connected subgraph of G. Thus, a disconnected graph G can be partitioned into connected subgraphs, which are

called the *components* of G. The components are the maximal connected subgraphs of a graph. For example, the graph G of Fig. 1 has two components, one with the four vertices $\{u, v, x, y\}$ and the other with the single vertex w.

Associated with any pair of vertices u and v of a graph G is the distance $d_G(u, v)$ [or just $d(u, v)$ when the graph G is obvious], which is the length of a shortest path from u to v. A natural and interesting problem is the determination of a good procedure or algorithm for finding $d(u, v)$. A brute-force approach of checking all possible paths from u to v would not be very satisfactory, since the number of such paths could be quite large—in fact, of order of magnitude $(n-2)!$ for a graph G of order n.

A good "shortest-path algorithm" was discovered by Dijkstra and determines the distance from a fixed vertex of a graph G of order n to each of the remaining vertices, and it works in order of magnitude n^2. Also, this algorithm will handle the more general structure of a weighted graph G in which each edge e of G is assigned a real number $w(e)$, called its weight. For a weighted graph, the length of a path is the sum of the weights of the edges of the path. It should be noted that any graph G can be considered as a *weighted graph* by assigning weight 1 to each edge of G. The algorithm of Dijkstra is of greedy design and uses a breadth-first search of the graph. It starts by finding the vertex in G which is "closest" to a fixed vertex v. At each step the "next closest" vertex is found using the information about the distance from v to vertices whose distances from v have already been determined. Each of the n steps involves at most n comparisons and n additions of real numbers, and thus n^2 is an upper bound on the order of magnitude of the algorithm.

The "shortest-path algorithm" can be used to determine the components of a graph. All vertices in the same component of a fixed vertex v of a graph would have finite distance from v, while the remaining vertices would have infinite distance. Although there are more efficient methods for finding the components of a graph, the shortest path algorithm is a good one for this purpose.

Connectedness in a graph which is being used as a model for a transportation or communication system is clearly desirable. However, connectedness alone might not be sufficient. If the connected graph G which represented a communication network had a vertex v such that $G-v$ was disconnected (v is called a *cut-vertex* of G), then it would be critical that no failure occur at v, for this could result in a failure of the entire system. Thus, we need some measure of the extent of the connectedness of a graph. With that objective in mind we introduce some additional concepts.

If S is a collection of vertices of a graph G, then $G-S$ is the graph obtained from G by deleting the vertices of S. If G is connected but $G-S$ is disconnected, then S is called a *cut-set*, and the set S is said to separate vertices x and y if they are in different components of $G-S$. The connectivity (sometimes called *vertex connectivity*) $\kappa(G)$ of a graph G is the minimum number of vertices in a cut-set of G. The only graphs without cut-sets are complete graphs, and there the connectivity is one less than the order of the complete graph. A graph G is k-connected if $\kappa(G) \geq k$. Thus, for example, a graph of order at least 3 and with no cut-vertices is 2-connected. It is easily checked that $\kappa(P_n) = 1, \kappa(C_n) = 2, \kappa(K_n) = n-1$, and $\kappa(K_{m,n}) = n$ when $m \geq n$. One would expect that as the connectivity of a graph becomes larger, there would be an increase in the number of "alternative" paths between pairs of vertices. The following results verify this expectation.

Theorem 2.1 (Menger): *For distinct nonadjacent vertices u and v of a graph, the maximum number of internally disjoint (vertex disjoint except for u and v) paths between u and v is equal to the minimum number of vertices that separate u and v.*

Theorem 2.2 (Menger-Whitney): *A graph is k-connected if and only if for each pair of distinct vertices of the graph there are at least k internally vertex disjoint paths between them.*

Vertex connectivity has an analog, *edge connectivity*, denoted by $\kappa_1(G)$. Each of the previous definitions and results concerning vertices has a natural edge analog. For example, if e is an edge of a connected graph G and the graph $G-e$ (the graph obtained from G by deleting the edge e) is disconnected, then e is called a cut-edge. An edge-cut-set is a collection of edges which, when deleted, disconnect the graph, and the edge connectivity is the minimum number of edges in an edge-cut-set. There are results analogous to those of Menger and Whitney which also say that high edge connectivity is equivalent to the existence of many edge disjoint paths between any pair of vertices.

III. TREES

A tree is usually defined as a connected graph which contains no cycles. However, this very useful class of graphs has many other characterizations. The following theorem gives five equivalent statements, each of which could be used as a definition of a tree.

Theorem: 3.1 *The following statements are equivalent for a graph of order n:*

(i) G is connected and has no cycles.
(ii) G is connected and has size $n-1$.
(iii) G has no cycles and has size $n-1$.
(iv) G is a graph in which every edge is a cut-edge.

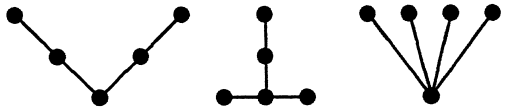


FIGURE 6 Trees of order 5.

(v) G is a graph with a unique path between each pair of distinct vertices.

All nonisomorphic trees of order 5 are pictured in Fig. 6.

Any connected graph G has a *spanning tree*. This is easy to observe. If an edge e is on a cycle of a connected graph, then its deletion will not disconnect the graph. Thus, by successive deletion of edges on cycles of appropriate connected subgraphs of G, one eventually obtains a connected subgraph of G which has no cycles. This graph is a spanning tree T of G. This tree is not in general unique (unless G is itself a tree), but it will have size $n-1$ if G has order n, and there will be no smaller size connected spanning subgraph of G. The number of spanning trees in a graph can be quite large, as the following result of Cayley indicates.

Theorem 3.2 (Cayley): *For $n \geq 2$, the graph K_n has n^{n-2} nonidentical spanning trees.*

There is a more general result on the number of spanning trees of a graph, and this result can be used to prove the theorem of Cayley. Let G be a graph and let $A(G)$ be the adjacency matrix of G. The matrix $D(G)$, called the *degree matrix*, is the diagonal matrix with the degrees of the vertices on the diagonal. Thus, the sum of the entries of each row (column) of the matrix $D(G)-A(G)$ is 0. This implies that all of the cofactors of this matrix $D(G)-A(G)$ are the same. This is the setting for the following result.

Theorem 3.3 (Kirchhoff, matrix tree): *The number of nonidentical spanning trees in any graph G is the value of any cofactor of the matrix $D(G)-A(G)$.*

Consider the problem of building a transportation system that connects (but not necessarily each pair directly) a collection of cities. The cost of constructing the link between each pair of cities is known, and we want to minimize the total cost of the system by selecting the appropriate links. This translates directly into finding a minimal cost spanning tree of a weighted graph. The result of Cayley indicates that the brute-force method of considering all possible spanning trees is not very efficient. An intuitive and good algorithm [of order $n^2/ln(n)$] for finding such a minimal spanning tree was developed by Kruskal. First, the edges of the graph are sorted in increasing order by weight. A subgraph is built by successively adding edges from this sorted list (starting with those of smallest weight), subject only to the condition that an edge is never added if it creates a cycle. This will generate a minimal cost spanning tree.

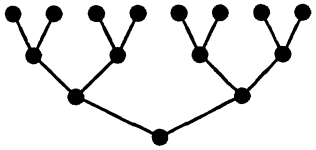


FIGURE 7 Complete binary tree.

Trees are also useful as information structures. One example of this is a binary tree used as a search tree. The *complete binary tree* T of depth k can be defined inductively as a tree with a root vertex v which is adjacent to two vertices v_L and v_R, called the left and right children of v, such that v_L and v_R are roots of disjoint complete binary trees of depth $k-1$ in $T-v$. Figure 7 shows a complete binary tree of depth 3. The tree T has $2^{k+1}-1$ vertices, and the paths from the root to other vertices of the tree have length at most k. Elements from an ordered set with $n=2^{k+1}-1$ elements can be stored at the vertices of the tree in such a way that for any complete binary subtree the elements in any "left tree" are less than the root and elements in the "right tree" are greater than the root. Thus, with the binary search tree the number of comparisons needed to search for an element from this ordered set with n elements has an upper bound of $\log_2(n+1)$, which is considerably less than a naive straightforward approach would require.

IV. EULERIAN AND HAMILTONIAN GRAPHS

In the first paper on graph theory, Euler (1736) considered the problem of traversing the seven bridges of Königsberg (see Fig. 8a) without crossing any bridge twice. He showed that it was not possible. This is equivalent to showing that the multigraph G of Fig. 8b does not contain a trail which uses all of the edges of G.

An *(closed) eulerian trail* of a graph G is a (closed) trail which uses all of the edges of the graph. A graph which contains a closed eulerian trail is called *eulerian*.

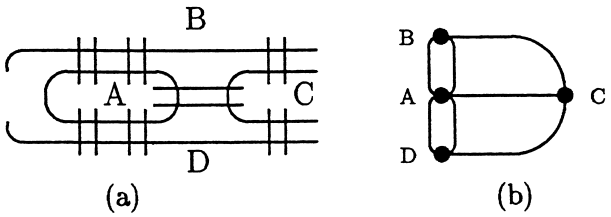


FIGURE 8 Königsberg bridges and multigraph.

Euler showed that the graph G of Fig. 8b has no eulerian trail. For a graph to have such a trail, it is clear that the graph must be connected and that each vertex, except for possibly the first and last vertex of the trail, must have even degree. These conditions are also sufficient, as the following result states.

Theorem 4.1 (Euler): *Let G be a connected graph (multigraph). Then, G has a closed eulerian trail if and only if each vertex has even degree, and G has an "open" eulerian trail if and only if there are precisely two vertices of odd degree.*

A consequence of Theorem 1.1 is that a graph has an even number of vertices of odd degree. There is a useful immediate corollary of Theorem 4.1. If a connected graph G has $2k$ vertices of odd degree, then the edges of G can be "covered" with k trails, and this is the minimum number of trails which will suffice. This observation is the basis of many puzzles and games.

Also, related to eulerian graphs is the *Chinese postman problem*, which is to determine the shortest closed walk that contains all of the edges in a connected graph G. Such a walk is called for obvious reasons a postman's walk. If G has size m, then the postman's walk will have length m if and only if G is eulerian. At the other extreme, this shortest walk will have length $2m$ if and only if G is a tree. There is the obvious extension of the Chinese postman problem to weighted graphs and minimizing the sum of the weights along the postman's walk. There are several good algorithms for solving this problem.

A graph G is *hamiltonian* if it contains a spanning cycle, and the spanning cycle is called a hamiltonian cycle. The name is derived from the mathematician Sir William Rowan Hamilton, who in 1857 introduced a game, whose object was to form such a cycle. In Euler's problem the object was to visit each of the edges exactly once. In the hamiltonian case the object is to visit each of the vertices exactly once, so the problems seem closely related. However, in sharp contrast to the eulerian case, there are no known necessary and sufficient conditions for a graph to be hamiltonian, and the problem of finding such conditions is considered to be very difficult. There are numerous sufficient conditions for the existence of a hamiltonian cycle and a few necessary conditions. The following is an example of one of the better-known sufficient conditions.

Theorem 4.2 (Ore): *If for each pair of nonadjacent vertices u and v of a graph G of order $n \geq 3$, $d(u) + d(v) \geq n$, then G is hamiltonian.*

A traveling salesman wishes to visit all of the cities on his route precisely one time and return to his home city in the smallest possible time. The *traveling salesman problem* is to determine the route which will minimize the time (distance) of the trip. Another version of the same problem is presented by a robot that is tightening screws on a piece of equipment on an assembly line. An order for tightening the screws should be determined so that the distance traveled by the arm of the robot is minimized. The corresponding graph problem in both cases is to determine a minimum-weight hamiltonian cycle in a complete graph, with weights assigned to each edge. The weight assigned to an edge would represent the time or cost of that edge. A brute-force approach of examining all possible hamiltonian cycles could be quite expensive, since there are $(n-2)!$ possibilities in a complete graph of order n. Although there are good solutions for special classes of graphs, no good algorithm is known for determining such a hamiltonian cycle in the general case; in fact, the traveling salesman problem is known to be *NP-complete*. This means that it is not known if a good algorithm exists, but the existence of a good algorithm to solve this problem would imply the existence of good algorithms to solve many other outstanding problems, such as the graph isomorphism problem. Some of these problems will be mentioned in later sections. Although there is no known good algorithm which always gives a minimum solution, there are procedures which give reasonable solutions most of the time.

V. COLORINGS OF GRAPHS

A coloring (sometimes called a vertex coloring) of the vertices of a graph is an assignment of one color to each vertex. If k colors are used, it is called a k-coloring, and if adjacent vertices are given different colors, it is a proper coloring. The minimum number of colors in a proper coloring of a graph G is called the (vertex) *chromatic number* of G and is denoted by $\chi(G)$. The chromatic number of many special graphs is easy to determine. For example, $\chi(K_n) = n$, $\chi(C_n) = 3$ if n is odd, and $\chi(B) = 2$ for any bipartite graph B with at least one edge. Therefore, all paths, all cycles of even length, and all trees have chromatic number 2, since they are bipartite. In general, the chromatic number of a graph is difficult to determine; in fact, from an algorithm point of view it is an NP-complete problem just like the traveling salesman problem.

In a proper coloring of a graph the set of all vertices assigned the same color must be independent. A proper k-coloring is thus a way of partitioning the vertices V of a graph G into k independent sets, and thus, the chromatic number is the minimum number of independent sets in such a partition.

Consider the problem of storing n chemicals (or other objects) when certain pairs of these chemicals cannot be

stored in the same building. There is a natural model associated with this problem, which is a graph G of order n. The vertices are the chemicals, and the edges are pairs of objects that cannot be stored together. The chromatic number $\chi(G)$ is the minimum number of buildings needed to safely store the objects, since the buildings partition the objects into sets corresponding to independent sets of vertices.

The chromatic number $\chi(G)$ is related to other graphical parameters which are easy to determine, such as $\Delta(G)$, the maximum degree of a vertex of G. The maximum degree gives an upper bound on the chromatic number.

Theorem 5.1 (Brooks): *For any graph G, $\chi(G) \leq \Delta(G)+1$, and if G is a connected graph which is not an odd cycle or a complete graph, then $\chi(G) \leq \Delta(G)$.*

A maximal complete subgraph of a graph is called a *clique*. If a graph G contains a clique of order m, then clearly $\chi(G) \geq m$. However, the determination of the order of the largest clique in a graph is also an NP-complete problem. In addition, the chromatic number of a graph can be very large without the graph even containing a complete graph with three vertices or any small cycle. There is no relation between $\chi(G)$ and the largest clique of a graph G that would simplify the problem of determining $\chi(G)$. The *girth* of a graph G is the length of a shortest cycle in the graph. The following result indicates that a graph can be very sparse and still have a large chromatic number.

Theorem 5.2 (Erdős, Lovász): *For any pair of positive integers k and m, there is a graph with chromatic number at least k and girth at least m.*

This result is interesting, but the proof technique that was used may be of even more interest. The probabilistic method, which was introduced by Paul Erdős, is the basis for the proof. It its simplest form, this powerful technique merely counts the number of graphs which do not satisfy a certain property and verifies that this number is less than the total number of graphs. Therefore, there must be at least one graph with the desired property. This results in a proof of the existence of a graph, but it does not necessarily exhibit such a graph.

There is an edge analog to the vertex coloring of a graph. An assignment of a color to each edge of a graph is called an edge coloring, and it is called a k-edge coloring if at most k colors are used. A proper edge coloring is one in which adjacent edges are assigned different colors. The *edge chromatic number* $\chi_1(G)$ of a graph G is the minimum k for which there is a proper k-edge coloring of G. It should be noted that this concept applies to any multigraph without loops. A lower bound for $\chi_1(G)$ is $\Delta(G)$, since each of the edges incident to a fixed vertex must have different colors. An upper bound is given by the following result.

Theorem 5.3 (Vizing): *For any graph G, $\chi_1(G) = \Delta(G)$ or $\Delta(G)+1$.*

The theorem of Vizing implies that a graph G falls into one of two categories, depending on whether $\chi_1(G) = \Delta(G)$ or not. "Most" graphs are in the first category ($\chi_1 = \Delta$). In particular, all bipartite graphs are in this category. If n is odd, then the cycle has C_n edge chromatic number 3, but maximum degree 2, so there are graphs in the second category. No characterization for graphs in either of these two categories is known.

There are two scheduling problems that can be considered as graph coloring problems. We briefly describe each of the problems along with a solution. The first problem involves a vertex coloring and the second problem involves an edge coloring.

A schedule of classes is to determined that will accommodate the requests of a group of students. If each class is offered at a different time, then each student will be able to take the classes he or she wanted. This could result in a large number of class periods, with some being at very undesirable times. The problem is to determine the minimum number of time periods in which the classes can be scheduled so that each student will get the schedule he or she requested. Consider a graph G with the vertices being the classes to be scheduled. Two classes are joined by an edge if some student would like to take both classes. If two classes are scheduled at the same time, then no student requested to take both of these classes, so such classes must be independent in the graph G. This suggests that the chromatic number $\chi(G)$ is the minimum number of time periods needed for the scheduling. This fact is not difficult to verify.

A similar scheduling problem deals with the scheduling of teachers for classes. A collection of teachers T are to be scheduled to teach a set of classes C. The number of sections of each class for which each teacher is responsible is known. The problem is to determine a schedule which uses a minimum number of time periods. The assumption is made that, in any period, a teacher can teach only one class and each class is taught by only one teacher. Consider the bipartite graph B (actually it is a bipartite multigraph without loops) with T as the vertices in one part and C as the vertices in the other part. If a teacher t in T is scheduled to teach m sections of a class c in C, then place m edges between the vertex t and the vertex c. The construction of a schedule is really a proper edge coloring of the bipartite graph B using the time periods as colors. This observation can be used to verify that the minimum number of periods that must be used in the scheduling is the edge chromatic number $\chi_1(B)$, and this is the maximum degree $\Delta(B)$.

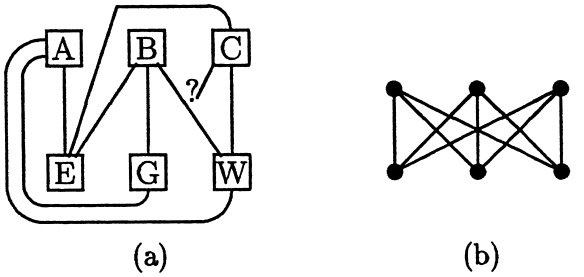


FIGURE 9 Houses-and-utilities problem.

VI. PLANAR GRAPHS

The houses-and-utilities problem is a classical puzzle whose solution can be described using graph theory. Suppose there are three houses (A, B, and C), and three utilities (E, G, and W for electricity, gas, and water) as pictured in Fig. 9a. The problem is to connect each of the utilities to each of the houses without any of the utility lines crossing. Pictured in Fig. 9a is an attempt to do this that failed. The graph theoretical form of this problem is to determine if the complete bipartite graph $K_{3,3}$ (Fig. 9b) can be drawn or pictured in the plane in such a way that the lines which represent the edges do not intersect except at the vertices of the graph. This leads to the concepts of plane and planar graphs.

A *planar graph* is one that can be drawn in the plane with points representing the vertices, and "nice smooth curves or lines" that do not intersect representing the edges. Once such a graph is drawn or embedded in the plane, it is called a *plane graph*. Thus, the three-utilities problem is equivalent to determining if $K_{3,3}$ is a planar graph. It is also a question that has a surprisingly uncomplicated answer, not only for the special graph $K_{3,3}$, but for any graph.

Figure 10 is a diagram of a plane graph G. Associated with this diagram are points (vertices of G), curves (edges of G), and regions (faces of G). The regions are the connected portions of the plane that remain after the points and curves of the diagram are deleted. Also note that one of these regions, namely R_4, is unbounded. If p, q, and r are the number of vertices, edges, and faces of a plane graph G, respectively, then $p = 6$, $q = 8$, and $r = 4$ for the graph in Fig. 10. An induction proof can be employed to verify that the following linear equality is always satisfied by these parameters.

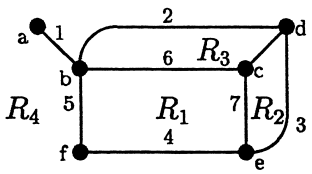


FIGURE 10 Plane graph.

Theorem 6.1 (Euler formula): *If G is a plane graph with p vertices, q edges, and r faces, then $p - q + r = 2$.*

The above result is a useful and powerful tool in proving that certain graphs are not planar. The boundary of each region of a plane graph has at least three edges, and of course each edge can be on the boundary of at most two regions. Thus, the number of edges and regions in a plane graph satisfies the inequality $r \leq 2q/3$. This inequality and Theorem 6.1 imply the relationship $q \leq 3p - 6$ between the number of vertices and edges in a planar graph. This type of reasoning and the observation that in a bipartite plane graph each region will have at least four edges on its boundary give the following result.

Theorem 6.2: *If G is a connected planar graph with p (≥ 3) vertices and q edges, then $q \leq 3p - 6$. If, in addition, G is a bipartite graph, then $q \leq 2p - 4$.*

The second inequality of Theorem 6.2 implies that the bipartite graph $K_{3,3}$ is not planar, since $9 > 2(6) - 4$. The first inequality of Theorem 6.2 implies that the complete graph K_5 is also not planar, since $p = 5$ and $q = 10$ for K_5. In some sense any graph that is not planar must contain either K_5 or $K_{3,3}$. We now describe what we mean by "in some sense." To *subdivide* an edge uv of a graph is to replace an edge with a vertex w and two edges uw and wv. This is like placing a vertex in the middle of an edge. A *subdivision* of a graph is obtained by successively subdividing the edges of a graph. Clearly, a subdivision of a graph is planar if and only if the graph is planar. We can now give a characterization of planar graphs.

Theorem 6.3 (Kuratowski): *A graph is planar if and only if it contains no subdivision of K_5 or $K_{3,3}$.*

One of the most famous problems in mathematics is the four-color problem, which was first proposed in 1852. Consider a map, such as the map of Europe. We would like to color the countries (which we will assume to be connected) so that two countries that have a common boundary (not just a point) will have different colors. Is it possible to color any such map with at most four colors? It is easy to show that six colors will suffice, since any planar graph must have a vertex of degree at most 5. Also, there is a clever but elementary proof which shows that five colors is sufficient. However, the question of whether four colors are enough remained an open question for many years. Many mathematicians (both professional and amateur) worked on this problem, and many incorrect proofs were generated.

A map M with five countries is pictured in Fig. 11a. Associated with this map is a planar graph G, whose

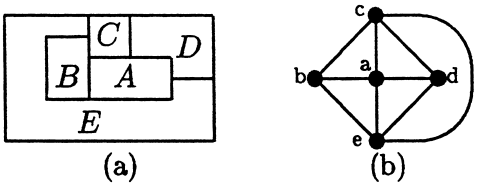


FIGURE 11 Map and plane graph.

vertices are the countries and whose edges join countries with a common boundary. An embedding of G is given in Fig. 11b. It is important to note that the graph G is planar. Coloring the map M so that adjacent countries have different colors is the same as giving a proper vertex coloring of the planar graph G of Fig. 11b. Associated with any map is a planar graph, and conversely, associated with a plane graph is a map. Thus, solving the four-color problem is equivalent to showing that no planar graph has chromatic number greater than four. In 1976 a positive solution to this problem was obtained by Appel and Haken by reducing the problem to a finite number of classes of planar graphs, and then by analyzing these cases with the aid of a computer. Since that time, the number of cases left to computer verification has been decreased, and also independent computer programs have been created to verify the remaining cases. However, all present proofs are still dependent on computer verification of a large number of cases.

Theorem 6.4 (Appel-Haken): *For any planar graph* G, $\chi(G) \leq 4$.

For many applications, planarity is a very desirable property to have. If the graph that represents an electrical circuit is planar, then the circuit can be printed on a board. Also, there are good algorithms (in fact, linear in the number of edges of the graph) for testing graphs for planarity and for embedding planar graphs in the plane.

For graphs which are not planar, there are several measures of how "nonplanar" they are. We mention three of the most common measures. One of these is the genus of a graph. The surface obtained from a sphere by adding m handles (like a handle on a coffee cup) is a *surface of genus* m. Embeddings for these surfaces is defined in the same way as it is for the plane: curves which represent edges must not intersect except at points which represent vertices. A graph G has *genus* m if it can be embedded in a surface of genus m but not in a surface of genus $m-1$. Any finite graph can be embedded in a surface of sufficiently large genus, so every graph has finite genus. Embedding a graph in the plane is equivalent to embedding the graph on a sphere. Since a sphere is a surface of genus 0, the planar graphs are precisely the graphs of genus 0.

If a graph G is not planar, then it can be "broken" into subgraphs which are planar. The minimum number of planar subgraphs of G, whose union of edges is all of the edges of G, is called the *thickness* of G. Therefore, a planar graph has thickness 1.

If a graph G is not planar, then any drawing of it in the plane would require that at least two of the edges "cross" at some point other than at a vertex. The minimum number of such crossings possible for some representation in the plane is called the *crossing number* of G. Clearly, planar graphs have crossing number 0.

In general, the determination of the genus, thickness or crossing number of a graph is an extremely difficult problem, even for special classes of graphs. For example, both the genus and thickness of all complete graphs are known, but the proofs are not easy. Also, the crossing number is not known for complete graphs with more than 10 vertices.

VII. FACTORIZATION THEORY

A spanning subgraph of a graph G is called a *factor* of G. If $E(G)$ is the edge disjoint union of the edges of factors of G, then these factors form a *factorization* of G. Thus a factorization is a partitioning of the edges of the graph. A factor of a graph which is r-regular is called an *r-factor*, and any factorization of a graph with r-factors is called an *r-factorization*. A hamiltonian cycle in a graph is an example of a 2-factor. Of particular interest are 1-factors of graphs, also called *perfect matchings*. A perfect matching of a graph of order n (clearly n must be even) thus consists of $n/2$ independent edges (edges which are not adjacent). A collection of independent edges, which may not necessarily span the graph, is called a *matching*. A matching with the most edges is called a maximum matching. In a cycle C_{2k} of even length the alternate edges in the cycle form a perfect matching in the cycle. There are thus two such perfect matchings, and they form a 1-factorization of the cycle.

Factorizations of complete graphs have been studied extensively. For example, K_{2n} has a 1-factorization. Thus K_{2n} will have an r-factorization if and only if $2n$ is a multiple of r. Of course K_{2n+1} cannot have a 1-factor because of its order, but it does have a 2-factorization. In fact, it has a factorization consisting of hamiltonian cycles. In Fig. 12a, a 1-factorization of K_4 is displayed, and a 2-factorization of K_5 is given in Fig. 12b. There are many other kinds of interesting factors and factorizations of complete graphs, such as factoring with paths or trees.

One of the earliest general results in this area involves a characterization of graphs which have 2-factorizations. Clearly, a graph G cannot have an r-factorization unless it is m-regular for some multiple m of r. In the case when $r = 2$, this condition is also sufficient.

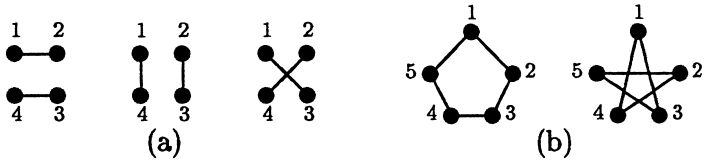


FIGURE 12 (a) 1-Factorization of K_4. (b) 2-Factorization of K_5.

Theorem 7.1 (Petersen): *A graph has a* 2*-factorization if and only if it is* m*-regular with* m *even.*

A similar characterization for 1-factorizations in bipartite graphs exists.

Theorem 7.2 (Kőnig): *A bipartite graph has a* 1*-factorization if and only if it is regular.*

Perfect matchings of graphs have been investigated more extensively than any other factors. Fortunately, there is a very useful characterization of bipartite graphs which have perfect matchings. If G is a bipartite graph with parts A and B, then a matching M is said to *saturate* A if each vertex in A is incident to an edge of M. Thus, if $|A| = |B|$, any matching which saturates A is a perfect matching. Also, if M is a matching which saturates A, then for any subset S of A, the neighborhood $N(S)$ of S (vertices not in S which are adjacent to some vertex of S) must have as many vertices as S, because just the matching M implies this. This condition is also sufficient to imply the existence of a perfect matching.

Theorem 7.3 (Philip Hall): *If* G *is a bipartite graph with parts* A *and* B*, then there is a matching which saturates* A *if and only if* $|N(S)| \geq |S|$ *for all* $S \subseteq A$.

A characterization of all graphs which have perfect matchings also exists, but it is more complicated. If S is a separating set for a graph G, then $G - S$ may have several components. Any component C with an odd number of vertices (called an odd component) cannot within itself have a perfect matching. So, any perfect matching of G would have at least one edge joining a vertex of C with a vertex of S. Thus, if G has a perfect matching, the number of odd components cannot exceed the number of vertices in S. This condition on the number of odd components is also sufficient for the existence of a perfect matching.

Theorem 7.4 (Tutte): *A graph* G *has a perfect matching if and only if for each* $S \subseteq V(G)$ *the number of odd components in* $G - S$ *does not exceed* $|S|$.

Although both Theorem 7.3 and Theorem 7.4 are useful results, they cannot be applied directly to obtain good algorithms for finding perfect matchings (or maximum matchings) in graphs. This is true because there are 2^n different subsets in a set with n elements. However, good algorithms do exist for finding maximum matchings in graphs. One such algorithm for bipartite graphs uses the idea of an alternating path. If M is a matching of a graph G, then an M-alternating path is a path whose edges alternate between edges in M and edges not in M. If the first and last vertices of the path are not in M, then the path is an M-augmenting path. If P is an M-augmenting path, then replacing the edges of P which are in M with the edges of P not in M will give a larger matching in G. Thus a maximum matching M never has an M-augmenting path. Berge showed that this was a necessary and sufficient condition for a matching to be maximum.

The algorithm based on this result searches for augmenting paths by constructing a tree rooted at a vertex which is not saturated by the the matching. Also, all of the paths in this tree are alternating. Either this tree will give an augmenting path, which will give a larger matching, or the matching will be a maximum matching. The condition of Hall can be shown to be violated if the maximum matching is not a perfect matching.

There are many applications of matchings in graphs. An obvious one is the *assignment problem*. Assume there are n workers and n jobs, and each worker can perform certain of the jobs. The assignment problem is to determine if it is possible to assign each person a job (two workers can not be assigned the same job), and if so, make the assignment. The graph model is a bipartite graph with the workers in one part and the jobs in the other part. An edge is placed between a worker and any job he can perform. The assignment problem reduces to finding a maximal matching in this bipartite graph. If this matching is a perfect matching, then an assignment can be made. Otherwise, it is not possible. Since there is a good algorithm for finding maximal matchings in bipartite graphs, there is a good algorithm for solving the assignment problem.

The assignment problem can also be considered when the number of jobs and the number of workers are not the same. However, for example, in this case when the number of workers exceeds the number of jobs, the objective is to assign all of the jobs. This will unfortunately leave some of the workers unemployed. The graph model is still the same bipartite graph, and the objective is to find a matching which saturates the vertices associated with the jobs. The same algorithm used when the number of jobs is the same as the number of workers applies in this case.

A generalization of the assignment problem is the *optimal assignment problem*. Here there are also n workers and

n jobs. However, in this case, the ith worker can perform the jth job with some efficiency c_{ij}. The problem is to assign each worker a job such that the sum of the efficiencies is a maximum. Again, a brute-force approach of examining all possible assignments is not efficient, since there are $n!$ such possibilities. However, there is a good algorithm for solving the optimal assignment problem which utilizes finding maximum matchings in a series of appropriately defined bipartite graphs. Associated with a given assignment of jobs there is a bipartite graph. A maximum matching is determined in this bipartite graph. If the maximum matching is not a perfect matching, then it can be used to obtain a better assignment which determines a new bipartite graph. If the matching is perfect, then the assignment is optimal and the algorithm terminates.

VIII. GRAPH RECONSTRUCTION

A famous unsolved problem in graph theory is the Kelly-Ulam conjecture. A graph G of order n is *reconstructible* if it is uniquely determined by its n subgraphs $G - v$ for $v \in V(G)$. In Fig. 13 there is an example of the four graphs obtained from single vertex deletions of a graph of order 4, and the graph they uniquely determine.

Reconstruction Conjecture (Kelly-Ulam): *Any graph of order at least* 3 *is reconstructible.*

The initial but equivalent formulation of the conjecture involved two graphs. If G and H are graphs with $V(G) = \{u_1, u_2, \ldots, u_n\}$ and $V(H) = \{v_1, v_2, \ldots, v_n\}$, and if $G - u_i \cong H - v_i$ for $1 \le i \le n$, then $G \cong H$. Note that to say that a graph G is reconstructible does not mean that there is a good algorithm which will construct the graph G from the graphs $G - v$ for $v \in V$. A positive solution to the conjecture might still leave open the question of the complexity of algorithms that would generate a solution to the problem.

Although it is not known in general if a graph is reconstructible, certain properties and parameters of the graph are reconstructible. It is straightforward to reconstruct from the vertex-deleted subgraphs both the size of a graph and the degree of each vertex. Let G be a graph of size q with vertices $\{v_1, v_2, \ldots, v_p\}$, and for each i let q_i be the size of the graph $G - v_i$. Each edge in G would appear in precisely $p - 2$ of the vertex deleted subgraphs, hence

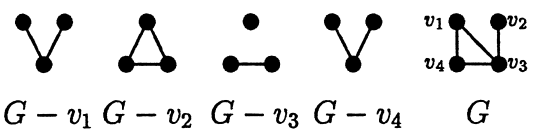


FIGURE 13 Graph reconstruction.

$$q = \left(\sum_{i=1}^{p} q_i \right) \Big/ (p - 2).$$

Also, clearly the vertex v_i has degree $q - q_i$. An immediate consequence of these facts is that any regular graph is reconstructible.

Several properties dealing with the connectedness of a graph are reconstructible, including the number of components of the graph. A subgraph of a graph is a *block* if it is a maximal 2-connected subgraph. The blocks of a graph partition the edges of a graph, and the only vertices that are in more than one block are the cut-vertices. If a graph has at least two blocks, then the blocks of the graph can also be determined. However, this does not mean the graph can be reconstructed from the blocks.

There are many special classes of graphs which are reconstructible, but we list only three well-known classes.

Theorem 8.1: *The following classes of graphs are reconstructible:*

(i) Disconnected graphs
(ii) Trees
(iii) Regular graphs

Corresponding to the "vertex" reconstruction conjecture is an *edge reconstruction conjecture*, which states that a graph G of size $m \ge 4$ is uniquely determined by the m subgraphs $G - e$ for $e \in E(G)$. Such a graph is said to be *edge-reconstructible*. Just as in the vertex case, the edge conjecture is open. The two conjectures are related, as the following result indicates.

Theorem 8.2 (Greenwell): *If a graph with at least four edges and no isolated vertices is reconstructible, then is is edge-reconstructible.*

Theorem 8.2 implies that trees, regular graphs, and disconnected graphs with two nontrivial components are edge reconstructible. There are also results which show that graphs with "many" edges are edge-reconstructible. For example, Lovász has shown that if a graph G has order n and size m with $m \ge n(n-1)/4$, then G is edge-reconstructible.

Intuitively, the edge-reconstruction conjecture is weaker than the reconstruction conjecture. This is confirmed by Theorem 8.2. However, there is another way of relating the two conjectures. Associated with each graph G is the line graph $L(G)$ of G. The vertices of $L(G)$ are the edges of G and two vertices of $L(G)$ (which are edges of G) are adjacent in $L(G)$ if and only if they were adjacent edges in G. The following result relates reconstruction and edge reconstruction.

Theorem 8.3 (Harary, Hemminger, Palmer): *A graph with size at least four is edge-reconstructible if and only if its line-graph is reconstructible.*

The line graphs of some special classes of graphs are easy to determine. For example, the line graph of a star $K_{1,n}$ is K_n, a complete graph, and the line graph of a cycle C_n is the cycle C_n of the same length. Therefore, the graphs K_3 and $K_{1,3}$ have isomorphic line graphs, namely, K_3. With this one exception, the line graphs of nonisomorphic connected graphs are also nonisomorphic. However, this does not imply that every graph is the line graph of some graph. In fact, there are numerous characterizations of line graphs. In particular, no graph which has an induced subgraph isomorphic to $K_{1,3}$ can be the line graph of a graph.

Since not every graph is the line graph of some graph, Theorem 8.3 does not imply that the edge reconstruction conjecture and the vertex reconstruction conjecture are equivalent.

IX. EXTREMAL THEORY

There are a large number of optimization problems and results in graph theory which could be considered in a discussion of extremal graph theory. However, we will restrict our consideration to the Turán-type extremal problem of determining the maximum number of edges a graph can have without containing a certain subgraph. For example, consider the problem of determining how many edges there can be in a graph of order n which does not contain a K_3 (triangle). For n even, the complete bipartite graph $K_{n/2,n/2}$ has $n^2/4$ edges and no triangle, and, in fact, no odd cycle of any length. Also, any graph with n vertices and $(n^2/4)+1$ edges (for n even) can be shown to contain a triangle. Thus, when n is even, the maximum number of edges in a triangle-free graph of order n is $n^2/4$.

The original problem of Turán, who initiated this kind of investigation, was to determine the maximum number of edges a graph of order n can have without containing a complete graph K_p. The case $p=3$ was just considered. The general case, while more complicated, is entirely parallel.

Multipartite graphs are a class of graphs which occur in this type of study. A graph G is called a *k-partite graph* if the vertices V can be partitioned into k parts such that the only edges in G join vertices in different parts. Thus, bipartite graphs are simply 2-partite graphs. If all edges between each pair of k parts are in the graph, the graph is called a *complete k-partite graph*. In any k-partite graph the vertices in each part are independent, and if the graph is a complete k-partite graph then the chromatic number of the graph is clearly k. Let $T(n, k)$ denote the complete k-partite graph of order n in which the difference in the number of vertices in any pair of parts is at most 1. That is, if $n = km + r$ with $0 \le r < m$, there will be r parts with $m+1$ vertices and the remaining $k-r$ parts will have m vertices. Let $t(n, k)$ denote the size of the graph $T(n, k)$. Turán proved the following result.

Theorem 9.1 (Turán): *The maximum number of edges in a graph of order n which does not contain a K_p is $t(n, p-1)$, and the graph $T(n, p-1)$ is the only graph of that order and size with no K_p.*

For a fixed graph F, which we will call the forbidden graph, the extremal number $\mathrm{ex}(n, F)$ is the maximum size of a graph of order n which does not contain an F as subgraph. The collection of graphs of order n and size $\mathrm{ex}(n, F)$ are called the extremal graphs for F and is denoted by $\mathrm{Ex}(n, F)$. The extremal problem consists of finding $\mathrm{ex}(n, F)$ and also $\mathrm{Ex}(n, F)$, if possible. In the case when the forbidden graph is K_p, the result of Turán states that $\mathrm{ex}(n, K_p) = t(n, p-1)$ and $\mathrm{Ex}(n, Kp) = \{T(n, p-1)\}$. Note that

$$t(n, p-1) = \left(\frac{p-2}{p-1}\right)\binom{n}{2}$$

when n is a multiple of $p-1$ and is close to that for all other values of n. If p is thought of as the chromatic number of the complete graph K_p, then the extremal result for an arbitrary graph F is closely related to the result for the complete graph. In fact, the chromatic number $\chi(F)$ determines the order of magnitude of the extremal number if the graph is not bipartite. In the following result, an "error function" is needed. Denote by $o(n^2)$ a function of n with the property that $\lim_{n\to\infty} o(n^2)/n^2 = 0$.

Theorem 9.2 (Erdős-Simonovits): *For any nontrivial graph F with chromatic number $k \ge 3$,*

$$\mathrm{ex}(n, F) = \left(\frac{k-2}{k-1}\right)(n^2) + o(n^2).$$

Also, any graph in $\mathrm{Ex}(n, F)$ can be obtained from $T(n, k-1)$ by deleting or adding at most $o(n^2)$ edges.

If F is not a bipartite graph, then the first term in the expression for $\mathrm{ex}(n, F)$ in Theorem 9.2 is of order n^2, so the error function $o(n^2)$ becomes insignificant for sufficiently large values of n. However, when F is a bipartite graph, this first term is 0, and thus Theorem 9.2 gives little information. The extremal problem for a bipartite graph is called the *degenerate extremal problem*. In this case, there are many interesting open questions, and no general asymptotic result like Theorem 9.2 is known in the degenerate case. Examples which give lower bounds for the number $\mathrm{ex}(n, F)$ are difficult to find when F is a bipartite graph, and in many cases involve designs. We mention two results which give upper bounds. Sharp lower bounds are not known for either of the classes of graphs considered

below, except for a few small-order cases such as C_4, C_6, C_{10}, and $K_{3,3}$.

Theorem 9.3 (Kővári-Sós-Turán): *If $r \leq s$, then there exists a constant c_{rs} (depending only on r and s) such that*

$$\text{ex}(n, K_{r,s}) \leq c_{rs} n^{[2-(1/r)]}.$$

Theorem 9.4 (Erdős): *There is a constant c_k such that* $\text{ex}(n, C_{2k}) \leq c_k n^{[1+(1/k)]}$.

In the special case when the forbidden bipartite graph F is a tree, more is known about $\text{ex}(n, F)$. When $p-1$ divides n, a disjoint union of complete graphs K_{p-1} implies that $\text{ex}(n, Tp) \geq (p-2)n/2$ for any tree with p vertices. Erdős and Sós conjectured that $(p-2)n/2$ is an upper bound as well for all values of n. This is known to be true for stars and paths.

The extremal problem can be generalized to consider a family of forbidden graphs, not just one graph. However, the extremal results for a family of forbidden graphs is essentially the same as for one forbidden graph. The important parameter in the case of a family of graphs is the smallest chromatic number of a graph which is in the family. Therefore, if there is at least one subgraph in the family which is bipartite, then the extremal number will be $o(n^2)$. Also, if a graph G of order n has more than $\text{ex}(n, F)$ edges, then it has at least one copy of F. An interesting problem is to determine the number of copies of F that G must have as a function of the number of edges k in G.

An area of extremal theory which predates theory related to the Turán type of theory is *Ramsey theory*. In Ramsey theory, pairs of "forbidden" graphs are considered, say, F_1 and F_2, and one is interested in determining whether for each graph G of order n, F_1 is a subgraph of G, or F_2 is a subgraph of the complement $\bar{G}$ of G. A consequence of a result of Frank Ramsey proved in 1930 is that this is always true if n is sufficiently large. The smallest n for which it is true is called the Ramsey number of the pair (F_1, F_2) and is denoted by $r(F_1, F_2)$. In general, these numbers are difficult to determine. In the classical case when F_1 and F_2 are complete graphs, only a few of the Ramsey numbers are known; in fact, Table I gives all of the known Ramsey numbers $r(K_m, K_n)$ for $3 \leq m \leq n$.

TABLE I Ramsey Numbers $r(K_m, K_n)$

$m \backslash n$	3	4	5	6	7	8	9
3	6	9	14	18	23	29	36
4		18					

X. DIRECTED GRAPHS

A *directed graph* or *digraph* D is a finite collection of elements, which are called vertices, and a collection of ordered pairs of this vertices, which are called arcs. Thus, a digraph is similar to a graph except that each arc in a digraph has a direction, while an edge in a graph does not. Just as in the case for graphs, the vertices and arcs of D will be denoted by V and E, respectively. However, in this case the arc $e = uv$ does not join vertices u and v, but it joins u to v. Thus, $uv \neq vu$. The terms adjacent with and incident with will be replaced by the terms adjacent to, adjacent from, incident to, and incident from. The order and size of a digraph is the same as for a graph.

Most of the concepts for graphs have obvious analogs for digraphs, for example, *subdigraph* and directed walks of various types. However, the degree of a vertex v is split into two parts: $d^+(v)$ is the number of arcs incident from v and is called the *outdegree*, and $d^-(v)$ is the number of arcs incident to v and is called the *indegree*. The degree $d(v) = d^+(v) + d^-(v)$. Associated with each digraph D is an underlying graph G obtained from D by removing the directions from the edges of D and then removing any one of any pair of multiple edges that are produced from removing the directions.

The concept of connectivity is more complicated for directed graphs, and several types of connectivity arise. We will mention two that are most commonly used. A digraph D is connected if the underlying graph is connected. A vertex v is *reachable* from a vertex u if there is a directed path from u to v. The digraph is *strongly connected* or *diconnected* if each pair of distinct vertices is reachable from each other. Clearly, diconnectedness implies connectedness, but the digraph in Fig. 14a shows that the two concepts are not the same. Just as connectivity can be used to partition the vertices of a graph into components, diconnectedness can be used to partition the vertices

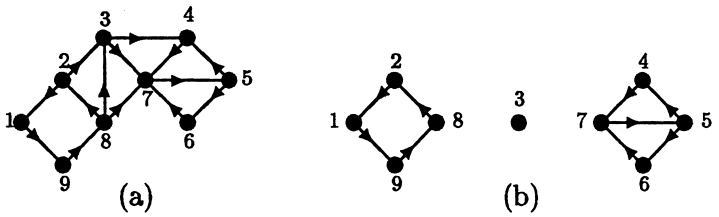


FIGURE 14 (a) Digraph. (b) Components.

of a digraph into dicomponents. Any pair of vertices in the same dicomponent are mutually reachable. The dicomponents of the digraph D in Fig. 14a are shown in Fig. 14b. Thus, the digraph in Fig. 14a is an example of a graph which is connected, but yet it has three diconnected components.

The lengths of paths or cycles in the underlying graph of a digraph give no information about the length of directed paths or cycles. However, it is surprising that the chromatic number of a digraph (which is the same as the chromatic number of the underlying graph) does provide some information, as the following two results indicate. The first of these results can be obtained as a corollary of the second.

Theorem 10.1 (Gallai-Roy): *Every digraph D contains a directed path of length at least $\chi(D)$.*

Theorem 10.2 (Bondy): *Every diconnected digraph D contains a cycle of length at least $\chi(D)$.*

Just as in the case for graphs, there are natural questions concerning the existence of special subdigraphs in digraphs. Conditions which imply the existence of eulerian directed trails, cycles, or other subdigraphs in digraphs are many times analogs of conditions in graphs. The following are just two such examples from among many.

Theorem 10.3 *A digraph D has a closed eulerian directed trail if and only if $d^+(v) = d^-(v)$ for all $v \in V$.*

Theorem 10.4 (Woodall): *If D is a digraph of order n such that $uv \notin E(D)$ implies that $d^+(u) + d^-(v) \geq n$, then D has a directed hamiltonian cycle.*

Consider a city which has had both one-way and two-way streets, but because of the additional traffic flow must make each street one-way. Is it possible that the direction of the streets can be chosen so that a motorist can get from any one point in the city to any other point in the city? Given a graph G, an *orientation* of the graph is an assignment of a direction to each of the edges of the graph. Thus, the oriented graph obtained in this way is a digraph. The graph theory form of the initial problem is to determine for which graphs there is an orientation which makes the resulting digraph diconnected. If the graph contains a *bridge* (an edge which disconnects the graph), then clearly no such orientation exists. The lack of such a bridge, which means that at least two edges must be deleted to disconnect the graph and so the graph is 2-edge-connected, is also a sufficient condition.

Theorem 10.5 (Robbins): *Every 2-edge-connected graph has an orientation which gives a diconnected digraph.*

A digraph obtained from orientating a complete graph is called a tournament. See Fig. 15 for an example of a

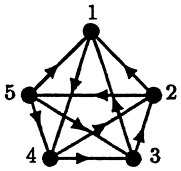


FIGURE 15 Diconnected tournament.

tournament which is diconnected and in fact has a directed hamiltonian cycle. The use of the term "tournament" is a natural one. If a "tennis tournament" were played with each pair of participants playing each other (round robin tournament), then the results of the tournament could be modeled by this type of directed graph. A directed edge from u to v indicates that u defeated v. Tournaments are a class of digraphs that has been studied extensively. If T is a tournament of order n, then the chromatic number $\chi(T)$ is n. Therefore Theorem 10.1 and Theorem 10.2 have the following immediate consequences.

Theorem 10.6 (Re'dei): *Every tournament has a directed hamiltonian path.*

Theorem 10.7 (Camion): *Every tournament of order at least 3 has a directed hamiltonian cycle if and only if it is diconnected.*

Theorem 10.6 implies that in any round-robin tournament the players can always be ordered such that the ith player defeated the $(i+1)$th player. In this case the first player could be considered the "best" player. However, the hamiltonian path is not necessarily unique, so there could be another ordering which would give a different first person who could also claim to be the "best." Theorem 10.7 implies that the players can be cyclically ordered such that the ith player defeated the $(i+1)$th if and only if for each pair of players A and B there is a sequence from A to B such that each player defeated the next player in the sequence. In this case any person could make an argument for being the "best" player in the tournament.

Most of the algorithms dealing with graphs have digraph analogs. Also, in general, the complexity of the corresponding algorithms are approximately the same. If there is a good algorithm in the graph case, then the corresponding digraph algorithm will usually be good, and if the graph algorithm is NP-complete, then the corresponding digraph algorithm will be NP-complete. For example, there is a good algorithm to find the dicomponents of a digraph just as there is a good algorithm to find the components of a graph. The existence of hamiltonian cycles is NP-complete in both the directed and undirected case.

XI. NETWORKS

The term "network" is used in several different contexts and has many meanings. For example, there are communication networks and there are electrical networks. We will generally restrict our consideration to one kind of network, which we now define. The purpose is to build a model for determining how to maximize the flow of goods from one point to another in some transportation system. A *network* is a digraph D with two specified vertices s (which is called the *source*) and t (which is called the *sink*), and a function c (called the capacity) which assigns to each arc of the network a non-negative integer. A *flow* on D is a function f from $E(D)$ into the non-negative integers such that the following two conditions are satisfied:

$$f(e) \le c(e) \qquad \text{for all} \quad e \in E, \quad \text{and}$$

$$\sum_{uv \in E} f(uv) = \sum_{vw \in E} f(vw) \qquad \text{for all} \quad v \in D - s, t.$$

The first condition merely states that the flow cannot exceed the capacity of the arc, and the second condition requires that the flow into an intermediate vertex is the same as the flow out of the vertex. The object is to find a maximum flow from the source s to the sink t. The value to be maximized is the net flow out of s, which is $\sum_{sv \in E} f(sv) - \sum_{vs \in E} f(vs)$. An example of a network with a flow is given in Fig. 16. The first number of the pair by each arc is the flow and the second number is the capacity. The net flow in this example is 4.

A *cut* in a network D is a partition of the vertices V into two subsets S and $\bar{S}$ with $s \in S$ and $t \in \bar{S}$. The capacity of the cut is the sum of the capacities of the edges from S to $\bar{S}$. The following is called the max-flow min-cut theorem.

Theorem 11.1 (Ford-Fulkerson): *In a network the value of a maximum flow is equal to the capacity of a minimum cut.*

There is a good algorithm for finding a maximum flow in a network. It is based on starting with any flow—say, the 0 flow—and successively increasing this flow by finding a path (with possibly some arcs reversed) from source to the sink on which the flow can be increased. If at any step this cannot be done, the algorithm generates a cut whose capacity is equal to the flow. This, of course, implies that the flow is maximum.

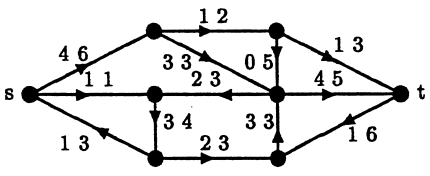


FIGURE 16 Network.

The networks considered here were single-source and single-sink networks. This is not a restriction, since it is simple to go from a multiple-source and multiple-sink network to a network with a single source and a single sink. A new network can be constructed by adding a "super source" which is adjacent to all of the sources and a "super sink" which is adjacent from all of the sinks. A solution to the maximum flow problem for this single-source and single-sink network will give a solution to the maximum flow problem for the multiple-source and multiple-sink network.

There are many applications and consequences to the max-flow min-cut theorem. The matching result of Hall and the various forms of Menger's theorems for both graphs and digraphs can be proved using this result.

Digraphs with labels on the directed edges, such as the network just described, can be used as models for many situations. In operations research, activity networks are useful as models for the planning of activities of large projects. A digraph with a single source and sink (representing the beginning and end of the project) will have directed edges which represent the activities of the project. The directed edges also indicate which activities must be completed before others can be started. The label on each edge will represent the time required to complete that activity. Algorithms which calculate path lengths can be used to determine optimal schedules for such projects and to determine which activities are critical in keeping such a schedule.

XII. RANDOM GRAPHS

Paul Erdős first introduced probabilistic methods to graph theory in the 1940s, and this powerful technique of studying "random graphs" has had an extraordinary impact on the development of graph theory. There are several models for studying random graphs, but we will just discuss one. Consider the collection of all labeled graphs (the vertices are distinguishable) of order n on the vertices $\{1, 2, \ldots, n\}$. The complete graph on the set of vertices has $N = \binom{n}{2}$ edges, and the number of different labeled graphs is 2^N, since there is the choice of either choosing or rejecting each of the N edges. Given a positive number p with $(0 \le p \le 1)$, this collection of graphs becomes a probability space, which we will denote by $\mathcal{G}(n, p)$, by choosing each edge with independent probability p. Therefore the probability that a *random graph* in $\mathcal{G}(n, p)$ is a fixed graph G that has m edges is $p^m(1-p)^{N-m}$.

Using the probability space $\mathcal{G}(n, p)$, it is possible to prove the existence of a graph with certain properties without actually exhibiting the graph. This is done by counting the number of graphs that do not satisfy a

specified property and showing that this number is smaller than the total number of graphs. An example of this type of result is the following result of Erdős concerning an exponential lower bound for Ramsey number $r(K_s, K_s)$.

Theorem 12.1 (Erdős): *There is a positive constant c such that $r(K_n, K_n) > cs^{s/2}$.*

The lower bound of the previous theorem has been improved using random graphs (probabilistic techniques), but at this point no graphs have actually been described that satisfy the bound. It was mentioned earlier that given positive integers k and m there is a graph G with the chromatic number $\chi(G) \geq k$ and girth $g(G) \geq m$. Again, the major tool used in this proof is the probabilistic method.

If R is a graphical property, then the probability that a graph G in $\mathcal{G}(n, p)$ has property R will be denoted by $P[R \cap \mathcal{G}(n, p)]$. We say that *almost all graphs* have property R if $\lim_{n\to\infty} P[R \cap \mathcal{G}(n, p)] = 1$. There is the corresponding definition for *almost no graphs*. As the value of p increases, the expected number of edges of a random graph G in $\mathcal{G}(n, p)$ will also increase. Thus, if p is a fixed positive real number, then the expected number in edges in a random graph G in $\mathcal{G}(n, p)$ is $p\binom{n}{2}$, which is substantial. Therefore, these graphs are dense, so it is not surprising that the following is true.

Theorem 12.2: *If p is a fixed positive number $(0 < p < 1)$, then for any fixed positive integer k, almost all graphs [in the probability model $\mathcal{G}(n, p)$] have the following properties (among others):*

(i) Diameter 2
(ii) Contain any finite subgraph H as an induced subgraph
(iii) k-Connected
(iv) Hamiltonian
(v) Genus exceeding k
(vi) Chromatic number exceeding k

It is obvious that a p increases, the number of edges of graphs in a random graph in $\mathcal{G}(n, p)$ increases, but it is not obvious that random graphs undergo significant structural changes as the number of edges increases. This behavior can be described in terms of *threshold functions* for a graphical property in the probability space $\mathcal{G}(n, p)$ for the probability $p = p(n)$. If R is a graphical property then $t(n)$ is a threshold function for property R when $\lim_{n\to\infty} p(n)/t(n) = \infty$, then almost all graphs have property R and when $\lim_{n\to\infty} p(n)/t(n) = 0$, then almost no graphs have property R. The following two results are examples of threshold functions for the property of being "connected" and the property of being "hamiltonian."

Theorem 12.3: *Assume that $\lim_{n\to\infty} g(n) = \infty$, and let $p_1(n) = [\ln n - g(n)]/n$ and $p_2(n) = [\ln n + g(n)]/n$. Then, almost no graphs in $\mathcal{G}[n, p_1(n)]$ are connected, and almost all graphs in $\mathcal{G}[n, p_2(n)]$ are connected.*

The previous theorem implies that $t(n) = (\ln n)/n$ is the threshold function for connectivity, and the next result implies that $t(n) = [(\ln n)/n + \ln\ln n]/n$ is the threshold function for hamiltonicity.

Theorem: 12.4: *Assume that $\lim_{n\to\infty} g(n) = \infty$, and let $p_1(n) = [\ln n + \ln\ln n - g(n)]/n$ and $p_2(n) = [\ln n + \ln\ln n + g(n)]/n$. Then, almost no graphs in $\mathcal{G}[n, p_1(n)]$ are hamiltonian, and almost all graphs in $\mathcal{G}[n, p_2(n)]$ are hamiltonian.*

There are corresponding threshold functions results for a wide range of graphical properties.

SEE ALSO THE FOLLOWING ARTICLES

ALGEBRA, ABSTRACT • DISCRETE MATHEMATICS AND COMBINATORICS • GRAPHICS IN THE PHYSICAL SCIENCES • PROBABILITY

BIBLIOGRAPHY

Alon, N., and Spencer, J. (1992). "The Probabilistic Method," Wiley, New York.

Beineke, L. W., and Wilson, R. J., eds. (1983). "Selected Topics in Graph Theory 2," Academic Press, London.

Beineke, L. W., and Wilson, R. J., eds. (1988). "Selected Topics in Graph Theory 3," Academic Press, London.

Biggs, N. L., Lloyd, E. K., and Wilson, R. J. (1976). "Graph Theory 1736–1936," Clarendon Press, Oxford, U.K.

Bollobás, B. (1998). "Modern Graph Theory," Springer-Verlag, New York.

Bollobás, B. (1978). "Extremal Graph Theory," Academic Press, New York.

Chartrand, G., and Lesniak, L. (1996). "Graphs and Digraphs," 3rd ed., Chapman & Hall, London.

Chartrand, G., and Oellermann, O. (1993). "Applied and Algorithmic Graph Theory," McGraw-Hill, New York.

Clark, J., and Derek Holton, D. (1991). "A First Look at Graph Theory," World Scientific, Singapore.

Gould, R. (1988). "Graph Theory," Benjamin/Cummins, Menlo Park, CA.

Graham, R. L., Grötschel, M., and Lovász, L., eds. (1995). "Handbook of Combinatorics, I, II," MIT Press, Cambridge, MA, and North Holland, Amsterdam, The Netherlands.

Palmer, E. (1985). "Graphical Evolution," Wiley, New York.

West, D. B. (1996). "Introduction to Graph Theory," Prentice-Hall, Upper Saddle River, NJ.

Wilson, R. J., and Watkins, J. (1991). "Graphs, An Introductory Approach," Wiley, New York.



Graphics in the Physical Sciences

Daniel B. Carr
George Mason University

I. Univariate Guidelines
II. Bivariate Guidelines
III. Multivariate Visualization
IV. Closing Remarks

GLOSSARY

Box plot Graph displaying the outliers, adjacent values, quartiles, and median. The distribution caricature is especially useful as a terse summary and for comparing the distributions of different groups.

Conditioned plots Multiple juxtaposed panels where only the cases that satisfy constraints associated with a panel appear in or define the contents of the panel. Controls for the variation in related variables. Useful for reducing three-, four-, and five-dimensional relationships to a sequence of planar views.

Conditioned maps Multiple panels of juxtaposed maps where only the regions with variables that satisfy constraints associated with a panel appear highlighted in panel. Useful or controlling the variation of a variable display on a map due to one or more independent variables.

Glyph plots Graphs that encode variables as something other than position along an axis. Sometimes useful for showing several variables.

Linked plots A set of panels representing different variables of the same cases. The representation of the variables of the same case are connected by lines, color, point position, or other features that allow matching. Useful for assessing higher-dimensional relationships using mode dimensional views.

Mean difference plot A scatterplot that plots the difference of paired values on the y-axis and the mean of pair values on the x-axis.

Multidimensional scaling An algorithmic method that represents points in a lower-dimensional space while attempting to preserve the distances between all point pairs. An infrequently used method for embedding data in lower dimensions.

Paired comparison plot A scatterplot of values that have been paired because they are measurements on the same entity or process in two different circumstances.

Parallel coordinate plots Graph with parallel axes for each of the variables. Lines connect the coordinates of the same case between adjacent axes. Is especially useful in showing many variables, and graphically highlighting subsets of data.

Principal components A method of constructing linear combinations of the data to produce new orthogonal variables. The first new variable has the largest variance. The second new variable has the second largest variance and so on. Commonly used for approximating the data in a lower-dimensional space by dropping the last of the new variables.

Quantile plot Graph that relates the ordered observations in a batch of data to cumulative probabilities.

Quantile–quantile plot Graph for comparing two distributions. Good for several tasks like judging relative tail thickness, ratios of standard deviations, and changes in mean.

Scatterplot matrix Graph displaying scatterplots of all pairs of variables, arranged in square array with individual variables corresponding to rows and columns. Especially useful in the early stages of studying joint relationships across several variables.

Three-dimensional scatterplot Graph with points plotted against three orthogonal axes. Perceived via a stereo pair view or via motion parallax. Especially useful in seeing three-dimensional structure. Useful as a framework for showing additional coordinates using glyphs.

THE PURPOSE of graphics in the physical sciences is to help scientific and public understanding of the phenomena being studied. The graphics and design issues for the respective audiences differ. This article emphasizes graphics for scientists who are often interested in the discovery and analysis of data.

Scientists involved in discovery can face many challenges. These may include

- The complexity of the phenomena being studied,
- The difficulty of parsimoniously conceptualizing this complexity,
- The logistic and political impediments to collecting adequate, representative data,
- The limits of computational resources,
- The limits of human perception and cognition for understanding multivariate summaries.

Graphics can help in conceptualizing and characterizing the phenomena being student. The Crick and Watson discovery of the DNA's double-helix structure was a major step forward. Now the exponentially expanding studies of genes can boggle the mind. Graphical bookkeeping (for example, see biochemical pathways at http://www.expasy.ch/cgi-bin/show thumbnails.pl.) helps scientists cope with the ever-increasing details. The possibilities of combinatorial chemistry and genetics are hard to imagine. The physical scientists monitoring the earth also face huge datasets. Consider one multispectral dataset with 30-m resolution over the 8 million km^2 of the continental United States. A quick calculation indicates that it takes over 7000 screen images to examine this data on a 1024×1280 monitor. That encodes all of the multispectral information for a 30-m region into the color of a pixel. To appreciate the multispectral resolution requires many times more images. The human visual system may be able to handle on the order of 10 ~7 bits per second of visual input, so flashing through the images may not take all that long. However, encoding the information in a way that helps scientists is a real challenge. Norretranders estimates that consciousness is limited to about 16 bits a second. If care is not taken none of what is seen will get into those precious 16 bits or whatever gets in might be distorted.

The brief monochrome article cannot do justice to methods available for quantitative visualization in the physical sciences. One book of over 400 pages catalogues monochrome chart types. The possibilities grow combinatorially from this when considering the addition of color, texture, interactivity, and further composition of forms.

This article presents some of the basic methods associated with the field of statistical graphics, along with a discussion of related cognitive and perceptual principles that help to provide guidance when moving beyond the basic forms. The principles are, of course, not correct in every detail, but provide a reasonable basis for selecting among the options. Before proceeding, a few comments on data and estimates are appropriate.

The varied sources of physical science data include laboratory studies, field samples, and electronically collected data such as satellite imagery and computer simulation. Most data need to be transformed into estimates that are suitable for scientific analysis. Each type of data and collection circumstances poses it own issues in addressing the processes of producing estimates that are worthy of evaluation. Recurrent types of issues include calibration of instruments, scaling of variables, adequacy of variables as surrogates for the desired variables of interest, adjustments for covariates, representative coverage of the population or phenomena of interest, and validation of simulated or indirect estimates.

Graphics can be no better than the estimates presented (unless they are being used to show how bad the estimates are). Readers should look for indications of estimate quality either in the accompanying text or in the graphics. The display of confidence bounds for estimates provides a good indication. The lack of confidence bounds is sometimes a warning that estimates have not been assessed with respect to accuracy (bias) and precision (variability).

The heart of graphics is visual comparison. One common task is to assess estimated density patterns. This involves comparing local densities from different parts of a plot. Another task compares data to a conjectured functional relationship. Other common tasks include comparing estimates against reference values and comparing two or more sets of estimates against each other. Much of graphics design concerns facilitating accurate comparisons.

I. UNIVARIATE GUIDELINES

Assuming accurate comparison is important, it is advantageous to encode estimates so that humans can decode them accurately. Cleveland and McGill (1984) discuss perceptual accuracy of extraction and indicate preferred methods for univariate encoding. In this research subjects judged relative magnitudes of graphically encoded variables. As presented here the results put the graphical encoding methods into three classes, labeled best, good, and poor.

The two best encoding methods represent variables using position along a common scale as shown in Fig. 1 and position along identical nonaligned scales. Humans do well in judging the position of a point relative to scale. Locating the position of objects is a fundamental visual task. Mapmakers have long used position along a scale as the fundamental encoding for spatial coordinates. Length, angle, and orientation are good encodings. Figure 2 shows that transforming line segments into a standard position converts the task of judging length into a task of judging the position of one endpoint against a scale. While this is not necessarily what people do, the example suggests that judging line length is more complicated than judging position.

Figure 3 shows angle encoding. Rotation of the angles puts them in position for comparison against equivalent angular scales shown in gray. The transformation suggests that while angle comparisons work pretty well, they are more complicated than direct comparison against angle scales. Area, volume, point density, and color saturation are poor encodings. Those familiar with experimental results involving Steven's Law will not be surprised about poor results for the area and volume encodings. Steven's Law states that the perceived magnitude of a stimulus follows a power law.

$$p(x) = a\,x^{b}$$

where x is the magnitude of the true stimulus (for example, length, area, or volume), and where the constants a and b depend on the type of stimulus. The range of characteristic exponents b for length is [.9 , 1.1], area s

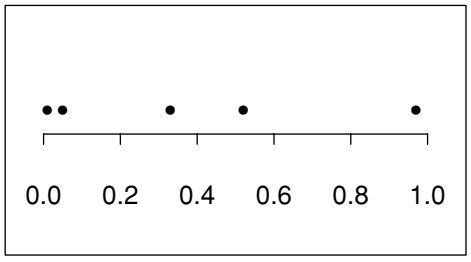


FIGURE 1 Position along a scale. This is the preferred encoding for a continuous variable.

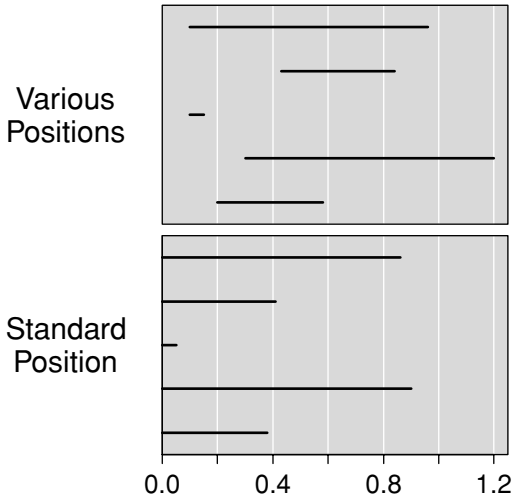


FIGURE 2 Transforming lengths. This changes the decoding task from taking differences to judging position along a scale.

[.6, .9] and volume [.5, .8]. With an exponent of about 1, people's perception of length tends to be direction proportional to object length. However, we tend judge area and volume nonlinearly. Consider comparing areas, one of 4 square units and the other of 1 square unit. With an exponent of .75, the ratio of perceived magnitudes is 2.8 to 1, rather than convert 4 to 1. We underjudge the large areas relative to small areas. If everyone had the same exponent, graphical encoding could adjust for systematic human bias. However, the range of values for the exponent "b" vary substantially from person to person. Providing a

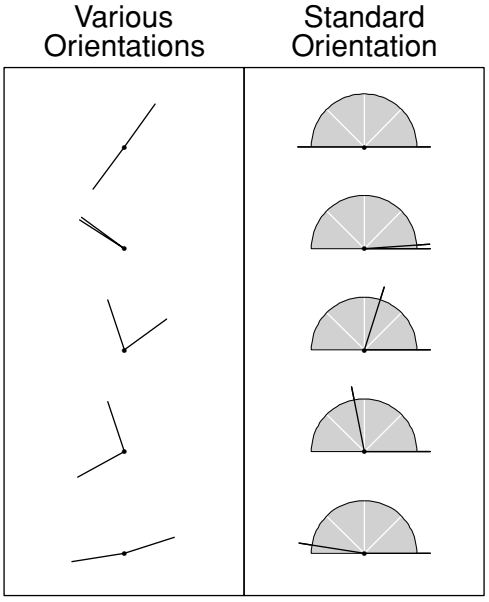


FIGURE 3 Transforming angles. This reduces the decoding task from taking differences to judging position along an angular scale.

set of reference symbols in a legend helps people calibrate to the intended interpretation, but the preferred strategy is to use better encodings whenever possible.

Weber's law, a fundamental law in human perception, also has significant ramifications in terms of accurate human decoding. A simple example gives the basic notion of the law. The probability of detecting that a 1.01-in. line is longer than a 1-in. line is nearly the same as the probability of detecting that a 1.01-ft line is longer than a 1-ft line. In absolute terms .01 in. is much smaller than .01 ft. Using a finer resolution scale allows more accurate judgments on an absolute scale. For example the tic marks on a ruler help us make more accurate assessments. The graphical equivalent uses grid lines to provide a finer resolution scale for more precise comparisons. In interactive graphics, zooming in effectively provides a finer scale. Computer human interface implementations often provide sliders that allow the reader to change the reference scale and make accurate judgments.

We render most graphics on a plane. We could show values of a continuous variable as points on a line. However, we do not usually do so. We may see gap between points quite well, but we do not judge point density very well. Consequently the standard approach for this simple case is to estimate the density and show it using a bivariate plot.

II. BIVARIATE GUIDELINES

Edward Tufte notes that it took over 5000 years to mankind to generalize from early clay tablets maps to representing general variables using a scatterplot. A scatterplot is an excellent representation since the two orthogonal axes allow two coordinates to be independently encoded as position along a common scale. The popular press in the United States still considers the scatterplot too complicated for the general public, but in the sciences the scatterplot is the standard for representing continuous bivariate data. Common bivariate activities include assessing univariate distributions, comparing univariate distributions, and looking for functional relationships.

A. Univariate Density Estimates

When we are interested in data density it is advantageous to compute and show the density as accurately as possible. Figure 4 illustrates the construction of a kernel density estimate based on a sample of five univariate values. The locations of the white triangles relative to the x-axis indicate the magnitudes of observed values. The basic idea is that each observed value is a surrogate for values in a neighborhood. We then construct a relative likelihood for a neighborhood about the value. Figure 4 shows the five

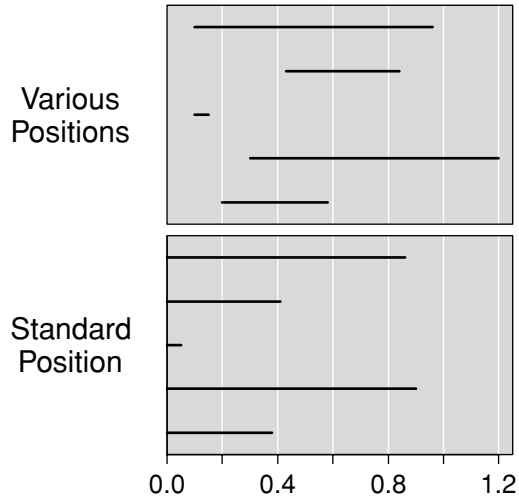


FIGURE 4 Density estimate. This shows construction of a kernel density estimate from five points by averaging bumps.

likelihoods (or kernels) as bell-shaped curves, one in each of the top panels. For each location where we want to estimate the data density (the x locations of white lines in Fig. 4) we simply average values from the five constructed likelihoods. The white lines indicate the locations of the density estimates. In the bottom panel each white line is the average height of all the white lines directly above it. (When panels have no white lines directly above, they contribute zero to the average.) The construction is straightforward.

Scott (1992) provides the theory behind density estimation. For a valid density estimate, the kernel needs to integrate to 1. The hard part is in deciding how wide to make the kernel. Scott describes methods for making this decision.

B. Quantile Plots

The integral of the estimated density up to a value approximates that the cumulative probability of observations begin less than or equal to that value. A cumulative distribution plot plots the sorted values on the x-axis and cumulative probabilities on the y-axis. A quantile plot is essentially the same plot transposed. For quantile plots the x-axis shows cumulative probabilities and the y-axis shows the sorted data values. Figure 5, a quantile plot, shows the pairs (p_i, q_i) where p_i is the cumulative probability and q_i represents the ith largest observation. Cleveland's convention connects the point pairs as shown.

There are different methods for approximating cumulative probabilities. The order statistics approach here follows Cleveland (1993) and uses the expression $(i - .5)/n$ for $i = 1, \ldots, n$ to calculate the cumulative probabilities, where n is the sample size. The probability corresponding

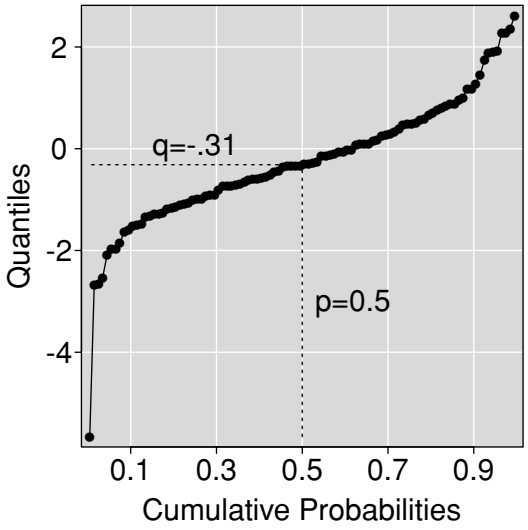


FIGURE 5

to the smallest observation in a sample of size 10 would be $(1 - .5)/10 = .05$. An interpretation is that if the observations were from a random sample of size 10, the probability of a future observation being smaller than the currently observed smallest observation is .05.

Another common textbook approach for calculating cumulative probabilities is the empirical method. For this method the cumulative probability for a value is the fraction of observations less than or equal to the value. This approach treats the curve as a step function jumping at each observation. The step function jumps from 0 to $1/10$ at the smallest observation in a sample of size 10. This suggests that the probability of a future observation being smaller than the smallest one already observed is 0. The sample does not allow determination of the true probability. The probability .05 seems reasonable. As the sample gets larger the discrepancy between the two calculation approaches get smaller.

Interpretation is straightforward. For any probability covered on the x-axis it is possible to determine a quantile. To obtain the .5 quantile (or estimated median) go straight up from .5 on the x-axis to the curve and then straight across to the y-axis to obtain x units. Starting with .25 and .75 yields corresponding quantiles are also known as the 1st and 3rd quartiles, or 25th and 75th percentiles, respectively. Similarly one can go from quantiles to cumulative probabilities. Since scientists use such plots to describe convenience collections from a population as well as random samples from a population, the trickiest interpretation task is often to decide if the probabilities really extend to a larger population.

Quantile plots are helpful on maps and provide a frame of reference for observing change over time. For example, one can tell if the 50th percentile (the median) has

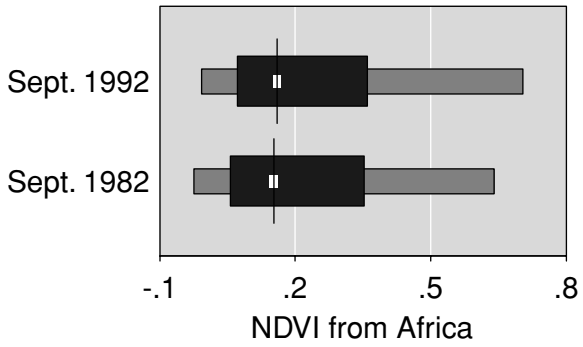


FIGURE 6 Box plots. This compare two distributions of values taken from Fig. 8. Vertical lines show the medians. First and third quartiles bind the thicker gray rectangles. Adjacent values bind the thinner gray rectangles. No outlines appear in the example. When interior white rectangles do not overlap, medians are not significantly different.

changed. For maps it is possible to save space by representing selected cumulative probability and quantile pairs using a parallel coordinate approach.

C. Box Plots

The box plot is a distribution caricature that has achieved wide acceptance. While used to represent individual distributions, the common use is to compare distributions. Figure 6 shows two boxplots. The features shown include the median, quartiles, and adjacent values. The notion of adjacent values sets a bound on what will be called outliers and warrant a note on how it is calculated. The upper adjacent value is the largest observation less than the 3rd quartile $+1.5$ times the interquartile range (the 3rd quartile -1st quartile). The lower adjacent value is determined correspondingly. Outliers, if any, are more extreme than the adjacent values. Not all box plot variations are the same. Some just show the maximum and minimum values rather than adjacent values and outliers. The design variation in Fig. 6 has an additional feature. It uses a white line to provide comparison intervals for the medians. If two comparison intervals do not overlap the medians are significantly different.

D. Quantile–Quantile Plots

Some authors consider the quantile–quantile (QQ) plots as the preferred graphic to make detailed continuous distribution comparisons. For theoretical distributions, the cumulative distribution function, $F()$, provides the correspondence between the probability and quantile pairs via $p = F(q)$. When $F()$ is a strictly increasing function the quantile function, Q(), is the inverse of $F()$ and $Q(p) = q$. Familiar pq pairs from the standard normal distribution

are (.5, 0) and (.025, −1.96). Comparison of two distributions, denoted 1 and 2, proceeds by plotting quantile pairs ($Q_1(p)$, $Q_2(p)$) over a range of probabilities, such as from .01 to .99 in steps of .01. For two distributions of observed data, the calculations described for the quantile plots (previously described) are appropriate. Figure 6 shows a QQplot for two sets of data. The x-axis shows quantiles from Set 1 and the y-axis shows quantiles from Set 2. Sometimes statistician chose $Q_1(p)$ to be from a theoretical family of distributions, such as the normal family, to see if parametric modeling is reasonable using the normal (Gaussian) family of distributions.

A merit of QQplots is that in simple cases they have a nice interpretation. If points fall on a straight line then the distributions have the similar shape differing only in the first 2 moments. This is the case in Fig. 7 since the robust fit (thin) line matches the quantiles quite well.

The slope and intercept of the approximating straight line tell about the discrepancies in the second moment (standard deviation) and first moment (mean). The slope of the thin line tells about the ratio of standard deviations. The thick line in the figure is the reference line for identical distributions. The lines are not quite parallel in Fig. 7 so the standard deviations are not quite the same. Multiplying Set 2 quantiles by a factor and then adding a constant changes the slope and intercept of the robust fit line. Graphical input and visual assessment provides a quick way to find the two numbers that make the lines match. These two numbers then indicate the ratio of the standard deviations and the defference in the means after the Set 2 scale adjustment. In Fig. 7 the lines are nearly parallel so a reasonable guess is that the distribution differ in mean by about .5.

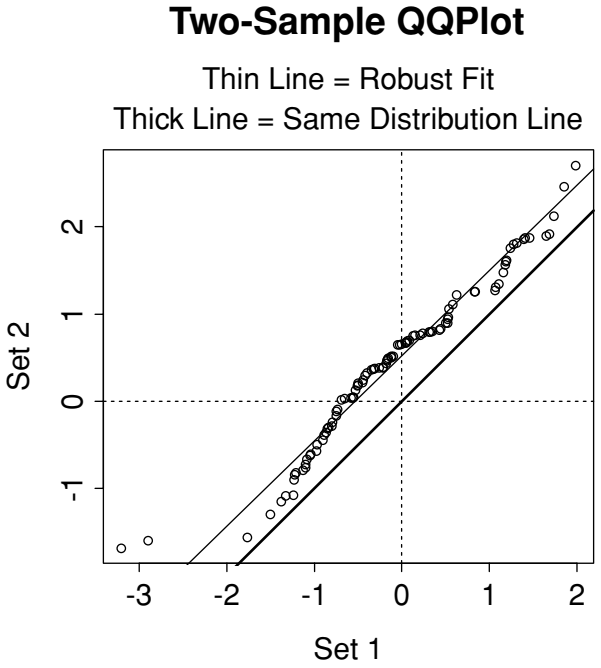


FIGURE 7 Quantile–Quantile (QQ) Plot. This compares distributions. When the points fall near a straight line, they have the same shape and differ only in the standard deviation and mean.

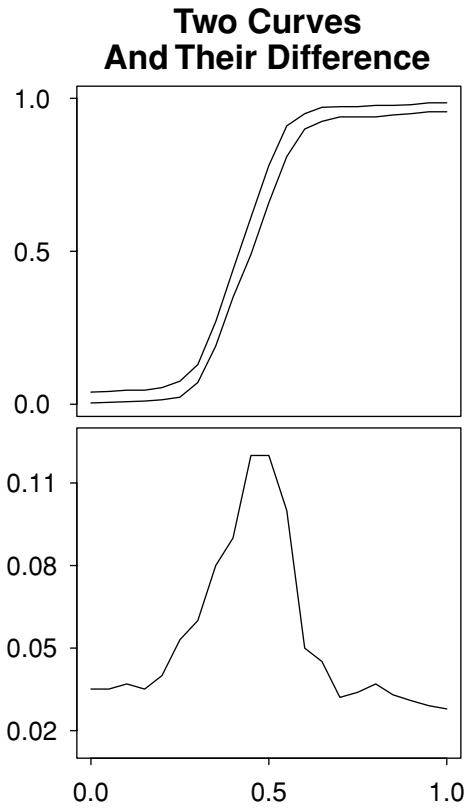


FIGURE 8 Explicit difference of two curves. Humans tend to see closest differences between curves, not differences parallel to the y-axis.

QQplots avoid the visually deceptive procedure of superimposing two cumulative distribution functions or two survival curves. As Fig. 8 suggests, we are really poor at judging the distance between curves. Our visual processing assesses the closest differences between curves rather than the correct vertical distances. This deficiency applies to comparing spectra and time series as well. Adding grid lines can help, but it is often better to plot the difference explicitly or make distribution comparisons using QQplots.

E. Direct Comparison of Two Distributions and the Mean Difference Plot

Before and after comparisons are common in science. The basic idea is to control for variation in experimental units by studying the change of experimental unit values. This differs from QQ plots in that the study unit is the basis for pairing values rather than cumulative percentages. Figure 9 shows a paired comparison plot for two

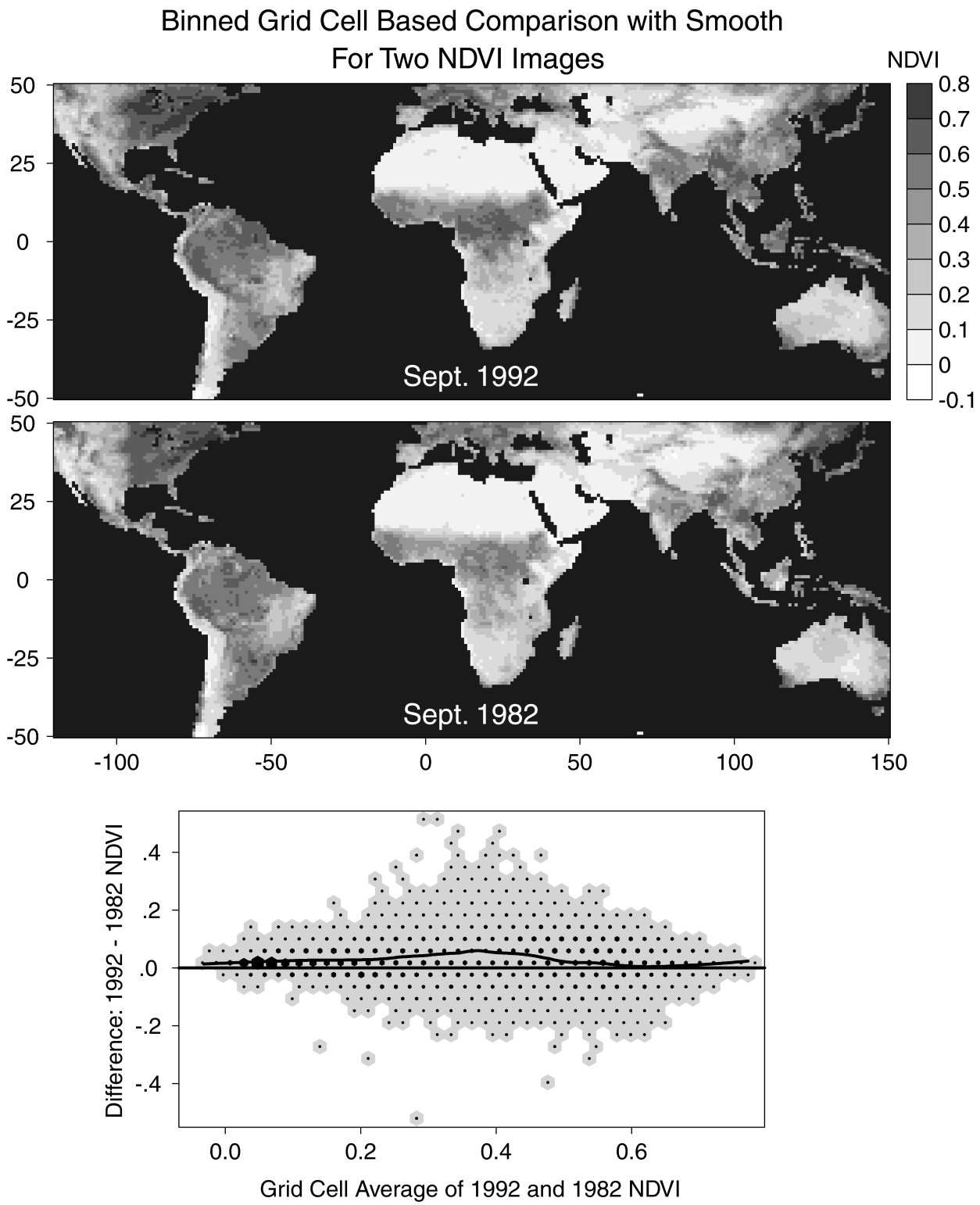


FIGURE 9 Pair Difference Plot (bottom panel). Top panels represent the Device Vegetation Index (NDVI) values on a 1° grid using gray levels. The images were taken about a decade apart. The bottom panel shows the difference of location paired values. The mean and difference plots show the mean on the x-axis and the difference on the y-axis. The reference line is $x = 0$. This plot handles many thousands of points by hexagon binning. Large hexagon symbols show more counts. While there are many exceptions, the smooth line indicates a tendency for 1992 values to be larger.

low-resolution satellite images of the same region. The top two images show NDVI values on a 1° grid. NDVI stands for normalized device vegetation index and provides a measure of greenness. The traditional reference line for equality is a 45° line through origin. John Tukey suggested making the reference line horizontal by plotting differences on the y-axis and mean on the x-axis. The bottom of Figure 9 shows a mean difference plot. Making transformations to simplify the visual reference and reduce the visual calculation burden is an important graphic design principle.

The data for the panels in Fig. 9 consist of over 27,000 point pairs. The overplotting in the bottom panel would cause a problem. Since we do not judge density well the bottom panel bins the data into hexagon bins and represents the count with the size of the hexagon symbol plotted. Such binning is a fast but rudimentary way to obtain a bivariate density estimate. Carr lists the merits of hexagon binning over square binning. An alternative view estimates the density surface on a higher-resolution bivariate grid and displays the surface with perspective views or contours.

The line in the bottom panel of Fig. 9 shows a smooth of the data. The upward shift indicates increased greenness in 1992. Symbols below the zero reference line for no change indicate that the change is a decrease in some locations. The smooth indicates that the increase tends to be highest for intermediate values of NDVI.

F. Functional Relationships and Smoothing

When y is considered a function of x, common practice is to enhance scatterplots of (x, y) pairs by adding a smooth curve. This is done in Fig. 9 to see if the difference is related to the average of the two values. To avoid the considerable human variability in sketching a fit, the standard procedure is to model the data using a procedure that others can replicate. Figure 10 shows a scatterplot with the smooth line generated using loess (see Cleveland *et al.* for more details). Loess fits the data using weighted local regression. That is, the regression uses data local to $x0$ to predict a value at $x0$. Points closest to $x0$ receive the greatest weight. The use of many local regressions produces a set of pairs (x, y) that are connected to produce the smooth. Each local regression in the smooth shown in Fig. 10 used a linear model in x and included the closest 60% of the observations to the prediction point $x0$. Such a smooth is reproduceable. The smooth in Fig. 10 draws further attention to the distinction between ocean and land states and additional modeling is appropriate.

Smoothing is an extremely important visual enhancement technique. It helps us see the structure through the noise. The decomposition of data into smooth and residual parts is fundamental technique in statistical modeling.

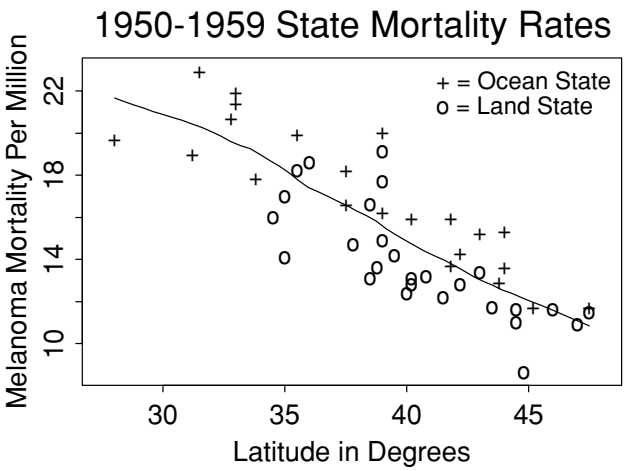


FIGURE 10 A smooth. The smooth curve suggests a functional relationship. The two types of points suggest there are two different functional relationships involved.

Hastie and Tibshirani provide a good introduction to a variety of smoothing methods.

Numerous smoothers are available. Historically, many researchers used cubic spline as smoothers. Cubic splines have a continuous second derivative and that is sufficient to make curves appear smooth to humans. The elegant mathematical formulation behind splines increased their popularity within the statistical community. However, there is no *a priori* best smoother. New methods, such as the wavelet smoothing, keep appearing in statistical software. Recently developed wavelet smoothers are supposed to be better than many smoothers at tracking discontinuities in the functional form. However, the old local median smoothers still do well at handling discontinuities.

Smoothers typically have a smoothing parameter that needs to be estimated or specified by the user. With computational power at hand, cross-validation methods have become increasingly popular as a community standard. This reduces the judgment burdens on the analyst. However, this does not guarantee a match between an empirical curve and a hypothesized true but unknown underlying curve.

III. MULTIVARIATE VISUALIZATION

Visualization in the physical sciences is inherently multivariate. Scientists are interested in the relationships among many attributes and the attributes have space–time coordinates. The purpose of multivariate graphics is to show multivariate patterns and to facilitate comparisons. As in low dimensions, the patterns often concern data distributions or models with at least one dependent variable.

After converting attributes and their space–time coordinates to images for evaluation, human visual comparisons

typically fall into three categories, comparison of external images to each other, comparisons of external images to external visual references, and comparison of external images to the analyst's internal references. Internal references include scientific knowledge, statistical expectations, and process models. The visual investigation process often involves converting internal references into external visual references subject to further manipulation. With external images and references available, the next step often involves graphical transformation to simpler forms in terms of our perceptual–cognitive processing abilities.

Multivariate graphics must deal with the problem of showing higher-dimensional relationships and, as in lower dimensions, often have to deal with the noise that obscures patterns and the overplotting due to many points.

A. Graphical Design Principles and Resources

The description indicates the increased complexity as we move to higher dimensions. At the same time, Kosslyn warns that "The spirit is willing but the mind is weak." We should approach the multivariate challenge prepared to do battle. Our arsenal of tools includes design principles that help us in uncharted territory. Some of our tools include:

- Distributional caricatures such as box plots to help us deal with large data sets
- Map caricatures that let us show small multiples
- Modeling to reduce noise and complexity
- Layering and separating to constrain and guide the information flow
- Partitioning and sorting to simplify appearance and facilitate comparisons
- Linking low-dimensional views to get insight into higher-dimensional relationships.

The basic formats for comparison graphics include juxtaposition, superposition, or the direct display differences. The art of multivariate graphics is to select the methods and enhancements that work best in view of the phenomena's complexity view and in view of human perceptual and cognitive limitations.

This entry suggests just a few representation tools and principles. Good reference include books by Cleveland, Kosslyn, Wilkinson, and Ware. The books by Tufte provide wonderful examples of putting design principles into practice. MacEachren provides a good introduction on mapping and Cleveland and McGill provide an early survey on dynamic multivariate graphics that go beyond the scope of the entry. As more work is done in computing environments, issues of the computer human interface become increasingly important. Card *et al.* edited a book of readings that gathers many important concepts.

The entry only briefly describes color since it is not an option here. However, color provides an important tool. Different hues provide a good way to distinguish categories of a categorical variable. The cartographic guidance limits categories or hues to six or fewer. Humans are very sensitive to a second dimension of color, a dark-to-light scale that is referred to in literature by terms such as value, lightness, or brightness. Gray levels provide ordered scale so it is thinkable to represent a continuous variable using value or gray level. Humans are less sensitive to the dimension of color called saturation. A saturation scale going from an achromatic color such as medium-gray to a vivid red is also an ordered scale. However, humans can make fewer saturation distinctions than lightness distinctions. Generally color is a poor choice for representing a continuous variable. Humans perceive that hue and brightness are integral dimensions so we should not use them to encode two or three variables. Brewer *et al*. discuss additional considerations that apply to the color-vision impaired.

Since most people can work with four chunks of information, this article suggests attempting to fit the guidelines into four broad categories of quantitative design principles:

- Use encodings that have high perceptual accuracy of extraction
- Provide context for appropriate interpretation
- Strive for simple appearance
- Involve the reader

These organizing categories contain some conflicting guidelines. For example, a long list of caveats may provide the context for appropriate interpretation but conflict with simple appearance and reader involvement. Composing graphics that balance among the guidelines remains something of an art form. The communication objectives influence the balance.

B. Communication Objectives

Multivariate graphics can have many different communication objectives. Four common objectives include providing an overview, telling a story, suggesting hypotheses, and criticizing a model. In providing an overview, broad coverage is important. Achieving clarity often involves hiding details and working with caricatures. Similarly, in telling a story the predetermined message must shine through. Scientists often fail to tell simple stories because they are eager to list caveats and a host of details that qualify the basic results. Interactive web graphics can alleviate the archival side of this problem providing ready access to metadata, supplemental documents, and gigabyte databases. However, it still takes careful design to draw readers to the details.

This entry emphasizes graphics of discovery. Discovery objectives include suggesting hypotheses and criticizing models. The graphics in this article can often be used to display residuals from models and to review patterns in the residuals. For discovery, it is often crucial to see through the known and miscellaneous sources of variation. In the context of mortality mapping, John Tukey said, "the unadjusted plot should not be made." Today mortality maps control for known variation by being sex and race specific. The maps control for age by limiting the age range or by showing adjusted statistical rates. In the physical sciences there are often known factors that contribute to variation that warrant making adjustments. Thus the graphics of discovery often show model residuals.

C. Surface Views

In some ways multivariate visualization is similar to lower-dimension visualization. The general tasks are still to show densities and functional relationships. The difference is that more variables including spatial and temporal indices are called into play. The density of bivariate points constitutes a third coordinate. The basic idea in bivariate kernel density estimation is similar to that for univariate estimation: average local likelihoods. The key difference is that local neighborhood is bivariate. The result is surface $z = f(x, y)$. Figure 11 shows the contours of a bivariate density surface. Figure 12 shows a wireframe perspective view.

Estimating functional relationships of the form $z = f(x, y)$ also follows the pattern established with one

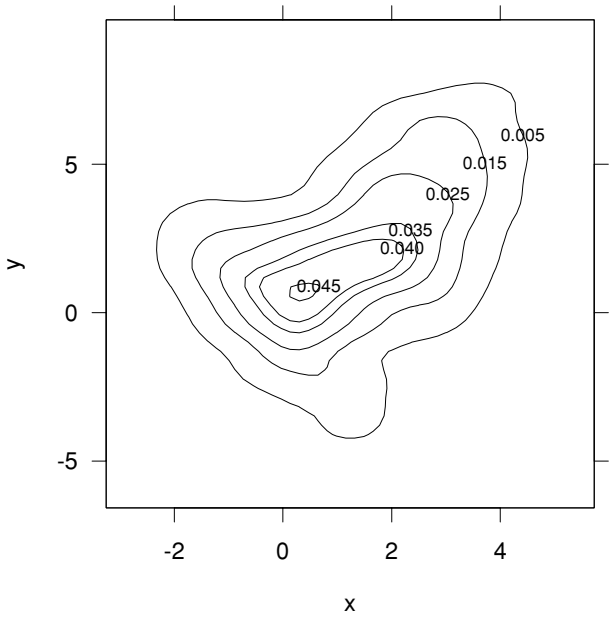


FIGURE 11 A contour plot. This plot represents the computed density of bivariate points.

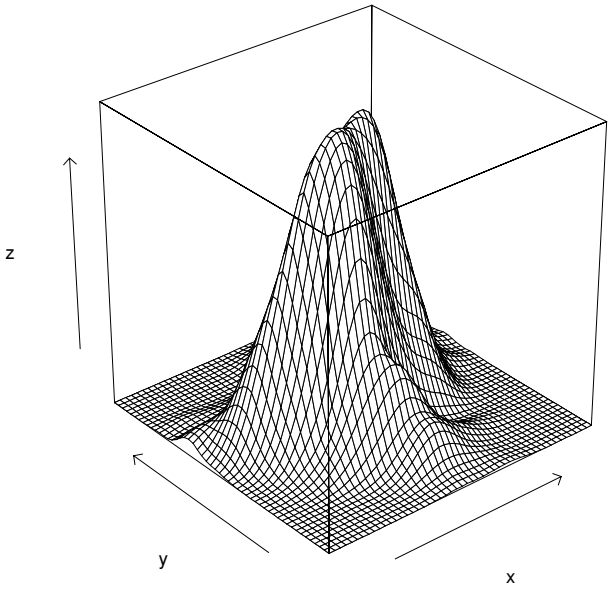


FIGURE 12 A perspective view or wireframe plot. This surface shows the density of bivariate points.

less variable. In the context of maps some approaches to modeling address the issue of spatial correlation. Common monochrome views again include contour and wireframe plots. Fully rendered color surfaces with highlights provide better a view of the wireframes. Pairing of contour and surface plots can aid understanding. The surface tends to provide a good overall impression (except for what is hidden) while the contours help to locate features, such as local extrema, on the plane. Wireframes and translucent surfaces can be superimposed on maps or contour plots.

Perspective views appeal to many people and have also been used to show local values as the animated flyovers of three-dimensional bar charts as well as surfaces. How perspective foreshortening complicates accurate decoding of values and comparisons. For a detailed study of surfaces, Cleveland recommends dropping back a dimension. By conditioning on intervals of x or y it is possible to return study strips of the surface using two-dimensional plots.

For comparing surfaces there are three standard approaches. One is to superimpose surfaces, for example, rendering them in translucent color. A second is to juxtapose displays of two surfaces. The third option of calculating and showing the difference of two surfaces is often best.

D. Distance Judgments and 3-D Scatterplots

In the multivariate context, accurate interpoint distance judgments are crucial to geometrically based visual interpretation. For two points in a scatterplot the cognitive task becomes one of judging the length of an implicit line between the two points. As indicated previously, humans

perceive length in a plane with good perceptual accuracy of extraction. For higher-dimensional representations, assessing interpoint distances becomes increasingly difficult. The position here is that our ability to judge distance in three-, four-, and five-dimensional representations provides a strong basis for ranking point encoding methods.

At first thought, a stereo three-dimensional scatterplot might seem ideal for representing continuous three-dimensional data. Judging distance between two points then equates to judging the length of a line segment. On closer inspection, there is substantial change from two-dimensions to three-dimensions. Depth perception for the third coordinate derives from horizontal binocular discrepancies (parallax). The horizontal discrepancies involve only a small fraction of our full horizontal field of view. Horizontal visual acuity within this small interval determines how many distinct depth planes we can resolve. This cannot match the horizontal resolution across the full field of view. Stereo three-dimensional plots are less than ideal since humans do not judge depth as accurately as they judge vertical and horizontal position. While stereo three-dimensional distance judgments are not as accurate as two-dimensional distance judgments, we live in a three-dimensional world and have strong intuitions about three-dimensional relationships. For those with good depth perception the testable conjecture here is that three-dimensional scatterplots allow the most accurate assessment of interpoint distances of the available encoding for three continuous variables.

Historically most analysts created depth views via rotation (motion parallax). This approach is very powerful. The main drawbacks are that moving points are harder to study than stationary points and that interacting with moving data is awkward. Using both stereo and rotation maintains the depth when rotation stops. Overplotting obscures points and some authors suggest we can only see two and one-half dimensions. Mostly one sees the edges of the data cloud. Different viewing angles bring out different edges. Slicing or sectioning helps to reveal the inside of the cloud. In the virtual reality (VR) settings one can fly through data clouds and touch points (or density features) to gain additional detail. Since we do not judge density accurately there is good reason to estimate trivariate densities and study the hypersurfaces $w = f(x, y, z)$. Scott provides instructive graphics that simultaneously show three contours of hypersurfaces.

E. Scatterplot Matrices, Parallel Coordinates, and Stereo Ray Glyphs

If we have points on a surface $w = f(x, y, z)$ there are many different ways to view the points. It is instructive to consider how the different views reveal that the data are from a surface and not points distributed throughout four

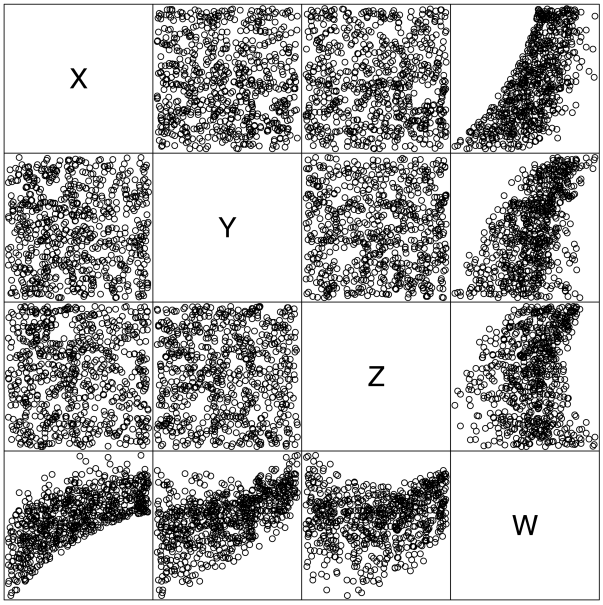


FIGURE 13 A scatterplot matrix. The four-dimensional points fall on a hypersurface defined by $w = f(x, y, z)$. The point cloud edges in last row and column suggest the date are a lower-dimensional object embedded in four dimensions. The cloud interiors do not show much.

dimensions. The three views considered here are scatterplot matrices, parallel coordinates, and stereo ray glyphs.

Figure 13 shows a scatterplot matrix. The panels in the matrix show scatterplots for all pairs of variables. Some variations of the scatterplot matrix show only the upper or lower triangle of panels since otherwise the same point pairs appear (transposed) in a second panel. If fact, thoughtful construction allows two data sets to be juxtaposed using upper and lower triangles. Returning to the task at hand, the edges of the cloud in the panels of the last row and column of the scatterplot matrix suggest there is some kind of relationship $w = f(x, y, z)$.

Figure 14 shows a parallel coordinate plot. This has four parallel axes. The plot scales each variable to the interval [0 1] and equates this to the length of the corresponding axis. Each component of a four-dimensional point is plotted on the respective axis and the components are connected. The clue about the relationships is limited to the line patterns in the last pair of axes.

Figure 15 shows a stereo ray glyph plot in a side-by-side stereo pair view. The three coordinates are represented by the point's three-dimensional location. The angle of the ray represents the fourth coordinate. The small size of a side by side stereo plot complicates views, but the smooth variation of the ray angle locally in three-dimensional space is evident. The locally smooth variation of the ray angle indicates there is a relationship.

As a demonstration of the inferior nature of nonpositional glyphs in low dimensions, one can generate, say

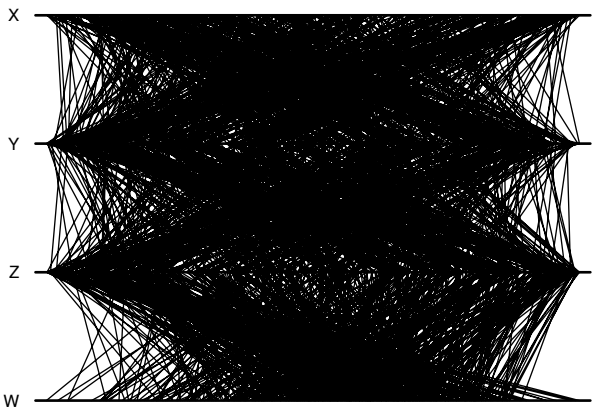


FIGURE 14 A parallel coordinate plot. The same four-dimensional points as in Fig. 13. The weak hints of a dimension reducing constraint appear in only in lines between the the z and w axes. The plot is useful for graphical input and for viewing many coordiates.

a thousand points of three-dimensional data embedded in four dimensions. That is, select the triples (u, v, w) randomly, say from a normal distribution, and then let $x1 = f1(u, v, w)$, $x2 = f2(u, v, x)$, $x3 = f3(u, v, w)$, and $x4 = f4(u, v, w)$ where the four functions are simple distinct polynomials. The structure in the stereo ray glyph plot will immediately suggest the existence of a constraint and the fact the data are not four-dimensional. The ability to observe smooth variation can easily be masked by noise in real data. Still it should be clear that the scatterplot matrix and the parallel coordinate plot do not help us to see in three dimensions as well as three-dimensional scatterplots do. Our brains are not wired for the task of constructing higher-dimensional views from projections in lower dimensions.

Few researchers have seriously tackled the visualization of six-dimensional data. Bayly and co-authors provide a notable exception. They successfully used colored stereo ellipsoids to evaluate problems in improving an electrostatic potential model. Their article includes color side-by-side stereo figures. In a long sequence of efforts they failed to obtain insight using a wide variety of encodings. Their eventual success suggests that using position along a scale to represent three variables has merit.

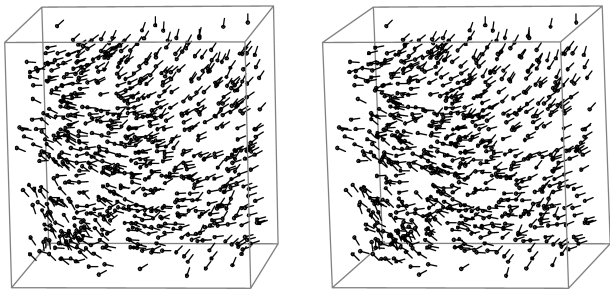


FIGURE 15 A stereo ray glyph plot. The ray pointing down to show small values and up to show large values. Some people can fuse the stereo pair images without a viewer. The smooth local variation of the ray angle strongly suggests that lower dimensional structure is embedded in four dimensions.

As we move beyond four dimensions, our ability to judge interpoint distances deteriorates quickly. We are forced to use clustering algorithms and other computational methods to bring out patterns. Simple graphical methods do not work well except when there is a very simple geometric structure embedded in high-dimensional space. For complex multivariate structure, Furnas and Buja indicate how we can lower the viewing dimensionality by slicing or sectioning (adding linear constraints). We can also condition on categorical variables. This divide-and-conquer approach can reveal a plethora of patterns, but it is only the rare individual that can integrate a catalog of low-dimensional views into a coherent higher-dimensional framework. Few people claim to fully understand the richness of even four-dimensional structures.

F. Glyphs and Layouts

Glyphs provide one way to represent observations with several variables as points. The climate description maps of the U.S. Environmental Data Service have some interesting glyphs. These include a wind rose glyph showing 16 variables representing the percentage of time the wind is blowing in the major compass directions, and bar chart glyphs show 12 variables indexed by month of the year. In both these cases the glyph controlling variables are statistical summaries. It is not a requirement that variables controlling glyph characteristics be direct measures.

Chernoff faces seem to be the glyphs that draw the most attention. Different variables control different facets of caricature faces. Examples appear in Fig. 16. Since a significant portion of our brain is devoted to face recognition, the encoding idea intrigues many people. While humans get a gestalt impression that they can often relate to an emotion, it is difficult to respond to the individual components of the face. Few glyphs see much use. The star glyph, a variant of a parallel coordinate plot for a single case, seems to appear with some frequency.

A simple example plotting a few hundred faces or stars representing points on a hypersurface is instructive. The hypersurface relationship remains obscure. That is discouraging. In contrast to the map glyphs introduced previously, glyphs are often laid out arbitrarily on a regular grid. This wastes using the favored representation, position along a scale, for two coordinates. Position using possibilities that can lead to overplotting include using the first two principal components of multivariate data, and

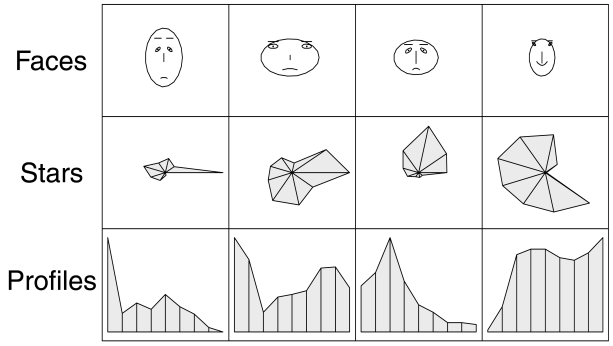


FIGURE 16 Three types of glyphs. Here a face, profile, and star glyph represent nine coordinates. In this case the coordinates represents a gene's mRNA production levels at nine animal ages.

multidimensional scaling provides another option. There are additional options that avoid overplotting.

G. Linked Plots

Linking points across plots provides a way to connect cases whose variables are represented in different plots. Linking provides a weaker binding of the multivariate observations than glyphs. Linking methods include linking by lines, colors, names, pointers, and spatial linking by juxtaposition. The following discussion emphasizes line linking and color linking.

The parallel coordinate plot (see Fig. 14) provides an example of linking with lines. The lines connect components of the same case across the axes showing the different variables. In many applications of line linking the lines overplot and this complicates decoding the plot.

Painting points on a screen provides a common way of linking points using color. A common application involves scatterplot matrices. Painting some points in one panel selects the corresponding cases. The color then highlights views of the selected cases in the other panels. This notion extends to different types of panels. When there are spatial coordinates one of the panels is often a map.

H. Conditioned Plots

Conditioned plots partition the estimates (or data) into sets based on classes defined for conditioning variables. The different plots for the sets then appear as juxtaposed panels. The visualization task is then to study how the distributions or function relationships shown in the panels vary across the conditioned panels. Tukey and Tukey provide an early example called a casement display. Conditioned plots appearing as a one-way or two-way layout are typically two-dimensional plots, but they can be three-dimensional wireframe plots or other higher-dimensional plots. Many people readily understand one- and two-way conditioning and the corresponding layout of panels. For example, defining each of two conditioning variables as low, medium, and high leads to a 3×3 layout of panel showing other variables of interest. Thus conditioned plots are one of the more effective ways to study relationships involving three to five dimensions.

IV. CLOSING REMARKS

This entry provides a description of several templates for low-dimensional graphics. When datasets have higher dimensions there are a variety methods that will reduce the dimensions of many datasets to something that we can view using the templates. The methods include principal components, projection, and sectioning. The methods have some limitations including the sizes of data sets to which they apply.

Increasingly, the size, detail, and complexity of data sets overwhelm our direct visualization capacity. Our dimension reduction and case reduction methods are stretched thin and overwhelmed at times. The number of plausibly important views of data exceeds the time we have to look and study. In thinking about the future, John Tukey coined the term "cognostics" (diagnostics interpreted by a computer rather than by a human). The idea was to compute features of merit and have computers rank plots by their by potential interest to humans. Researchers have developed projection pursuit methods that help us to find interesting views in lower dimensions. There is much more to be done along this line. The work can be frightening because our algorithms can miss details like the hole in the ozone layer. On the other hand we miss a lot because our viewing is not optimized. We also miss a lot because we do not really put computers to work to find the representations that are optimal for our understanding! While our visualization methods fall further behind the increasing amounts of data, it remains the case that we can do much more than we are doing now.

SEE ALSO THE FOLLOWING ARTICLES

COLOR SCIENCE • FLOW VISUALIZATION • GRAPH THEORY • IMAGE PROCESSING

BIBLIOGRAPHY

Bayly, C. I., Cieplak, P., Cornell, W. D., and Kollman, P. A. (1993). "A Well-Behaved Electostatic Potential Based Method Using Charge

Restraints for Deriving Atomic Charges: The RESP Model," *J. Phys. Chem.* **97,** 10269–10280.

Brewer, C. A., MacEachren, A. M., Pickle, L. W., and Herrmann, D. (1997). "Mapping Mortality, evaluation color schemes for choropleth maps," *Annals of the Association of American Geographers* **87**(3), 411–438.

Card, K., Mackinlay, J. D., and Schneiderman, B. (1999). "Information Visualization Using Vision To Think," Morgan Kaufmann Publishers, Inc., San Francisco, CA.

Carr, D. B. (1991). Looking at large data sets using binned data plots. "Computing and Graphics in Statistics" (A. Buja and P. Tukey, eds.), pp. 7–39, Springer-Verlag, New York.

Cleveland, W. S. (1993). "Visualizing Data," Hobart Press, Summit, NJ.

Cleveland, W. S. (1994). "The Elements of Graphing Data," Hobart Press, Summit, NJ.

Cleveland, W. S., and McGill, R. (1984). "Graphical perception: Theory, experimentation, and application to the development of graphics methods," *J. Am. Stat. Assoc.* **79,** 531–554.

Cleveland, W. S., and McGill, M. E. (eds.). (1988). "Dynamic Graphics for Statistics," Chapman and Hall, New York.

Furnas, G. W., and Buja, A. (1994). "Prosection views: Dimensional inference through sections and projections," *J. Comp. Graph. Stat.* **3**(4), 323–353.

Hastie, T. J., and Tibshirani, R. J. (1990). "Generalized Additive Models," Chapman and Hall, New York.

Kosslyn, S. M. (1994). "Elements of Graph Design," W. H. Freeman and Company, New York.

MacEachren, A. M. (1994). "Some Truth with Maps: A Primer on Symbolization & Design," Association of American Cartographers, Washington, D.C.

Scott, D. W. (1992). "Multivariate Density Estimation; Theory, Practice and Visualization," Wiley, New York.

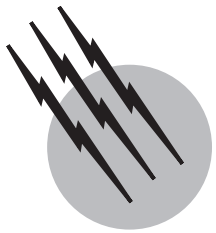
Tufte, E. R. (1983). "The Visual Display of Quantitative Information," Graphics Press, Cheshire, CN.

Tufte, E. R. (1990). "Envisioning Information," Graphics Press, Cheshire, CN.

Tufte, E. R. (1997). "Visual Explanations," Graphics Press, Cheshire, CN.

Wilkinson, L. (1999). "The Grammar of Graphics," Springer-Verlag, New York.

Wood, D. W. (1992). "The Power of Maps," The Guilford Press, New York.



Group Theory

Ronald Solomon
Ohio State University

I. Fundamental Concepts
II. Important Examples
III. Basic Constructions
IV. Permutation Groups and Geometric Groups
V. Group Representations and Linear Groups
VI. Finite Simple Groups
VII. Other Topics

GLOSSARY

Abelian group Group in which the operation is commutative.
Center Subgroup of a given group consisting of all elements which commute with every group element.
Commutator subgroup Subgroup of a given group generated by the subset of all commutators.
Cyclic group Group generated by a single element.
Homomorphism Function mapping a group into a group and preserving the group operation.
Isomorphism One-to-one homomorphism from one group onto another.
Linear representation of a group Homomorphism of a group into the group of all invertible linear operators on a fixed vector space.
Permutation representation of a group Homomorphism of a group into the group of all permutations of a fixed set.
Quotient group of a group G Group of all cosets of a fixed normal subgroup of G.
Representation of a group Homomorphism of the group into the full group of all symmetries or automorphisms of some fixed mathematical object.
Simple group Group whose only normal subgroups are the identity subgroup and the group itself.
Solvable group Group which has a normal series of subgroups such that each of the factors (quotient groups) of the series is abelian.
Subgroup Subset of a group which is a group under the same operation.
Sylow p-subgroup Subgroup of a group which is maximal with respect to the property that each element has order a power of the prime p.

A GROUP is a nonempty set G of invertible functions mapping a set X to itself such that whenever $f, g \in G$, then also $f^{-1} \circ g \in G$. (There is also a more abstract definition which we shall give in the next section.) Thus, in particular, G is a subset (subgroup) of the group $S(X)$ of all invertible functions whose domain and range is the set X.

Any such function may be thought of as a permutation or symmetry of the set X, and $S(X)$ is generally referred to as the symmetry group of X or as the symmetric group on the set X.

Historically, groups emerged first in the work of Lagrange, Ruffini, and Cauchy around 1800 as sets of permutations of a *finite* set X. In this context the definition may be simplified. A nonempty set G of permutations of a finite set X is a group if and only if G is closed under composition of permutations. Closure under functional inversion follows automatically. The condition of closure under composition puts severe restrictions on the group G, e.g., the remarkable fact discovered by Lagrange that if $|X| = n$ [and so $|S(X)| = n!$] and G is any subgroup of $S(X)$, then $|G|$ is a divisor of $n!$

The subject was greatly enriched by Evariste Galois, who demonstrated around 1830 how to associate to each polynomial equation a finite group (called its Galois group) whose structure encodes considerable information concerning the algebraic relationships among the roots of the equation. An analogous theory for differential equations was developed later in the century by Lie, Picard, and Vessiot. The relevant groups in this setting were no longer finite groups but rather continuous groups (more commonly known as Lie groups).

Starting around 1850, Cayley championed the idea of studying groups abstractly rather than via some specific representation. This notion simplifies numerous arguments and constructions, notably the construction of quotient groups. The first modern definition of a group was given by Dyck in 1882, who also pioneered the study of infinite groups.

I. FUNDAMENTAL CONCEPTS

We shall freely use the standard notation of naive set theory with regard to sets, subsets, elements, relations, functions, etc. By a binary operation $\circ$ on a set S, we shall mean a function $\circ : S \times S \to S$. It is customary to say that S is closed under the binary operation $\circ$ in this situation, and it is customary to write $x \circ y$ for the value of $\circ$ at the ordered pair $(x, y) \in S \times S$. The symbols $\mathbf{Z}$, $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{C}$ will denote the sets of integers, rational numbers, real numbers, and complex numbers, respectively.

We now give the promised abstract definition of a group.

Definition: *A group is an ordered pair* $(G, \circ)$ *such that* G *is a set closed under the binary operation* $\circ$ *satisfying*:

(1) $(x \circ y) \circ z = x \circ (y \circ z)$ *for all* $x, y, z \in G$.
(2) *There is an element* $e \in G$ *with* $e \circ x = x = x \circ e$ *for all* $x \in G$.
(3) *For any* $x \in G$, *there is an element* $x^{-1} \in G$ *such that* $x^{-1} \circ x = e = x \circ x^{-1}$.

This definition makes the group operation the prominent feature and is indifferent to the kind of mathematical objects which the elements of G might be.

The element e is unique and is called the identity element of G, and the element x^{-1} is uniquely determined by x and is called the inverse of the element x. Henceforth in this article we shall in general write xy for $x \circ y$. Also, we shall typically speak of a group G, with the operation $\circ$ being tacitly understood. It is important to note that in general $xy \neq yx$. A group in which the commutative law

$$xy = yx \qquad \text{for all } x, y \in G$$

holds is called an abelian group in honor of Nils Abel, who first recognized that a polynomial equation whose Galois group is abelian is solvable by radicals. Other groups are called nonabelian groups.

Sometimes when G is an abelian group, the operation is denoted by $+$ and the identity element by 0. Many important sets of numbers form abelian groups under $+$, for example, $(\boldsymbol{C}, +)$ and such subsets as $(\mathbf{Z}, +)$, $(\mathbf{Q}, +)$, and $(\mathbf{R}, +)$.

Groups of numbers like $(\mathbf{R}, +)$ seem quite different from groups of functions. However, we may identify $(\mathbf{R}, +)$ with the group $(\mathbf{T}, \circ)$ of all translation functions $T_a : \mathbf{R}^1 \to \mathbf{R}^1$ defined by

$$T_a(x) = a + x \qquad \text{for all } x \in \mathbf{R}^1.$$

More generally, as observed by Cayley and Jordan, if $(G, \cdot)$ is any abstract group, then G may be regarded as a group of invertible functions on the set G by regarding the element $g \in G$ as the function $\lambda_g : G \to G$ defined by

$$\lambda_g(x) = gx \qquad \text{for all } x \in G.$$

It is for this reason that our two definitions of group are essentially the same. The mapping $G \to S(G)$ taking g to λ_g is called the (left) regular representation of G.

Groups like $(\mathbf{Z}, +)$, $(\mathbf{Q}, +)$, and $(\mathbf{R}, +)$ are instances of subgroups of $(\mathbf{C}, +)$ in the following sense.

Definition: *Let* $(G, \circ)$ *be a group. A nonempty subset* H *of* G *is a subgroup of* G *if* $(H, \circ_H)$ *is itself a group, where* $\circ_H$ *is the restriction of the function* $\circ$ *to* $H \times H$.

It follows readily that the identity in H is the identity in G and that inverses of elements in H are the same as their inverses in G.

Given any subgroup H of a group G, we may define an equivalence relation on the group G by the rule

$$g \equiv_H g_1 \qquad \text{if } g^{-1}g_1 \in H.$$

The equivalence classes determined by this equivalence relation are called the left cosets of H in G. Thus the left coset containing the element g is just

$$gH = \{gh : h \in H\}.$$

The set $G\backslash H$ of all left cosets of H in G forms a space on which G acts by left multiplication, in the sense that to each element $g \in G$ there is naturally associated a function $L_g : G\backslash H \to G\backslash H$ by $L_g(xH) = gxH$ and the functional equation $L_{gg_1} = L_g \circ L_{g_1}$ holds for all $g, g_1 \in G$. The left regular representation of G is the special case where $H = \{e\}$. However, although the map $g \to \lambda_g$ is one-to-one, in general, many elements of G may give rise to the same function on $G\backslash H$.

The fact that all left cosets of H have the same cardinality yields the first important theorem about finite groups.

Lagrange's Theorem: *Let G be a finite group and H a subgroup of G. Then $|H|$ is a divisor of $|G|$.*

In a similar way, it is possible to define the right cosets Hg of a subgroup H. An important class of subgroups, first studied by Galois, is the normal subgroups, which are those for which the right and left cosets coincide.

Definition: *A subgroup N of a group G is a normal subgroup if and only if $gN = Ng$ for all $g \in G$. When N is a normal subgroup of G, we write $N \lhd G$.*

Given a group G and a normal subgroup N, we can construct a new group G/N, called the quotient group of G by N.

Definition: *Let G be a group and N a normal subgroup of G. The set $G\backslash N$ with the operation $(gN)(g'N) = gg'N$ is a group, called the quotient group G/N.* (*Note that it is conventional to replace $\backslash$ by $/$ when the set is regarded as a group.*)

The normality condition is necessary for the multiplication on $G\backslash H$ to be well defined. The identity subgroup {e} and the entire group G are always normal subgroups of every group G. A group G whose only normal subgroups are {e} and G is said to be a simple group. Simple groups play the role in the theory of finite groups that prime numbers play in the theory of numbers, as remarked by Michael Solomon, among others. All finite groups are "composed" from the simple ones.

At the other extreme, every subgroup of an abelian group is normal. If G is any group and $g \in G$, then the subset $\{g^n : n \in \mathbf{Z}\}$ forms a subgroup of G called the cyclic subgroup generated by g and often denoted $\langle g \rangle$. If $G = \langle g \rangle$, then G is called a cyclic group. It follows that an abelian simple group must be cyclic, indeed cyclic of cardinality p for some prime p.

The quotient construction affords important examples of cyclic groups. Namely, for each positive integer n, the subset $n\mathbf{Z}$ of all multiples of n in $\mathbf{Z}$ is a subgroup of the abelian group $\mathbf{Z}$ and hence we have cyclic groups $\mathbf{Z}/n\mathbf{Z}$, generated by the coset $1 + n\mathbf{Z}$. Cyclic groups arise in many natural contexts. For instance, the set $Z(n)$ of all complex nth roots of 1 forms a cyclic group of cardinality n under multiplication. Likewise, the group of all rotational symmetries of a regular n-sided polygon is a cyclic group C_n of cardinality n, under composition of functions. The groups $\mathbf{Z}/n\mathbf{Z}$, $Z(n)$, and C_n have identical properties as abstract groups, although their manifestations in the "real world" are different. Under such circumstances the groups are called isomorphic. The formal definition follows.

Definition: *The function $\Phi : (G, \circ) \to (H, \cdot)$ is an isomorphism (of groups) if $\Phi : G \to H$ is a bijective function satisfying*:

$$\Phi(g \circ g') = \Phi(g) \cdot \Phi(g') \qquad \text{for all } g, g' \in G.$$

Two groups which are isomorphic are indistinguishable as abstract groups, and the isomorphism may be regarded as a dictionary for translating from one group to the other. Much of the power of abstract group theory derives from the fact that a given theorem about abstract groups may have numerous important instantiations via different realizations of the abstract groups. For later reference, we note that an isomorphism $\Phi : G \to G$ between a group and itself is called an automorphism of G. The set of all automorphisms of G itself forms a group, Aut(G), under composition of functions. (We shall refer later to an automorphism of a field, which has a similar meaning.)

Returning to cyclic and abelian groups, we observe that we now know, up to isomorphism, every cyclic group and every abelian simple group.

Theorem:

(1) *Every cyclic group is isomorphic either to $\mathbf{Z}$ or to $\mathbf{Z}/n\mathbf{Z}$ for some positive integer n; and*
(2) *Every abelian simple group is isomorphic to $\mathbf{Z}/p\mathbf{Z}$ for some prime p.*

In contrast to this elementary result, the classification of the finite nonabelian simple groups was a very challenging and complex undertaking, and the classification of infinite simple groups is surely impossible. In any case, a strategy for the analysis of a general (finite) group is to "decompose" it into simple factors via an iterated quotient group construction. This is meaningful for finite groups thanks to the following theorem.

Jordan–Hölder Theorem: *Let G be a finite group. Then there is an integer $n \geq 0$ and a chain of subgroups*

$$G = G_0 \geq G_1 \geq \cdots \geq G_{n-1} \geq G_n = \{e\}$$

such that $G_{i+1} \lhd G_i$ for all $i < n$ and G_i/G_{i+1} is a nonidentity simple group for all i. The chain is not necessarily unique, but the list $G_0/G_1, G_1/G_2, \ldots, G_{n-1}/G_n$ is uniquely determined by the group G, up to changing the order of the groups in the list and replacing groups by isomorphic copies.

The chain of subgroups is called a composition series for G, and the set of quotient groups is called the set of composition factors of G. Continuing the analogy with numbers, the Jordan–Hölder theorem plays the role of the fundamental theorem of arithmetic. In contrast to a number, however, a group G is hardly ever determined up to isomorphism by its set of composition factors. Indeed, already in the elementary case, where the set of composition factors is $\{\mathbf{Z}/p\mathbf{Z}, \mathbf{Z}/p\mathbf{Z}\}$, there are two nonisomorphic possibilities for the group G, one cyclic and the other not.

II. IMPORTANT EXAMPLES

We begin with some important examples of abelian groups. We have already mentioned the abelian group $(\mathbf{C}, +)$ and some of its subgroups. Sometimes the operation in an abelian group is naturally thought of as multiplication. Indeed, the algebraic object called a field may be defined as a set F closed under two binary operations $+$ and $\cdot$ such that $(F, +)$ is an abelian group with identity element 0, $(F - \{0\}, \cdot)$ is also an abelian group (with identity element 1), and the two operations are intertwined by the distributive law:

$$x(y + z) = xy + xz \qquad \text{for all } x, y, z \in F.$$

Thus, for example, $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{C}$ are fields, and so in particular $(\mathbf{Q} - \{0\}, \cdot)$, $(\mathbf{R} - \{0\}, \cdot)$, and $(\mathbf{C} - \{0\}, \cdot)$ are abelian groups with identity element 1. An important example of a subgroup of $\mathbf{C} - \{0\}$ is the circle group S^1:

$$S^1 = \{a + bi \in \mathbf{C} : a^2 + b^2 = 1\}.$$

Further examples of abelian groups are afforded by any vector space V with the operation of vector addition.

There are two particularly important sets of examples of nonabelian groups. The first are the symmetric groups $S(X)$, where X is any set and $S(X)$ is the group of all bijective functions $f : X \to X$, with the operation being composition of functions. If $\Phi : X \to Y$ is a bijection of sets, then there is an induced bijection $\tilde{\Phi} : S(X) \to S(Y)$ by

$$\tilde{\Phi}(f) = \Phi \circ f \circ \Phi^{-1},$$

which is in fact an isomorphism of groups. $S(X)$ is a nonabelian group whenever $|X| \geq 3$.

If X is any set with $|X| = n$, then we often denote by S_n the symmetric group $S(X)$. Since the isomorphism type of $S(X)$ is completely determined by the cardinality of X, this notation is justified. It is clear that every permutation σ of a finite set X can be written as a product of transpositions (permutations which interchange two points and fix the rest). We call σ an even (odd) permutation if the product involves an even (odd) number of transpositions. (Although the product is not unique, the parity turns out to depend only on σ.) The set of all even permutations is a subgroup of $S(X)$ called the alternating group $\mathrm{Alt}(X)$ (or A_n) of cardinality $n!/2$.

The second important class of groups is the family of general linear groups $\mathrm{GL}(V)$, where V is a vector space over a field F and $\mathrm{GL}(V)$ is the group of all invertible linear operators on V, i.e., invertible linear transformations $T : V \to V$. Again the operation is composition of functions, and indeed $\mathrm{GL}(V)$ is clearly a subgroup of the symmetric group $S(V)$. $\mathrm{GL}(V)$ is nonabelian whenever $\dim(V) > 1$. As above, if $\Phi : V \to W$ is an isomorphism between the F-vector spaces V and W, i.e., an invertible linear transformation mapping V onto W, then, exactly as before, Φ induces an isomorphism of groups between the groups $\mathrm{GL}(V)$ and $\mathrm{GL}(W)$. Thus the isomorphism type of $\mathrm{GL}(V)$ is completely determined by the field F and the dimension of V as an F-vector space. Indeed, if $\dim(V) = n$ is finite, then any choice of basis for V determines a (noncanonical) isomorphism of groups between the group $\mathrm{GL}(V)$ and the multiplicative group $\mathrm{GL}(n, F)$ of all $n \times n$ invertible matrices with entries taken from the field F. (This is simply an extension of the standard identification of a linear operator $T : V \to V$ with an $n \times n$ matrix A, the identification depending on a choice of basis for V.) If $|F|$ is finite, then it may be shown that (up to isomorphism of fields) there is only one field of cardinality $|F|$ and thus the notation $\mathrm{GL}(n, q)$, where $q = |F|$, is common and unambiguous. This is no longer the case when $|F|$ is infinite.

Certain subgroups of $\mathrm{GL}(V)$ are of particular importance and were first studied in depth by C. Jordan and L. E. Dickson. They are often referred to as the classical linear groups or classical matrix groups. First is $\mathrm{SL}(V)$, the subgroup of all elements of $\mathrm{GL}(V)$ of determinant 1. The others are subgroups which preserve a geometric structure on the space V.

Definition. *A bilinear form on an F-vector space V is a function $b : V \times V \to F$ which is linear in both positions, i.e.,*

$$b(cu + dv, w) = cb(u, w) + db(v, w)$$

and

$$b(u, cv + dw) = cb(u, v) + db(u, w)$$

for all $u, v, w \in V$ and all $c, d \in F$. We say that the form is symmetric if $b(u, v) = b(v, u)$ for all $u, v \in V$. We say that the form is alternating if $b(v, v) = 0$ for all $v \in V$.

Definition: *A quadratic form on an F-vector space V is a function $Q: V \to F$ such that*

(1) $Q(av) = a^2 Q(v)$ *for all* $a \in F$, $v \in V$; *and*
(2) $b(u, v) = Q(u + v) - Q(u) - Q(v)$ *defines a bilinear form* (*necessarily symmetric*) *on* V.

If (V, b) is a space with a symmetric bilinear form and if u and v are vectors in V of length 1, i.e., $b(u, u) = 1 = b(v, v)$, then $b(u, v)$ may be thought of as the cosine of the angle between the vectors u and v. In all cases, if $b(u, v) = 0$ we say that u and v are orthogonal vectors. Note that if $b(v, v) = 0$ for all $v \in V$, then $b(u, v) = -b(v, u)$ for all $u, v \in V$, and so in both the alternating and symmetric cases, orthogonality is a symmertic relation on V.

Definition: *If (V, b) is a space with an alternating or symmetric form b, we let*

$$V^{\perp} = \{v \in V : b(u, v) = 0 \text{ for all } u \in V\}.$$

If (V, Q) is a space with a quadratic form Q and associated symmetric bilinear form b, then we let

$$\mathrm{Rad}(V) = \{v \in V^{\perp} : Q(v) = 0\}.$$

We say that a space (V, b) with an alternating form b is a symplectic space if $V^{\perp} = \{0\}$. It may be shown that a finite-dimensional symplectic space V must have even dimension and for every even positive integer $2n$ and field F, there is (up to isometry) exactly one symplectic F-space of dimension $2n$. We say that a space (V, Q) with a quadratic form Q is an orthogonal space if $\mathrm{Rad}(V) = \{0\}$. There may be many orthogonal spaces of a given dimension over a given field; the number depends on both the field and the dimension. The classical law of inertia of Sylvester asserts in particular that there are exactly $n + 1$ nonisometric orthogonal structures on a real vector space of dimension n. An orthogonal space (V, Q) is completely determined by its associated bilinear structure (V, b) when the field F has characteristic different from 2 (e.g., when F is a subfield of $\mathbf{C}$).

The symplectic and orthogonal groups are the isometry groups of symplectic and orthogonal spaces, respectively, in the following sense.

Definition: *If (V, b) is a symplectic space and $T: V \to V$ is a linear operator, then T is an isometry of (V, b) if*

$$b(u, v) = b(T(u), T(v)) \qquad \text{for all } u, v \in V.$$

The symplectic group $Sp(V)$ is the group of all isometries of the symplectic space (V, b). [*As (V, b) is unique up to isometry, it is unambiguous* (*up to isomorphism*) *to write* Sp (V).] *If (V, Q) is an orthogonal space and $T: V \to V$ is a linear operator, then T is an isometry of (V, Q) if*

$$Q(v) = Q(T(v)) \qquad \text{for all } v \in V.$$

The orthogonal group $O(V, Q)$ is the group of all isometries of the orthogonal space (V, Q).

An isometry is necessarily a bijection, whence $\mathrm{Sp}(V)$ and $O(V, Q)$ are subgroups of $\mathrm{GL}(V)$. Classical mechanics focusses on the case where $F = \mathbf{R}$ and Q is positive definite, i.e.,

$$Q(x_1, x_2, \ldots, x_n) = x_1^2 + x_2^2 + \cdots + x_n^2,$$

and so often in the physics literature the group $O(\mathbf{R}^n, Q)$ for this choice of Q is denoted simply $O(n)$. The Lorentz group, however, which arises in relativity theory, is an orthogonal group defined with respect to a four-dimensional indefinite quadratic form.

Symplectic isometries always have determinant 1. There is a somewhat larger group $\mathrm{CSp}(V)$ of symplectic similarities, i.e., linear operators $T: (V, b) \to (V, b)$ such that for all $v, w \in V$,

$$b(T(v), T(w)) = \lambda_T b(v, w),$$

for some nonzero scalar $\lambda_T \in F$. This group will reappear toward the end of Section V.

Orthogonal isometries have determinant ± 1. The subgroup of $O(V, Q)$ consisting of isometries of determinant 1 is called the special orthogonal group, $\mathrm{SO}(V, Q)$. In the case of $O(n)$, it is denoted $\mathrm{SO}(n)$ and interpreted as the group of all orientation-prserving isometries of $\mathbf{R}^n$ fixing the origin. For $n = 3$, SO(3) is the group of all rotations of $\mathbf{R}^3$ about an axis through (0, 0, 0). The earliest classification theorem for a class of nonabelian groups was the classification of finite subgroups of SO(3), achieved in the mid-nineteenth century.

Theorem: *A finite subgroup of* SO(3) *is a subgroup of the full group of symmetries of one of the following objects centered at* (0, 0, 0): *a regular polygon lying in a plane through* (0, 0, 0) *or one of the five regular polyhedra* (*Platonic solids*): *the tetrahedron*, *the cube*, *the octahedron*, *the dodecahedron*, *or the icosahedron.*

It follows that a finite subgroup G of SO(3) is either a cyclic group or is the dihedral group D_n of all symmetries of a regular n-gon or is the tetrahedral, octahedral, or icosahedral group (the group of rotational symmetries of the tetrahedron, octahedron, or icosahedron, respectively).

Finally, when the field F admits an involutory automorphism σ (i.e., $\sigma^2 = I \neq \sigma$), there is a further important

family of classical linear groups, namely, the unitary groups.

Definition: *Let F be a field and let σ be an involutory automorphism of F. A (σ-)hermitian form on an F-vector space V is a function $h: V \times V \to F$ which is linear in the first position and which satisfies*

$$h(v, w) = h(w, v)^{\sigma} \qquad \text{for all } v, w \in V.$$

A pair (V, h) with V an F-vector space and h a σ-Hermitian form on V is said to be a unitary space if $V^{\perp} = \{0\}$. We denote by $U(V, h)$ the group of all isometries of the unitary space (V, h) and call it the unitary group of (V, h). The subgroup of unitary isometries of determinant 1 is the special unitary group, $\mathrm{SU}(V, h)$.

Of course the most important of all involutory automorphisms is the complex conjugation map on $\mathbf{C}$. When h is the standard positive definite inner product on $\mathbf{C}^n$, the unitary group $U(\mathbf{C}^n, h)$ is often denoted simply $U(n)$. There are, however, other inequivalent hermitian forms on $\mathbf{C}^n$. There is a famous twofold covering map from SU(2) onto SO(3) called the spin covering because of its connection with the concept of spin in quantum mechanics.

If F is a finite field with $|F| = q^2$ for some prime power q, then F admits an involutory automorphism and it is possible to define unitary groups of isometries of finite-dimensional F-vector spaces. Up to isomorphism, in the finite field case, the unitary group is uniquely determined by the field F and the dimension n of the vector space. Unfortunately, if $|F| = q^2$, some group theorists denote this group $U_n(q^2)$, while others write $U_n(q)$. This is only one instance of the notational inconsistencies which litter the terrain of classical linear groups. The original notations of Jordan and Dickson have been largely abandoned or modified in favor of a "classical" notation dating from the time of Weyl. However, when Cartan (for the field $\mathbf{R}$) and later Chevalley and Steinberg (for arbitrary fields) constructed these groups from a Lie-theoretic standpoint, they adopted the notation of Killing, which is completely different from Weyl's notation. Moreover, even within the Weyl framework, there are treacherous inconsistencies. For instance, some write $\mathrm{Sp}_n(F)$ and others $\mathrm{Sp}_{2n}(F)$, both meaning the group of symplectic isometries of a $2n$-dimensional F-vector space. Let the reader beware.

III. BASIC CONSTRUCTIONS

In Section I we addressed the problem of decomposing a group into its constituent simple composition factors (when possible). Now we consider the opposite problem of "composing" two or more groups to create a new and larger group. One fundamental construction is the direct product construction.

Definition: *Let G and H be groups. The Cartesian product set $G \times H$ with the operation*

$$(g, h) \cdot (g', h') = (gg', hh')$$

is a group, called the direct product $G \times H$.

This construction can be iterated to define direct products $\prod G_i$ over arbitary index sets I. Another important generalization is the semidirect product. Before defining this, we must generalize the notion of an isomorphism of groups.

Definition: *Let $(G, \circ)$ and $(H, \cdot)$ be groups. A function $f: G \to H$ is a homomorphism of group provided that*

$$f(g \circ g') = f(g) \cdot f(g') \qquad \text{for all } g, g' \in G.$$

Thus an isomorphism of groups is a bijective homomorphism of groups. In general, failure of injectivity for homomorphisms is measured by the following subgroup.

Definition: *Let $f: G \to H$ be a homomorphism of groups. The kernel of f is defined by*

$$\mathrm{Ker}(f) = \{g \in G : f(g) = e_H\},$$

where e_H is the identity element of the group H.

The pre-image $f^{-1}(h)$ of each element in $f(G)$ is both a left and right coset of $\mathrm{Ker}(f) = f^{-1}(e_H)$ and so has the same cardinality as $\mathrm{Ker}(f)$. In particular, f is injective if and only if $\mathrm{Ker}(f) = \{e_G\}$, where e_G is the identity element of G. There is an intimate connection between normal subgroups and homomorphisms.

Theorem: *Let $f: G \to H$ be a homomorphism of groups with kernel K. Then K is a normal subgroup of G and f induces an isomorphism between G/K and $f(G)$. Conversely, if K is any normal subgroup of the group G, then the function $\pi: G \to G/K$ defined by*

$$\pi(g) = gK \qquad \text{for all } g \in G$$

is a homomorphism of groups with kernel K.

Now we can describe the semidirect product construction of groups.

Definition: *Let K and H be groups and let $\phi: H \to \mathrm{Aut}(K)$ be a homomorphism of groups. The semidirect product $K :^{\phi} H$ (or simply $K : H$) is the group whose underlying set is the Cartesian product set $K \times H$ with multiplication defined by*

$$(k, h)(k', h') = (k\phi(h)(k'), hh')$$

$$\text{for all } k, k' \in K, h, h' \in H.$$

The direct product $K \times H$ is the special case of the above construction in which ϕ is the homomorphism mapping every element of H to the identity automorphism

of K. In all cases, $\{(k, e_H) : k \in K\}$ is a normal subgroup of $K : H$, which is isomorphic to K. The subset $H_0 = \{(e_K, h) : h \in H\}$ is a subgroup of $K : H$ isomorphic to H, but is not in general a normal subgroup of $K : H$. Indeed, H_0 is also normal if and only if $K : H \cong K \times H$.

If all composite (i.e., not simple) groups could be constructed from proper subgroups by an iterated semidirect product construction, then the classification of all finite groups, or even all groups having a composition series, would at least be thinkable, if not doable. However, not, all composite groups can be so constructed, as is illustrated by the easy example of the cyclic group C_{p^2}. The obstruction to a composite group "splitting" as a semidirect product was first analyzed by Schur. The attempt to parametrize the set of all groups which are an "extension" of a given normal subgroup by a given quotient group was one of the motivating forces for the development of the cohomology theory of groups.

In the context of finitely generated abelian groups, a complete classification is nevertheless possible and essentially goes back to Gauss. (A group G is finitely generated if there is a finite subset S of G such that every element of G is expressible as a finite word in the "alphabet" S. We write $G = \langle S \rangle$ if G is generated by the elements of the subset S. In particular, every finite group G is finitely generated: you may take $S = G$.)

Theorem: *If A is a finitely generated abeliean group, then A is a finite direct product of cyclic groups.*

On the other hand, the enumeration of all finite p-groups is an unthinkable problem. For p a prime, we call a group P a p-group if all of the elements of P have order a power of p. By theorems of Lagrange and Cauchy, P is a finite p-group if and only if $|P| = p^n$ for some $n \geq 0$. G. Higman and Sims have established that the number $P(n)$ of p-groups of cardinality p^n is asymptotic to a function of the form $p^{\frac{2}{27}n^3 + o(n^3)}$ as $n \to \infty$.

In a different direction, another important generalization of the direct product is the central product. First we introduce an important normal subgroup of a group G.

Definition. *Let G be a group. The center, $Z(G)$, of G is defined by*

$$Z(G) = \{z \in G : zg = gz \text{ for all } g \in G\}.$$

Definition: *Let H, K, and C be groups and let $\alpha : C \to Z(H)$ and $\beta : C \to Z(K)$ be injective homomorphisms. Let*

$$Z = \{(\alpha(c), \beta(c)) \in H \times K : c \in C\}.$$

Then the central product $H \circ_C K$ is the quotient group $(H \times K)/Z$.

Sometimes the central product is (ambiguously) written $H \circ K$. Often this is done when $Z(H) \cong Z(K)$ and α and β are understood to be isomorphisms. Like the direct product, all of these product constructions can be iterated.

In some ways dual to the center of G is the commutator quotient.

Definition: *Let G be a group. If $x, y \in G$, then the commutator $[x, y] = x^{-1}y^{-1}xy$. The commutator subgroup $[G, G]$ of G is defined by*

$$[G, G] = \langle [x, y] : x, y \in G \rangle,$$

i.e., $[G, G]$ is the subgroup of G generated by all commutators of elements of G. Then $[G, G]$ is a normal subgroup of G and the commutator quotient $G/[G, G]$ is the largest abelian quotient group of G.

Note that G is an abelian group if and only if $Z(G) = G$ if and only if $[G, G] = \{e\}$. Sometimes $G/[G, G]$ is called the abelianization of G.

This leads to the definition of the following important classes of groups.

Definition: *A nonidentity finite group G is quasi-simple if $G = [G, G]$ and $G/Z(G)$ is a (nonabelian) simple group. A finite group G is semisimple if $G = G_1 \circ G_2 \circ \cdots \circ G_r$, where each G_i is a quasisimple group.*

In the contexts of Lie groups and algebraic groups, a group is called simple if every proper closed normal subgroup is finite. Thus the group $SL(n, \mathbf{C})$, whose center is isomorphic to the finite group of complex nth roots of 1, would be called a simple group by a Lie theorist and a quasi-simple (but not simple) group by a finite group theorist. There is a notion of semisimple group in the categories of linear algebraic groups and Lie groups which coincides with the definition above for connected groups.

Some slightly larger classes of groups play an important role in many areas. The following definitions are not standard. The concept of a connected reductive group is fundamental in the theory of algebraic groups, in which context it is defined to be the product of a normal semisimple group and a torus (a group isomorphic to a product of GL_1's). The second definition below approximates this notion in the category of all groups.

Definitions: *A group G is almost simple if G has a normal subgroup E which is quasi-simple and $G/Z(E)$ is isomorphic to a subgroup of* Aut(E). *A group G is almost semisimple if G has a normal subgroup $E = S \circ Z(E)$ with $S = S_1 \circ S_2 \circ \cdots \circ S_r$ semisimple (with each S_i quasi-simple) and $G/Z(E)$ is isomorphic to a subgroup of* Aut(S) *normalizing each S_i.*

Most of the classical groups described above are almost semisimple. Indeed, S is a quasi-simple group in most cases. Also, the Levi complements of parabolic subgroups

(stabilizers of flags of totally singular subspaces) of the classical linear groups are usually almost semisimple, exceptions arising over small fields or because of the peculiar structure of certain four-dimensional orthogonal groups. In a vast number of cases of importance in mathematics and the physical sciences, the symmetry group (or automorphism group) of an interesting structure is either an almost semisimple group or has the form $G = V : H$ as the semidirect product of an abelian group V and an almost semisimple group H. For example, the group of all rigid motions of Euclidean space $\mathbf{R}^n$ has the structure $\mathbf{R}^n : O(n)$. Similarly, if V is an affine space, then the group $\mathrm{AGL}(V)$ of all affine transformations of V has the structure $\mathrm{AGL}(V) = V : \mathrm{GL}(V)$.

The proof of the classification of finite simple groups necessitated in fact the classification of all finite almost-simple groups. Thus, as a consequence of the classification (along with the classification of finite abelian groups), there is essentially a complete description of all finite almost-semisimple groups. The analogous result for infinite groups is out of reach, but if one restricts to the categories of algebraic groups or Lie groups, then a description of all connected reductive groups is again a part of the fundamental classification theorems.

IV. PERMUTATION GROUPS AND GEOMETRIC GROUPS

Definition: *A group G is said to have a permutation action (or, simply, to act) on a set X if there is a homomorphism $f : G \to S(X)$. If f is injective, then we may identify G with $f(G)$, a subgroup of $S(X)$. In this case we say that G is a permutation group. In any case we may write $g(x)$ for $f(g)(x)$.*

As noted earlier, the action of G on the set G by left multiplication defines an injective homomorphism of G into $S(G)$, called the left regular representation of G. Thus every group is a permutation group.

Definition: *If G acts on the set X and $x \in X$, we say that the G-orbit containing x is the subset*

$$x^G = \{g(x) : g \in G\}.$$

The G-orbits form a partition of the set X. If $X = x^G$, we say that G acts transitively on the set X. We say that the stabilizer in G of the point x is the subgroup

$$G_x = \{g \in G : g(x) = x\}.$$

The following is a version of Lagrange's theorem.

Theorem: *If G is a finite group acting on the set X and if $x \in X$, then*

$$|G| = |G_x|\,|x^G|.$$

If G acts transitively on the set X, then the permutation action of G on X is permutation isomorphic to the action by left multiplication of G on the left coset space $G\backslash G_x$. Conversely, for any subgroup H of a group G, the left multiplication action of G on $G\backslash H$ defines a transitive permutation action of G. Thus the classification of transitive permutation actions of a group G is equivalent to the classification (up to G-conjugacy) of subgroups of G. Here G-conjugacy refers to yet another important class of permutation actions of a group G.

Definition: *Let G be a group and X a subset of G. Let*

$$X^G = \{gXg^{-1} : g \in G\}.$$

The members of X^G are called the G-conjugates of X and X^G is called the G-conjugacy class of X. G acts transitively by conjugation on X^G. The subgroup G_X is denoted $N_G(X)$ and is called the normalizer in G of X.

Thus Lagrange's theorem in this context asserts that $|G| = |N_G(X)||X^G|$. (Another notational caution: For many mathematicians, X^G would be interpreted as the subset of X fixed pointwise by G. This is not the usage in this article.)

Definition: *A transitive permutation action of a group G on a set X is said to be imprimitive if there is a partition $\mathcal{B}$ of X into sets $B_i, i \in I$, with $|I| > 1$ and $|B_i| > 1$, satisfying: for all $i \in I$ and all $g \in G$, there exists $j \in I$ such that $g(B_i) = B_j$. The partition $\mathcal{B}$ is called a system of imprimitivity and the members B_i are called blocks. If there is no such system of imprimitivity for the action of G, the action is called primitive.*

In the theory of permutation groups, primitive permutation actions roughly play the role of simple groups in the abstract theory of groups: all permutation actions may be considered to be built up out of primitive permutation actions. Thus the classification of primitive permutation groups and primitive permutation actions has received considerable attention.

An elementary but basic fact is that if G acts transitively on the set X, then the stabilizer G_x of a point $x \in X$ is a maximal subgroup of G (i.e., there is no subgroup of G properly between G_x and G) if and only if the action of G on X is primitive. Thus the classification of primitive permutation actions is equivalent to the classification (up to G-conjugacy) of the maximal subgroups of the permutation group G.

An important class of examples of primitive permutation groups is the class of affine general linear groups $\mathrm{AGL}(V) = V : \mathrm{GL}(V)$ acting naturally on the affine space V as the group of all affine transformations of V. The

subgroup V of translation maps is a regular normal subgroup, i.e., V transitively permutes the vectors of V via translation and the stabilizer V_u of any vector u is $\{I\}$, the identity function (regarded as translation by $\mathbf{0}$). Thus $\mathrm{AGL}(V)_0 = \mathrm{GL}(V)$.

In the theory of permutation groups an important role is played by the socle of G, which is the product of all minimal normal subgroups of G. There exist rather detailed theorems concerning the structure of primitive permutation groups, in particular describing the structure of the socle and its intersection with the stabilizer of a point.

Theorem (O'Nan-Scott): *Let G be a primitive permutation group on a finite set X with $|X| = n$, and let B be the socle of G. Then one of the following holds*:

1. *X may be regarded as an affine space with $n = p^m$ for some prime p, B is the abelian group of all translations of X and G is a subgroup of the affine group* $\mathrm{AGL}(X)$; *or*
2. *$B = S^k$ is the direct product of $k \geq 1$ isomorphic nonabelian simple groups acting via the product action on the Cartesian product set $X = Y^k$, where S is a primitive simple subgroup of $S(Y)$; or*
3. *$B = S^k$ is the direct product of $k > 1$ nonabelian simple groups acting either via diagonal action with $|X| = |S|^{k-1}$ and with B_x a diagonally embedded copy of S, or via twisted wreath action with $|X| = |S|^k$ and $B_x = 1$.*

The O'Nan-Scott theorem focuses attention on the primitive permutation actions of nonabelian simple groups and indeed, in the wake of the classification of finite simple groups, considerable effort has been invested in the program of determining all maximal subgroups of all finite simple groups and all closed maximal subgroups of all simple algebraic groups. There exist detailed structure theorems for these maximal subgroups, as well as extensive specific information. Besides the case-by-case enumerations of maximal subgroups for the sporadic simple groups, most of this theory has been developed in the context of simple algebraic and Lie groups by Dynkin, Seitz, Liebeck, and others. Lifting theorems, originating with Steinberg, make it possible to a certain extent to transfer maximal subgroup questions concerning finite classical linear and exceptional groups to analogous questions about algebraic groups over the algebraic closure of the finite field.

Since the mid-nineteenth century there has been considerable interest in highly transitive permutation groups.

Definition: *Let G be a permutation group on a set X and let k be a positive integer. We say that G acts k-transitively on X if for any two ordered k-tuples $(x_1, \ldots, x_k)$ and $(y_1, \ldots, y_k)$ with $x_i \neq x_j$ and $y_i \neq y_j$ for $i \neq j$, there exists an element $g \in G$ with $g(x_i) = y_i$ for all i.*

Thus transitive is the same as 1-transitive. In the 1860s, Mathieu discovered two remarkable groups, M_{12} and M_{24}, which act 5-transitively on sets of cardinality 12 and 24, respectively. Despite considerable effort, the only known proof of the following theorem is as a corollary of the classification of finite simple groups.

Theorem: *Let G be a finite k-transitive permutation group on a set X for some $k \geq 4$. Then either $G = S(X)$ with $|X| \geq 4$ or $G = \mathrm{Alt}(X)$ with $|X| \geq 6$ or $G = M_{11}$, M_{12}, M_{23}, or M_{24}. [Here M_{11} (resp. M_{23}) is the stabilizer of a point in M_{12} (resp. M_{24}).]*

Highly transitive permutation groups lead to tight combinatorial designs which may often be interpreted as error-correcting codes or dense sphere-packing lattices. Specifically, Witt described designs (or Steiner systems) $S(5, 6, 12)$ and $S(5, 8, 24)$ associated with the Mathieu groups M_{12} and M_{24}, respectively. These yield the ternary and binary Golay codes, respectively.

There is also a complete description of finite k-transitive permutation groups for $k = 2$ and 3. One important class of examples of 2-transitive permutation groups is the class of projective general linear groups $\mathrm{PGL}(V)$. For V any vector space (over a field F) of dimension at least 2, we may form the projective space $P(V)$ whose objects are the k-dimensional subspaces of V, $k \geq 1$, with incidence being given by symmetrized containment of subspaces. Then $\mathrm{GL}(V)$ acts on the points of $P(V)$ and the kernel of the action is $Z(GL(V))$, the group of scalar linear transformations on V. The projective general linear group $\mathrm{PGL}(V) = \mathrm{GL}(V)/Z(\mathrm{GL}(V))$ then acts as a 2-transitive permutation group on the set $P(V)$. The image of $\mathrm{SL}(V)$ in the symmetric group on $P(V)$ is the subgroup of $\mathrm{PGL}(V)$, denoted $\mathrm{PSL}(V)$, and it too acts 2-transitively on $P(V)$.

The lines of projective spaces are coordinatized by the elements of the field F plus one extra "point at infinity." As no element of $\mathrm{PGL}(V)$ can map three collinear points to three noncollinear points, the action of $\mathrm{PGL}(V)$ cannot be 3-transitive unless the geometry contains only one line. This is precisely the case when $\dim(V) = 2$, and indeed in this case the action is 3-transitive. If V is two-dimensional over the finite field F with $|F| = q$, then we write $\mathrm{PGL}(V) = \mathrm{PGL}(2, q)$ and $\mathrm{PSL}(V) = \mathrm{PSL}(2, q)$. If $q = 2^m$, then $\mathrm{PGL}(2, q) = \mathrm{PSL}(2, q)$, while if q is odd, then the $\mathrm{PSL}(2, q)$ is a normal subgroup of $\mathrm{PGL}(2, q)$ of index 2.

Theorem: *Let G be a finite 3-transitive (but not 4-transitive) permutation group on a set X. Then one of the following holds*:

1. *X may be regarded as an affine space with* $|X| = 2^n$ *and either* $G = \mathrm{AGL}(X)$ *or* $|X| = 16$ *and* $G = X : A_7$; *or*
2. *The socle of G is* $\mathrm{PSL}(2, q)$ *and* $X = \mathrm{PG}(1, q)$ *is the projective line of order* q, $q \geq 2$; *or*
3. $G = M_{22}$ (*the stabilizer of a point in* M_{23}).

The description of finite 2-transitive (but not 3-transitive) permutation groups G is much more elaborate. Again the examples fall into two major classes. The first are affine groups $G = V : H$, with H acting as a subgroup of $\mathrm{GL}(V)$ transitive on the nonzero vectors of V. These were enumerated by Hering, using the classification theorem for finite simple groups. There are three infinite classes and several additional small examples. In the second class, the socle of G is a nonabelian simple group. These too have been completely enumerated using the classification theorem. The socles of the generic examples are $\mathrm{PSL}(n, q)$ acting on the points of projective space, $\mathrm{Sp}(2n, 2)$ acting on spaces of quadratic forms, and the split BN pairs of rank 1 acting on the coset space $G \backslash B$. There are also several isolated examples.

At one time it was suspected that every nonabelian finite simple group had a 2-transitive permutation action on some set. This is far from the case. However, a vast number of simple groups come close in the sense of being rank k permutation groups for $k \leq 3$, as defined below.

Definition: *A permutation group G acting on a set X is said to have a rank k action on X if G is transitive on X and a point stabilizer* G_x *has exactly k orbits on X.*

A 2-transitive group G is a rank 2 group in this sense. A considerable number of finite simple groups are rank 3 permutation groups. For example, the classical projective linear groups $\mathrm{PSp}(V)$, $\mathrm{PU}(V, h)$, and $\mathrm{PO}(V, Q)$ are rank 3 in their actions on the singular points of $P(V)$ [i.e., all points in the symplectic case and those points v such that $h(v, v) = 0$, resp. $Q(v) = 0$ in the unitary and orthogonal cases respectively], provided (in the unitary and orthogonal cases) that the geometries contain totally singular planes, i.e., planes on which the form is identically 0. There are tight combinatorial conditions on rank 3 permutation actions, developed by D. G. Higman and others. The study of rank 3 permutation groups led to the discovery of several of the so-called sporadic simple groups, as discussed in Section VI.

The action of a group G on the coset space $G \backslash H$ for a single subgroup H leads to a transitive permutation action. When a family $\{G_i\}_{i \in I}$ of subgroups and coset spaces $G \backslash G_i$ are considered simultaneously, this may be regarded as a geometry with objects of type i, $i \in I$, transitively permuted by G, and with incidence defined by

$$xG_i \text{ is incident with } yG_j \text{ for } i \neq j \text{ if } x\ G_i \cap yG_j \neq \emptyset.$$

Thus, if $G = \mathrm{GL}(V)$ and G_i is taken to be the full stabilizer of a fixed i-dimensional subspace V_i, $0 < i < \dim(V)$, then we recover the projective geometry $\mathrm{PG}(V)$. Analogous constructions using stabilizers of totally singular subspaces yield the classical polar geometries for $\mathrm{PSp}(V)$, $\mathrm{PO}(V, Q)$, and $\mathrm{PU}(V, h)$. This type of construction was generalized by Tits to yield the theory of buildings and BN pairs. Tits gave an elegant axiomatic treatment for a large class of geometries, which he dubbed "buildings." Most of these geometries have large (flag-transitive) automorphism groups, and the corresponding axiomatization of these groups underlies the theory of BN pairs or Tits systems.

V. GROUP REPRESENTATIONS AND LINEAR GROUPS

A linear representation of a group G is a homomorphism $\rho : G \to \mathrm{GL}(V)$ for some vector space V. Throughout this section all vector spaces will be assumed to be over a given field F. The systematic study of finite-dimensional group representations began in the 1890s, notably in the work of Frobenius. Many of the basic ideas are analogous to those in the theory of permutation representations. Thus the decomposition of a set into G-orbits has its correspondent in the decomposition of the finite-dimensional vector space V into G-invariant subspaces V_i:

$$V = V_1 \oplus \cdots \oplus V_n,$$

where each V_i is indecomposable in the sense that it cannot be written as the direct sum of two proper G-invariant subspaces. The Krull-Schmidt theorem asserts in this context the uniqueness of the isomorphism classes of G-invariant spaces in the unordered list $V_1, \ldots, V_n$. (In the case of infinite-dimensional representations of a Lie group G, this direct sum decomposition must often be replaced by a direct integral of G-invariant subspaces of a Hilbert space. We see this already in the classical Fourier integral which arises in the representation theory of the real line $\mathbf{R}^1$.)

Immediately at this stage the representation theory of finite groups bifurcates into two cases: ordinary representation theory, where the field F has characteristic 0, and modular representation theory, where the field F has characteristic p and p divides $|G|$. (When F has characteristic p and p does not divide $|G|$, the theory is essentially the same as the characteristic 0 theory.)

Maschke's theorem: *If G is a finite group and F is a field of characteristic* 0 (*e.g., a subfield of* **C**), *then every finite-dimensional G-space V is completely reducible*

in the sense that each indecomposable summand V_i is irreducible, i.e., V_i has no proper G-invariant subspaces.

On the other hand, every finite group of order divisible by a prime p has a p-modular representation space V which is indecomposable but not irreducible. The p-modular representation theory for cyclic p-groups is precisely the theory of unipotent Jordan canonical matrices. A unipotent Jordan block represents an indecomposable module and is irreducible only if it is a 1×1 block. Similar considerations play a significant role in the representation theory of Lie groups and algebraic groups. Thus a fundamental result in Lie theory asserts the complete reducibility of finite-dimensional complex representations of complex semisimple Lie groups. By contrast, nilpotent Lie groups typically have indecomposable finite-dimensional complex representations which are not irreducible.

For the remainder of this section we shall assume that V is finite-dimensional, although many of the concepts such as characters and induced representations have wider applicability.

In characteristic 0, a G-representation ρ is completely determined by its trace function or character χ_ρ, defined by

$$\chi_\rho(g) = \mathrm{Tr}(\rho(g)) \qquad \text{for all } g \in G,$$

where Tr is the trace of the matrix $\rho(g)$. As similar matrices have the same trace, χ_ρ is a class function, i.e., it is constant on G-conjugacy classes. The (ordinary) character theory of finite groups was developed by Frobenius, Burnside, and Schur around 1900. For abelian groups, complex characters are simply homomorphisms into $\boldsymbol{C}^\times$, and they arose in the number-theoretic investigations of Legendre, Gauss, and Dirichlet. It was Dedekind whose letter to Frobenius stimulated the extension of this theory to nonabelian groups.

The following concept goes back to Galois.

Definition: *A group G is solvable if and only if there is a series*

$$G = G_0 > G_1 > \cdots > G_{n-1} > G_n = \{e\},$$

with $G_{i+1} \triangleleft G_i$ for all $i < n$ and with G_i/G_{i+1} an abelian group.

For finite groups this is equivalent to the assertion that the composition factors of G are all cyclic of prime order. Galois established that a polynomial equation is "solvable by radicals" if and only if its Galois group is a finite solvable group, extending the result of Abel mentioned in Section I. Sylow proved that a finite group of prime-power cardinality is necessarily a solvable group. One of the great early achievements of character theory was the proof of the following result.

Burnside's p^aq^b Theorem: *If G is a finite group with $|G| = p^aq^b$, with p and q primes, then G is a solvable group.*

There is also a theory of Brauer characters in the context of modular representation theory, but the Brauer characters do not determine the modular representations uniquely, except in the case of irreducible modular representations.

A fundamental construction in representation theory, introduced by Frobenius, is the induced representation. First note that for a finite group G and a field F, the F-vector space $F[G]$ with basis G becomes a ring when endowed with the unique extension of the group multiplication to $F[G]$ via the distributive law. $F[G]$ may also be regarded as the algebra of F-valued functions on G with convolution product and, as such, has natural generalizations in the categories of Lie and algebraic groups. An F-vector space V with a G-action is an $F[G]$-module in a natural way.

Definition: *Let G be a finite group, H a subgroup of G, F a field and V an $F[H]$-module. The induced $F[G]$-module V^G is the tensor product $F[G] \otimes_{F[H]} V$ with the action of $F[G]$ via left multiplication.*

There is a similar construction in the categories of Lie groups and algebraic groups, which plays a fundamental role, for example, in Harish-Chandra's construction of unitary representations of semisimple Lie groups. In this context it is often best to study representations of the associated Lie algebra, in which case the universal enveloping algebra substitutes for the group algebra.

Induced modules are analogous to imprimitive sets in the theory of permutation groups. Indeed, a G-module V is called imprimitive if $V = U^G$, where U is an H-module for some proper subgroup H of G. Otherwise V is called a primitive G-module.

A new complication in representation theory relates to the size of the field F. This is the extension of the basic problem of finding eigenvalues in the field for a matrix or linear operator. Thus the theory works most smoothly over algebraically closed fields such as $\mathbf{C}$. Sometimes, however, it is necessary to work over smaller fields and due caution is necessary. For example, an $F[G]$-module V may be irreducible but not absolutely irreducible, in the sense that V becomes reducible upon suitable extension of the field of scalars. The theory of Schur indices and Brauer groups is important in this context.

There is a further important way of "analyzing" a G-module into simpler constituents. An $F[G]$-module V is said to be tensor decomposable if $V \cong U \otimes W$ with U and W $F[G]$-modules. [Here $\otimes$ denotes the ordinary tensor product over F and $g(u \otimes w) = g(u) \otimes g(w)$.] Otherwise V is tensor indecomposable. A fundamental theorem

of Steinberg, the tensor product theorem, describes all irreducible rational modules for simple algebraic groups and all irreducible modules for finite groups of Lie type in the defining characteristic (i.e., p-modular representations, where p is the defining prime for the group) as tensor products of certain basic modules and their Galois "twists."

As tensor decomposability parallels reducibility, so too there is a notion of a tensor-induced $F[G]$-module which is analogous to the notion of an imprimitive $F[G]$-module. We can call a module tensor-primitive if it is neither tensor decomposable nor tensor-induced. Thus the "prime" object in this context is an absolutely irreducible primitive and tensor-primitive $F[G]$-module.

In the context of linear representation theory, the socle is no longer the correct subgroup to consider. For example, whereas the socle of the permutation group S_n (when $n \geq 5$) is the nonabelian simple group A_n, which captures much of the structure of S_n, the socle of the general linear group $\mathrm{GL}(n, q)$, when n divides $q-1$, is a subgroup of the group Z of scalar matrices and captures very little of the structure of $\mathrm{GL}(n, q)$. Surprisingly it was not until 1970 that the correct replacement for the socle in the context of linear groups or general finite groups was defined. Before giving the definition, we must finally discuss the important class of nilpotent groups.

Definition: *For a group G, we define recursively $G^0 = G$ and $G^{i+1} = [G^i, G]$ for $i \geq 0$. [Thus $G^1 = [G, G]$ is the commutator subgroup of G.] A group G is said to be nilpotent (of nilpotence class n) if $G^n = \{e\}$ for some positive integer n (with n the smallest such positive integer).*

Thus an abelian group is a nilpotent group of nilpotence class 1. As G^i is a normal subgroup of G for all i and as G^i/G^{i+1} is abelian, every nilpotent group is a solvable group. However, the converse is false, as shown by the symmetric group S_3. Sylow proved that every finite p-group is nilpotent, and indeed, the following theorem characterizes finite nilpotent groups.

Theorem: *A finite group G is nilpotent if and only if G is a direct product of finite p-groups.*

In the category of linear algebraic groups, the quintessential (though certainly not the only) examples of connected nilpotent groups are the groups $U(n, F)$ of all upper triangular matrices with all diagonal entries equal to 1. We call a matrix unipotent if it is conjugate to a matrix in $U(n, F)$. We have the following theorems.

Theorem: *Let U be a subgroup of $\mathrm{GL}(n, F)$ all of whose elements are unipotent matrices. Then U is a nilpotent group and is conjugate to a subgroup of $U(n, F)$.*

Lie's Theorem: *Let G be a closed connected solvable subgroup of $\mathrm{GL}(n, \mathbf{C})$. Then G is conjugate to a subgroup of $B(n, \mathbf{C})$, the solvable group of all upper triangular matrices in $\mathrm{GL}(n, \mathbf{C})$.*

Thus every subgroup of $U(n, \mathbf{C})$ is nilpotent and every closed connected nilpotent subgroup of $\mathrm{GL}(n, \mathbf{C})$ is conjugate to a subgroup of $B(n, \mathbf{C})$. Lie's theorem does not extend to arbitrary fields. For example, $\mathrm{SO}(2)$ is a closed connected abelian subgroup of $\mathrm{SL}(2, \mathbf{R})$ which is not conjugate to a subgroup of $B(2, \mathbf{R})$.

Returning to abstract groups, it can easily be shown that the product of two normal nilpotent subgroups of a group G is again nilpotent. Hence if G is a finite group, then G has a unique maximal normal nilpotent subgroup, called the fitting subgroup of G, $F(G)$. Likewise the product of two normal semisimple subgroups of a group G is again semisimple and so a finite group G has a unique maximal normal semisimple subgroup denoted $E(G)$. Moreover, the subgroups $E(G)$ and $F(G)$ commute with each other elementwise.

Definition: *The generalized Fitting subgroup $F^*(G)$ of the finite group G is the central product $F^*(G) = E(G) \circ F(G)$.*

Notice that if $F(G)$ is an abelian group, then $F^*(G)$ is an almost semisimple group. In the general theory of finite groups (or more generally of groups having a composition series), the group $F^*(G)$ plays the role that the socle plays in the theory of permutation groups. Thus, returning to the example of $\mathrm{GL}(n, q)$, although the socle is typically a subgroup of the group of scalar matrices Z, $F^*(\mathrm{GL}(n, q)) = SL(n, q) \circ Z$ captures most of the structure of $\mathrm{GL}(n, q)$.

To further appreciate the fundamental significance of $F^*(G)$ we must introduce a basic concept.

Definition: *Let X be any subset of the group G. The set*

$$C_G(X) = \{g \in G : gx = xg \text{ for all } x \in X\}$$

is a subgroup of G called the centralizer in G of X.

Definition: *Let G be a group and N a normal subgroup of G. The conjugation action of G on the elements of N defines a homomorphism $\Phi : G \to \mathrm{Aut}(N)$. The kernel of this homomorphism is $C_G(N)$. Thus $G/C_G(N)$ is isomorphic (via Φ) to a subgroup of $\mathrm{Aut}(N)$.*

Theorem (Fitting–Bender): *Let G be a finite group. Then*

$$C_G(F^*(G)) = Z(F(G)).$$

Thus $G/Z(F(G))$ is isomorphic to a subgroup of $\mathrm{Aut}(F^*(G))$.

In particular, this result says that $|G|$ is bounded by a function of $|F^*(G)|$. Specifically, $|G| \leq (|F^*(G)|)!$. (This

upper bound is achieved only when $G = S_n$ for $n \leq 4$.) In fact, much more is true in most practical applications.

Returning briefly to linear groups, we recall that the fundamental case to analyze was the following: G is a subgroup of $\mathrm{GL}(V)$, where V is an absolutely irreducible primitive and tensor-primitive G-module. In this case the following structure theorem is valid. We need one further definition.

Definition: *Let p be a prime. An extraspecial p-group E is a finite p-group such that $|Z(E)| = p$ and $E/Z(E)$ is a nontrivial abelian p-group of exponent p.*

Extraspecial p-groups are the finite cousins of the Heisenberg groups, which play a prominent role in quantum mechanics.

Theorem: *Let G be a finite subgroup of* $\mathrm{GL}(V)$, *where V is an n-dimensional vector space over the field F. Suppose that for every normal subgroup N of G not contained in $Z = Z(\mathrm{GL}(V))$, V is an absolutely irreducible primitive and tensor-primitive $F[N]$-module. Then either*

1. *$F^*(G) = E \circ (Z \cap G)$ with E an extraspecial p-group for some prime p such that the field F contains primitive p*th *roots of* 1; *or*
2. *$F^*(G) = E \circ (Z \cap G)$ with E a quasi-simple finite group and $Z(E) \leq Z$.*

A version of this theorem figures prominently in Aschbacher's analysis of the maximal subgroups of the finite classical matrix groups.

The importance of finite groups G for which $F^*(G) = E$ is an extraspecial p-group (or a close relative thereof) affords an excuse for an illustration of the use of the Fitting–Bender theorem and of linear group methods in the study of abstract finite groups. First, let E be an extraspecial p-group and let $V = E/Z(E)$. Then V may be regarded as a finite-dimensional vector space over the finite field $\mathbf{F}_p$. Moreover, we may identify $Z(E)$ with the field $\mathbf{F}_p$. Then the function $b : V \times V \to \mathbf{F}_p$, defined by

$$b(u + Z(E), v + Z(E)) = [u, v] \qquad \text{for all } u, v \in E,$$

is a nondegenerate alternating form on the vector space V. Moreover, if p is odd, then there is a surjective homomorphism $\Phi : \mathrm{Aut}(E) \to \mathrm{CSp}(V)$, the conformal symplectic group of all similarities of the symplectic space (V, b). The kernel of Φ is the group of all inner automorphisms of E, i.e., the automorphisms induced by the conjugation action of E onto itself. This group, $\mathrm{Inn}(E)$ may be identified with $V = E/Z(E)$ and then we see that $\mathrm{Aut}(E) \cong V : \mathrm{CSp}(V)$. When $p = 2$, the squaring map $Q : V \to \mathbf{F}_2$, defined by

$$Q(u + Z(E)) = u^2 \qquad \text{for all } u \in E,$$

is a quadratic form on V whose associated bilinear form is b. Now there is a surjective homomorphism $\Phi : \mathrm{Aut}(E) \to O(V, Q)$ whose kernel is again $\mathrm{Inn}(E) \cong V$. Thus $\mathrm{Aut}(E)/\mathrm{Inn}(E) \cong O(V, Q)$. In this case, however, Griess proved that the group $\mathrm{Aut}(E)$ is not in general a semidirect product. In any case, we get the following theorem.

Theorem: *Let G be a finite group such that $F^*(G) = E$ is an extraspecial p-group. Set $V = E/Z(E)$. Then G/E is isomorphic to a subgroup of* $\mathrm{CSp}(V)$. *If $p = 2$, then in fact G/E is isomorphic to a subgroup of $O(V, Q)$, where Q is the quadratic squaring map. Moreover, G/E has no nontrivial normal p-subgroup.*

Thus the structure of G/E for such groups G may be analyzed by the methods of linear representation theory. In fact, a similar line of reasoning may be applied to any finite group G such that $F^*(G)$ is a p-group for some prime p. One of the earliest and deepest results in this vein is the Hall–Higman theorem, which illustrates the relevance of the final assertion of the preceding theorem.

Hall–Higman Theorem B: *Let V be a finite-dimensional vector space over a field of characteristic p, p an odd prime. Let G be a solvable subgroup of* $\mathrm{GL}(V)$ *having no nontrivial normal p-subgroup. Let x be an element of G of order p^n. Then the minimum polynomial $m_x(t)$ of x is $(t-1)^{p^n}$, except possibly if p is a Fermat prime and G contains a nonabelian normal 2-subgroup.*

Note that p^n is the largest possible size of a Jordan block for a linear transformation of order p^n acting on a vector space in characteristic p.

The general theory of p-modular representations of finite groups was developed primarily by Richard Brauer. The sharpest results are achieved when every p-subgroup of G is cyclic, for example, when $|G|$ is divisible by p but not by p^2. A related theme of great importance is the p-modular representation theory of a finite group of Lie type $G(q)$ defined over a field $\mathbf{F}_q$ of characteristic p. The methodology here is completely different from Brauer's theory and instead is analogous to the theory of (rational) representations of semisimple Lie groups or algebraic groups. Indeed, Steinberg proved that all irreducible representations of $G(q)$ come by restriction from the irreducible rational representations over $\bar{\mathbf{F}}_q$ of the algebraic group $G(\bar{\mathbf{F}}_q)$, where $\bar{\mathbf{F}}_q$ is the algebraic closure of $\mathbf{F}_q$ and $G(q)$ arises from $G(\bar{\mathbf{F}}_q)$ by taking fixed points of a Frobenius–Steinberg endomorphism. The group $G(\bar{\mathbf{F}}_q)$ is the characteristic p analog of the Lie group $G(\mathbf{C})$. Now, the irreducible continuous representations of $G(\mathbf{C})$ are parametrized by "highest weights" and the characters of these representations are given by the Weyl character formula. Curtis and Richen showed that an analogous parametrization holds for the finite groups of Lie type. An

analog of the Weyl character formula has been conjectured by Lusztig and proved generically, but not yet for all primes p.

For applications to topology and number theory, there have been efforts to study integral representations of a finite group G, i.e., homomorphisms of G into $\mathrm{GL}(n, R)$, where R is the ring of integers of some algebraic number field, e.g., $R = Z$. The situation here is quite subtle, and only small cases can be handled effectively.

VI. FINITE SIMPLE GROUPS

Around 1890, Wilhelm Killing published a series of papers classifying the finite-dimensional semisimple Lie algebras over $\mathbf{C}$, modulo a significant error corrected by E. Cartan. In particular, via Lie's correspondence, this gave a classification of the finite-dimensional simple Lie groups over $\mathbf{C}$, and this classification was shortly extended to the real field as well by Cartan. In addition to the Lie algebras associated with the classical linear groups over $\mathbf{C}$, Killing discovered five exceptional simple Lie algebras over $\mathbf{C}$, which he denoted E_2, F_4, E_6, E_7, and E_8, the subscript denoting the maximum dimension of a diagonalizable subalgebra, later known as a Cartan subalgebra. Later E_2 came to be known as G_2. In Killing's notation the classical families of Lie groups over $\mathbf{C}$ are: $A_n(\mathbf{C}) = \mathrm{PSL}(n+1, \mathbf{C})$, $B_n(\mathbf{C}) = \mathrm{PSO}(2n+1, \mathbf{C})$, $C_n(\mathbf{C}) = \mathrm{PSp}(2n, \mathbf{C})$, and $D_n(\mathbf{C}) = \mathrm{PSO}(2n, \mathbf{C})$.

The classification theorem of Killing and Cartan, which is a fundamental precursor of the classification of the finite simple groups, is the following result.

Theorem: *Let G be a finite-dimensional simple Lie group with trivial center. Then G is either* (*an adjoint version of*) *a member of one of the four infinite families of classical Lie groups $A_n(\mathbf{C})$, $B_n(\mathbf{C})$, $C_n(\mathbf{C})$, and $D_n(\mathbf{C})$; or G is isomorphic to the adjoint version of one of the five exceptional Lie groups $E_6(\mathbf{C})$, $E_7(\mathbf{C})$, $E_8(\mathbf{C})$, $F_4(\mathbf{C})$, or $G_2(\mathbf{C})$.*

In 1955 Chevalley published a paper constructing analogs of Killing's groups over all fields. In particular, when the fields are finite, these Chevalley groups are finite simple groups, except in a few cases over fields of cardinality 2 or 3. Shortly thereafter, Steinberg did the analog of Cartan's passage to real forms of Lie groups and constructed the so-called twisted Chevalley groups or Steinberg groups ${}^2A_n(q) = \mathrm{PSU}(n+1, q)$, ${}^2D_n(q)$, ${}^3D_4(q)$, and ${}^2E_6(q)$. If V is an even-dimensional vector space over a finite field $\mathbf{F}_q$, then there are exactly two nonisomorphic orthogonal groups on V: $O^+(V)$ and $O^-(V)$, where $O^+(V)$ is the isometry group of an orthogonal space of maximal Witt index (i.e., an orthogonal sum of hyperbolic planes). Then, for $n \geq 3$, $D_n(q)$, and ${}^2D_n(q)$ are the unique nonabelian simple composition factors of the groups $O^+(V)$, resp. $O^-(V)$, where V is an n-dimensional $\mathbf{F}_q$-space. The groups ${}^3D_4(q)$ and ${}^2E_6(q)$ were new families of nonabelian finite simple groups. ${}^2E_6$ had arisen already as a real form of E_6 in Cartan's classification, but the ${}^3D_4$ groups were truly new, arising as the fixed points on $D_4(F)$ of $\rho \circ \sigma$, where ρ is the special triality automorphism of the D_4 oriflamme geometry and σ is a field automorphism of F of order 3. [Note: only involutory automorphisms arise in the classification of real forms of Lie groups, since $(\mathbf{C} : \mathbf{R}) = 2$.]

As 1960 dawned, the known nonabelian finite simple groups were the alternating groups of degree at least 5, the Chevalley and Steinberg groups (including the classical linear groups), and the five Mathieu groups, which had been dubbed sporadic groups by Burnside around 1900. Each of these groups has even order, and indeed it had been conjectured as early as 1900 by Miller and Burnside that groups of odd order were necessarily solvable groups. Even more, each of these groups has order divisible by 6. In fact, they all have subgroups isomorphic to either $\mathrm{SL}(2, 3)$ or $\mathrm{PSL}(2, 3) \cong A_4$, with the exception of the 2-transitive simple permutation groups $\mathrm{PSL}(2, 2^n)$ for n odd, $n \geq 3$.

In 1960 Michio Suzuki discovered the first known nonabelian simple groups of order not divisible by 3, the infinite family of 2-transitive simple permutation groups $Sz(2^n)$, n odd, $n \geq 3$. Shortly thereafter, Ree showed that a modification of Steinberg's twist could be apply to the Chevalley groups B_2 and F_4 over fields of order 2^n, n odd, and to G_2 over fields of order 3^n, n odd, to produce the families of simple groups ${}^2B_2(2^n)$, ${}^2F_4(2^n)$, and ${}^2G_2(3^n)$, n odd. In particular, the groups ${}^2B_2(2^n)$ were precisely the Suzuki groups $\mathrm{Sz}(2^n)$. The other two families were new and were dubbed Ree groups. The finite Chevalley, Steinberg, Suzuki, and Ree groups are now often called the finite groups of Lie type, inasmuch as they can be uniformly constructed as the fixed points of Frobenius–Steinberg endomorphisms of simple algebraic groups over algebraically closed fields of prime characteristic, and these groups are analogous to the simple Lie groups over $\mathbf{C}$ classified by Killing and Cartan.

Meanwhile, in 1960–1962, Walter Feit and John G. Thompson wrote the most remarkable paper in the history of finite group theory, proving the Miller-Burnside conjecture.

The Odd-Order Theorem (Feit–Thompson): *Every finite group of odd order is solvable.*

This paper and the successor "N-Group Paper" of Thompson provided most of the fundamental ideas for the classification of the finite simple groups. The first tool in the analysis of finite simple groups was provided in 1872 by Ludvig Sylow.

Sylow's Theorem: *If G is a finite group with $|G| = p^n m$ for some prime p, with m not divisible by p, then G has subgroups of order p^k for all $k, 0 \le k \le n$, and all subgroups of G of order p^n are G-conjugate.*

The maximal p-subgroups of G are called Sylow p-subgroups of G. It is interesting to note that there are somewhat analogous results in the category of linear algebraic groups, namely, the conjugacy of Borel subgroups (maximal closed connected solvable subgroups) and of maximal tori (maximal connected semisimple abelian subgroups) of the linear algebraic group G. Note that a Borel subgroup $B = N_G(U) = UT$, where U is a maximal unipotent subgroup and T is a maximal torus. If G is defined over an algebraically closed field of characteristic p, then U is a maximal p-subgroup of G, in the sense that U is maximal with the property that every element of U has order a power of p.

Returning to finite group theory, the set of all normalizers and centralizers of nonidentity p-subgroups of the finite group G is called the set of p-local subgroups of G, and the study of G via the analysis of these subgroups is called p-local analysis. The "N-Group Paper" combined with the "Odd Order Paper" completed the classification of finite simple groups all of whose p-local subgroups are solvable groups. The strategy in brief is to choose a prime p and a Sylow p-subgroup P for which the p-rank (the dimension of an elementary abelian p-subgroup, thought of as a vector space over $\mathbf{F}_p$) is as large as possible. A theorem of Burnside, using monomial representations, guarantees that p may be chosen so that either the p-rank is at least 3 or $p = 2$ and $|G|$ is a multiple of 12. One now studies the normalizers of all nonidentity subgroups of P. If G is going to turn out to be a finite group of Lie type defined over a field of characteristic r, then almost always the chosen p will turn out to be r and the geometry defined by the coset spaces $G\backslash M_i$, where M_i is a p-local subgroup of G containing P, will be precisely the Tits building for the group G. If there are at least two maximal p-local subgroups M_i containing P, then the Tits building will have rank at least 2 and the geometry will be rich enough to characterize the group G. More accurately, finite buildings of rank at least 3 were classified in a major paper of Tits. Finite rank 2 buildings include all finite projective planes. However, the buildings of rank at least 2 which occur in the context of the simple group classification satisfy a "Moufang condition," which permits their complete identification. On the other hand, if there is a unique maximal p-local subgroup M of G containing P, then the "geometry" is simply the point set $G\backslash M$. This leads to a problem in permutation group theory.

Problem: *Classify finite transitive permutation groups G in which, for some fixed prime p, every nonidentity p-element fixes exactly one point.*

For $p = 2$, this problem was solved by a combination of independent results of Suzuki and Bender.

Theorem: *Let G be a finite transitive permutation group in which every involution (element of order 2) fixes exactly one point. Then either G has 2-rank at most 1 or $G \cong SL(2, 2^n)$, $\mathrm{Sz}(2^{2n-1})$, or $\mathrm{PSU}(3, 2^n)$ for some $n \ge 2$.*

For groups of odd order, the problem was solved by Feit and Thompson in the "Odd Order Paper." The theory of group characters plays a major and seemingly unavoidable role in this delicate analysis. The complete problem has been solved as a corollary of the classification of finite simple groups, but an independent solution for p odd remains elusive.

A somewhat different strategy for the classification of finite simple groups was pioneered by Richard Brauer starting around 1950. For groups of even order, Brauer championed an inductive approach focusing on the centralizer $C_G(t)$ of an involution t. A philosophical underpinning for this approach was provided by the following result.

Brauer–Fowler Theorem: *Let G be a finite simple group of even order containing the involution t. If $|C_G(t)| = c$, then $|G| < (c^2)!$.*

The Brauer–Fowler bound is useless in practice, but it does establish that given a finite group H with center of even order, the problem of finding all finite simple groups G containing an involution t with $C_G(t) \cong H$ is a finite problem. In practice, as Brauer, Janko, and their students demonstrated in the ensuing decades, the problem is not only finite, it is doable. Indeed, a posteriori, $C_G(t)$ usually determines G uniquely. Once Feit and Thompson proved that every nonabelian finite simple group has even order, Brauer's strategy became an inductive strategy for the remainder of the classification proof. Brauer's strategy is significantly different from the Feit–Thompson strategy in the following sense: since most primes are odd, a finite simple group of Lie type is almost always defined over a field of characteristic τ with $\tau \neq 2$. Thus, for most finite simple groups G, each involution t is a semisimple element in the sense that $F^*(C_G(t))$ is an almost semisimple group. Fortunately, the two approaches mesh and complement each other, and the current proof is a blending of both.

In the course of pursuing Brauer's strategy, Janko investigated many possible specific structures for the centralizer of an involution in a simple group. One of these structures, $C_2 \times A_5$, led him to a new simple group, J_1, the first sporadic group discovered since Mathieu. Shortly thereafter, Janko discovered two more sporadic groups, J_2 and J_3. The former was soon constructed as a rank-3 permutation group by M. Hall. The involution centralizer approach and the rank-3 permutation group approach were the main tactics leading to the discovery of 21 sporadic groups in the decade 1965–1975.

In a different vein, John Leech constructed in 1967 the Leech lattice, the densest sphere-packing lattice in $\mathbf{R}^{24}$, based on the Steiner system $S(5.8.24)$ for M_{24}. John Conway investigated the automorphism group .0 of the Leech lattice and discovered that $.0/\{\pm 1\} = \mathrm{Co}_1$ was a new finite simple group, as were two of its subgroups, Co_2 and Co_3. In 1974, pursuing the involution centralizer philosophy, starting with a centralizer C built out of the Leech lattice modulo 2, $\Lambda/2\Lambda$, and Co_1, Bernd Fischer and Robert L. Griess, Jr., independently discovered the largest of the sporadic simple groups, dubbed **M**, the Monster. The Monster was constructed by Griess as the automorphism group of a 196, 884-dimensional nonassociative commutative **C**-algebra, the Griess algebra. Twenty-one of the 26 sporadic simple groups are contained (as quotients of subgroups) in the Monster. The investigation of numerical mysteries (dubbed Monstrous Moonshine) associated with **M** led Frenkel, Lepowsky, Meurman, and Borcherds to develop the mathematical theory of vertex operator algebras, objects which were first studied by physicists.

The classification theorem for finite simple groups is the following theorem.

Classification Theorem: *Let G be a finite simple group. Then G is isomorphic to a member of one of the following families of simple groups*:

1. *The cyclic groups of prime order*;
2. *The alternating groups of degree at least* 5;
3. *The finite simple groups of Lie type*; *and*
4. *The* 26 *sporadic simple groups.*

There are a small number of unusual isomorphisms among the simple groups. For example, $A_5 \cong \mathrm{SL}(2, 4) \cong \mathrm{PSL}(2, 5)$, $A_6 \cong \mathrm{PSL}(2, 9) \cong [\mathrm{Sp}(4, 2), \mathrm{Sp}(4, 2)]$, and $A_8 \cong \mathrm{SL}(4, 2)$.

The proof is a massive inductive argument, i.e., one assumes througout that G is a minimal counterexample to the statement of the theorem, and so every composition factor of every proper subgroup of G (in particular, of every p-local subgroup of G) is one of the listed simple groups. Having chosen a prime p upon which to perform p-local analysis according to either the Feit–Thompson or Brauer strategy, one is confronted by the major problem of attempting to establish that for many of the maximal p-local subgroups H of G, either $F^*(H)$ is an almost semisimple group or $F^*(H)$ is a p-group. In the latter case we say that H is of parabolic type, since in a finite simple group G of Lie type in characteristic p, all subgroups which contain a Sylow p-normalizer (the so-called parabolic subgroups of G) are of parabolic type. Aschbacher and Timmesfeld were leaders in the analysis of such groups in the 1970s. Later, Stroth, Stellmacher, and Meierfrankenfeld also pursued this vein.

The principal obstruction to establish this structure theorem for maximal p-locals is the hypothetical existence of certain p'-subgroups (subgroups of order relatively prime to p) whose normalizers contain large p-subgroups. These were dubbed p-signalizers by Thompson, who developed the principal strategies for controlling them. These strategies were honed into a "signalizer functor theory" by Gorenstein and Walter with major contributions by Goldschmidt, Glauberman, Harada, Lyons, McBride, and others.

VII. OTHER TOPICS

Thus far the article has focused on finite groups, with some reference to Lie groups and linear algebraic groups. Certain other classes of groups have played a fundamental role in mathematics. Klein and Poincaré pioneered the study of discrete subgroups of PSL(2, **R**) and PSL(2, **C**). For instance, PSL(2, **R**) acts naturally as fractional linear transformations on the upper half-plane of $\mathbf{R}^2$, which may be identified with the hyperbolic plane with PSL(2, **R**) acting as orientation-preserving isometries. A discrete subgroup is then a subgroup Γ which does not contain arbitrarily small translations or rotations. A similar interpretation may be given to discrete subgroups of PSL(2, **C**) viewed as isometries of hyperbolic 3-space. An important class of examples of these discrete subgroups is given by PSL(2, **Z**) and its congruence subgroups, i.e., the kernels of the natural homomorphisms of PSL(2, **Z**) onto PSL(2, $\mathbf{Z}/\mathbf{n}\mathbf{Z}$), where $\mathbf{Z}/\mathbf{n}\mathbf{Z}$ is the ring of integers modulo n. These discrete groups may be regarded as the symmetry groups of tessellations (tilings) of hyperbolic space by congruent (hyperbolic) polygons. This theory has been generalized to the study of arithmetic subgroups of higher-dimensional Lie groups. Like the integers, these groups do not have a composition series and so they cannot be "built up from the bottom" like finite groups: they have no bottom from which to build. They are, however, usually generated by a finite set of elements or proper subgroups and may be viewed as a quotient of the universal or freest object so generated.

Thus, if G is generated by the set of elements S, then G is a quotient of the free group $F(S)$. A relator is a word which becomes 1 when interpreted in G rather than in $F(S)$. In general, if R is a set of relators, then there is a natural homomorphism from $F(S)/N(R)$ onto G, where $N(R)$ is the smallest normal subgroup of $F(S)$ containing all of the members of R. If that homomorphism is an isomorphism, then the generators S and relations R "present" the group G. In general, it is quite difficult to "understand" a group given by generators and relations. In fact, it is a theorem (the "Undecidability of the Word Problem") that there is no

recursive algorithm for deciding whether a given finitely presented group is the identity group.

A striking exception to the intractability of the "Word Problem" is the family of Coxeter groups $C(m_{ij})$ defined by the generators $\{s_i\}_{i\in I}$ and relators $(s_i s_j)^{m_{ij}}$, where $M = (m_{ij})$ is a symmetric matrix with $m_{ii} = 1$ for all i and with $m_{ij} \in \{2, 3, \ldots, \omega\}$ for $i \neq j$. Classical examples of Coxeter groups are discrete groups of isometries of spherical, Euclidean, or hyperbolic n-spaces. The spherical Coxeter groups are well understood and form the skeletons of Tits' spherical BN pairs. Beyond the finite dihedral groups, all but two of the spherical Coxeter groups are finite Weyl groups, whose classification is a key step in the modern proof of the Killing-Cartan classification theorem. The Euclidean Coxeter groups are also well understood and play a major role in the theory of affine buildings and p-adic groups. Closely associated to Coxeter groups are the braid groups and Artin groups, whose generators have infinite order but which satisfy relations similar to the Coxeter relations. They have featured in recent work in numerous areas, including knot theory, singularity theory, and inverse Galois theory, i.e., the problem of finding a polynomial (usually in $\boldsymbol{Q}[x]$) having a specified Galois group.

The braid groups arise as the fundamental groups of topological spaces, an important concept introduced by Poincaré. The elements of the fundamental group are equivalence classes of loops (closed paths) in the space with a distinguished base point. Multiplication is concatenation of loops and inversion is reversal of direction. Fundamental groups are naturally described via generators and relations, with the generators being certain loops and a relation being a closed path which can be contracted to the base point in the space. For example, if the space X is the union of n circles which intersect only at a common base point, then the fundamental group of X is the free group on n generators. The fundamental group acts naturally on the universal cover $\tilde{X}$ of the space X. In this example, $\tilde{X}$ is a regular tree (graph without circuits) of valency $2n$.

The study of fundamental groups of topological spaces attached along embedded subspaces motivates the group-theoretic study of so-called free products with amalgamation, HNN extensions, and group actions on trees, all of which is subsumed in the Bass–Serre theory of graphs of groups. The objective is understand the universal completion of a graph of groups, i.e., the largest possible group containing certain specified subgroups (attached to the vertices and edges of the graph) subject to certain relating homomorphisms mapping edge groups to vertex groups. The amalgamated product of two groups over a specified common subgroup is a "free" construction, which in all nontrivial cases leads to an infinite completion.

The free product $H * K$ of two groups H and K is the freest group generated by H and K with $H \cap K = \{e\}$. Underlying the Brauer–Fowler theorem on finite simple groups is the elementary fact that the free product $C_2 * C_2$ is an infinite dihedral group generated by an element x of infinite order and an involution y such that $yxy = x^{-1}$. All finite homomorphic images of $C_2 * C_2$ are finite dihedral groups (the symmetry groups of regular polygons) or are cyclic of order 1 or 2. By contrast, the free product $C_2 * C_3$ is isomorphic to PSL(2, $\mathbf{Z}$) and almost every finite simple group occurs as a homomorphic image of $C_2 * C_3$. PSL(2, $\mathbf{Z}$) is in turn the quotient of the 3-string braid group B_3 by an infinite cyclic central subgroup.

It follows from the theory of fundamental groups and covering spaces that the study of discrete subgroups (at least the torsion-free ones, i.e., those having no nontrivial elements of finite order) of PSL(2, $\mathbf{R}$) and PSL(2, $\mathbf{C}$) is essentially identical to the problem of finding Riemannian metrics of constant (sectional) curvature -1 on manifolds of dimensions 2 and 3, respectively. This suggests that one should be especially interested in studying fundamental groups of negatively curved Riemannian manifolds (or negatively curved spaces). Quite recently, Gromov has developed a theory of "word hyperbolic groups" which encompasses the group-theoretic aspects of negative curvature. For example, free groups as well as the fundamental groups of negatively curved manifolds are all word hyperbolic. In contrast to the general undecidability of the Word Problem, it can be shown that the Word Problem is solvable for word hyperbolic groups. Indeed, this has led to a study of the larger class of "automatic groups" for which the Word Problem is decidable by a finite-state automation.

Recently there have been projects to develop efficient computer algorithms for the identification of finite groups from data consisting either of a set of generating permutations or a set of generating matrices or an abstract set of generators and relators (a black-box group). Most of these algorithms use the classification of finite simple groups as well as numerous corollaries concerning simple groups with small linear or permutation representations. Computer algorithms played a relatively small role in the classification of finite simple groups. The existence of most of the sporadic simple groups can be established as a corollary of the existence of the Monster, which Griess constructed "by hand." Several of the sporadic groups were first constructed on a machine, however, and for a few this remains the only proof of their existence.

SEE ALSO THE FOLLOWING ARTICLES

GROUP THEORY, APPLIED • SET THEORY

BIBLIOGRAPHY

Borel, A. (1991). "Linear Algebraic Groups," 2nd ed., Springer-Verlag, Berlin.

Bridson, M., and Haefliger, A. (1999). "Metric Spaces of Non-positive Curvature," Grundlehren der mathematischen Wissenschaften 319, Springer-Verlag, Berlin.

Curtis, C., and Reiner, I. (1987). "Methods of Representation Theory I, II," Wiley-Interscience, New York.

Gorenstein, D. (1982). "Finite Simple Groups: An Introduction to Their Classification," Plenum, New York.

Gorenstein, D., Lyons, R., and Solomon, R. (1994). "The classification of the Finite simple groups, numbers 1–3," *AMS Math. Surv. Monogr.* **40** (1–3).

Huppert, B., and Blackburn, N. (1982). "Finite Groups II, III," Springer-Verlag, Berlin.

Knapp, A. (1996). "Lie Groups: Beyond an Introduction," Birkhauser, Boston.

Suzuki, M. (1977), "Group Theory I, II," Springer-Verlag, Berlin.



Group Theory, Applied

Nyayapathi V. V. J. Swamy
Oklahoma State University

Satyanarayan Nandi
Oklahoma State University

I. Brief History
II. Group Representation Theory
III. Continuous Lie Groups
IV. Applications in Atomic Physics
V. Applications in Nuclear Physics
VI. Applications in Molecular and Solid-State Physics
VII. Application of Lie Groups in Electrical Engineering
VIII. Applications in Particle Physics
IX. Applications in Geometrical Optics
X. The Renormalization Group

GLOSSARY

Brillouin zone Volume in k space (reciprocal lattice space) bounded by planes of energy discontinuities.
Character Trace of a matrix in a representation.
Color Physical property of the strong interactions [SU(3)] analogous to electric charge in electrodynamics.
Critical point The highest pressure and the highest temperature at which a liquid and gas can coexist in equilibrium. This is the point at which the difference between the density of the liquid phase and that of the gaseous phase becomes zero.
Group Finite or infinite set of elements which have certain properties defining a group.
Isospin Symmetry [SU(2)] which elementary particles obey if the electromagnetic interaction is neglected; this accounts for the charge multiplets which are observed in particle physics.
Lie algebra Certain commutation relations obeyed by the generators of a continuous Lie group.
Normal modes Fundamental vibrations into which any vibrational mode of molecule can be decomposed.
Particle physics Study of the fundamental constituents of matter and their interactions.
Quark model Model in which all hadrons (strongly interacting particles) are described in terms of fundamental particles called quarks.
Representation Set of matrices which represents the elements of a group.
SU(2), SU(3) Lie groups whose elements are unitary matrices in two and three dimensions with determinant $+1$.

APPLIED GROUP THEORY is the study of the theory of group representations and its applications in various areas of physics such as atomic physics, molecular physics, solid-state physics, and particle physics. Group theory is essentially a mathematical description of symmetry in nature, and it is very interesting that this symmetry plays an important role in understanding many experimental phenomena in physics.

I. BRIEF HISTORY

Ancient as well as modern works of art and architecture tell us that humans have long familiarity with symmetry. An Egyptian pyramid, a Greek temple, an Indian Taj Mahal, and a Chinese pagoda strike us as beautiful examples of symmetry. There is symmetry in works of nature as well, in living organisms, and in people. It is interesting to note, however, that while knowledge of symmetry existed for centuries, it was only toward the end of the eighteenth century that it was realized that symmetry has a scientific basis in the mathematical theory of groups. Since the mathematical structure and analysis of groups is the subject matter of another article in this *Encyclopedia*, we will discuss here only the applications of group theory in physics and engineering.

In a series of papers published toward the end of the nineteenth century and at the beginning of the twentieth century Frobenius and Schur laid the foundation of the theory of group representations for finite groups. The structure and representation of continuous groups is the work of Lie, Cartan, and Weyl. Wigner was probably the first to recognize the importance of Frobenius's work to quantum mechanics and modern physics. He applied representation theory to physical problems in many areas of physics. Subsequently several others played important roles in developing and applying group theory as a tool—Bargmann, Bethe, Casimir, Gell-Mann, Pauli, to name a few.

II. GROUP REPRESENTATION THEORY

It is necessary to discuss briefly some important theorems in representation theory, since all the applications in the physical sciences are based on these theorems. First, a Cayley multiplication table (Table I) defines a group abstractly. In the table the operation of "multiplication" is to be understood as, for instance, the correspondences $AB = D$ and $BA = F$. A word of caution is in order here. There is usually a notational difference between mathematicians and physicists. Here the operation AB is to be performed from right to left (i.e., the operation A to be made after the operation B, akin to the left multiplication of a matrix B by a matrix A). A function of an independent variable x is written $F(x)$. These notations are consistent with the customary usage in group theory books by physicists (e.g., Wigner, 1959). The set of six elements form a finite group or order $g = 6$ (G or S_3 or C_{3v}), and from the table one can see that it is closed with respect to multiplication. There exists an identity element (E) and every element has an inverse element such that the product of the element and its inverse (or vice versa) is the identity. If the multiplication is commutative (i.e., $AB = BA$) the group is known as an abelian group.

TABLE I Cayley Multiplication Table of C_{3v} or S_3

	E	A	B	C	D	F
E	E	A	B	C	D	F
A	A	E	D	F	B	C
B	B	F	E	D	C	A
C	C	D	F	E	A	B
D	D	C	A	B	F	E
F	F	B	C	A	E	D

This abstract group multiplication table can be "realized" in more than one way by different types of elements with a multiplication rule appropriately defined. For instance, the following numbers satisfy all the group properties under ordinary multiplications:

$$E = 1 = D = F, \qquad A = B = C = -1. \tag{1}$$

This, for instance, is an abelian group. Another realization of the group table is obtained by elements which are operations of permutations of three objects and this group is known as the symmetric group S_3. Explicitly, the elements are

$$\begin{matrix} E \\ \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix}, \end{matrix} \quad \begin{matrix} A \\ \begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{bmatrix}, \end{matrix} \quad \begin{matrix} B \\ \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}, \end{matrix}$$

$$\begin{matrix} C \\ \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{bmatrix}, \end{matrix} \quad \begin{matrix} D \\ \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{bmatrix}, \end{matrix} \quad \begin{matrix} F \\ \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}. \end{matrix} \tag{2}$$

The permutation D is defined to read "1 is replaced by 2, 2 by 3, and 3 by 1." AB is defined as the result of the permutation corresponding to A performed after the operation B. Yet another way of realizing this group, which is of importance in molecular physics, is by choosing for the elements certain geometric symmetry operations. This group is known as C_{3v}. These symmetry operations are

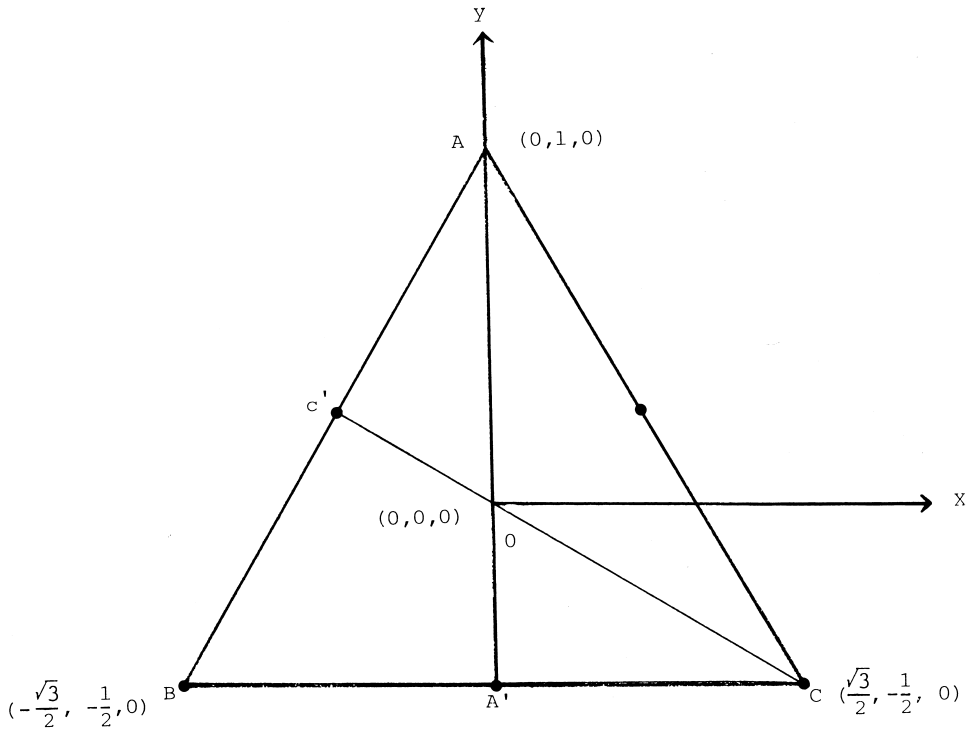


FIGURE 1 Equilateral triangle (C_{3v}) with coordinates of vertices. [Reprinted with permission from Swamy, N. V. V. J., and Samuel, M. A. (1979). "Group Theory Made Easy for Scientists and Engineers," Wiley-Interscience, New York. Copyright 1979 John Wiley and Sons.]

easily understood by applying them to the equilateral triangle in Fig. 1. In this A, B, C are the vertices, A', B', C' are the midpoints of the opposite sides such that AA', BB', CC' are the medians of the triangle. Let us assume the triangle to be in the XY plane of a rectangular coordinate system whose origin is at the centroid of the triangle O. Here the identity element means "leave it alone." Elements D,F correspond to rotations through 120° and 240°, respectively, about the Z axis through O perpendicular to the plane of the triangle. These are called C_3, C_3^2 operations, the subscript 3 referring to the angle of rotation $2\pi/3$ and the letter C stands for the cyclic axis. A, B, C are operations of reflections in three vertical mirror planes passing through AA', BB', CC' and containing the cyclic axis. Although we labeled the vertices to be able to understand the geometric symmetry operations of rotations and reflections, it is important to note that the operations bring the triangle into coincidence with itself, which implies that all the vertices are identical and hence indistinguishable. The ammonia molecule NH_3 has the symmetry of the C_{3v} group. The three hydrogen atoms are at the vertices of the triangle and the nitrogen atom sits somewhere on the cyclic axis and thus all the six geometric operations bring the molecule into coincidence with itself. The groups S_3 and C_{3v} are said to be isomorphic, which means that the elements E, A, B, ... of S_3 are in one-to-one correspondence with elements E, σ_v, σ_v', ... of C_{3v} uniquely such that group multiplication is preserved. $AB = D$ in the case of the group S_3 means that the product of the elements corresponding to A and B in C_{3v}, in that order, results in the element corresponding to D. Here multiplication means the group operation. Isomorphic groups have the same Cayley multiplication table. A many-to-one correspondence is known as homomorphism. In this case more than one element of one group corresponds to the same element in the other group, but products do correspond to products in either group. Thus the group of numbers in (1) is homomorphic to S_3 or C_{3v}.

We notice from the Cayley table for S_3 that subsets of elements (E), (E, A), and (E, D, F) themselves satisfy all the group properties. These are known as subgroups of the main group, of orders $h = 1$, 2, and 3, respectively. There is a theorem of Lagrange which says that the order h of a subgroup is a divisor of the order g of the group. Thus a group or order 11 can have only one subgroup, of order 1, the trivial group with only the identity element. Given an element A, one defines its conjugate of BAB^{-1}, where B is any element of the group. For instance, in Table I, $DAD^{-1} = DAF = DC = B$. Thus B is a conjugate of A, the operation of conjugation

resembling a similarity transformation in matrix theory. A set of self-conjugate elements is called a class, and a group is divided into classes. Thus, A, B, C form a class because the conjugate of A, XAX^{-1}, is again one of the elements of this set no matter which element of the group X is. For any finite group the identity element forms a class all by itself. A subgroup which has whole classes of the original group as its elements is known as an invariant subgroup. Thus for the above group the subgroup E, D, F is an invariant subgroup consisting of the classes $K_1(E)$ and $K_2(D,\ F)$. If a group does not have an invariant subgroup it is called a simple group. An isomorphism of a group with itself (i.e., a one-to-one correspondence between elements of the group preserving multiplication) is called an automorphism. If this is brought about by conjugation it is called an inner automorphism. For instance, the set E, A, B, C, D, F is an inner automorphism induced by the identity as the conjugating element. This is mapped into E, B, A, C, F, D if the conjugation is done by C. This then is an inner automorphism induced by C.

A. Representations and Characters

If a Cayley multiplication table is realized by choosing the elements as (unitary) matrices this set is known as a representation of the group. Frobenius and Schur did pioneering work in demonstrating quite a few important theorems in representation theory. For instance, the following set of 3×3 matrices is a representation Γ of S_3 or C_{3v}, of dimension 3:

$$\begin{matrix} E \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{matrix}, \quad \begin{matrix} A \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \end{matrix}, \quad \begin{matrix} B \\ \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{matrix},$$
$$\begin{matrix} C \\ \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{matrix}, \quad \begin{matrix} D \\ \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{matrix}, \quad \begin{matrix} F \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{matrix}. \tag{3}$$

Here obviously "multiplication" means matrix multiplication. It is easy to see, however, that another set of matrices, obtained from the one given through a similarity transformation, is also a representation of the group. Two sets of matrices which differ only through a similarity transformation are said to be equivalent representations of the group. Furthermore, this Γ is known as a reducible representation because the following one matrix S generates an equivalent representation through a similarity transformation:

$$\mathbf{S} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \\ 2/\sqrt{3} & -1/\sqrt{6} & -1/\sqrt{6} \end{bmatrix},$$

$$S\mathcal{D}(A)S^{-1} \equiv A' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= (1) \oplus \begin{matrix} B' \\ \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \end{matrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & \sqrt{3}/2 \\ 0 & \sqrt{3}/2 & -1/2 \end{bmatrix}$$

$$\times \begin{matrix} C' \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & -\sqrt{3}/2 \\ 0 & -\sqrt{3}/2 & -1/2 \end{bmatrix} \end{matrix} \begin{matrix} D' \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1/2 & -\sqrt{3}/2 \\ 0 & \sqrt{3}/2 & -1/2 \end{bmatrix} \end{matrix}$$

$$\times \begin{matrix} F' \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1/2 & \sqrt{3}/2 \\ 0 & -\sqrt{3}/2 & -1/2 \end{bmatrix} \end{matrix}. \tag{4}$$

(Boldface letters represent operators; letters with arrows represent vectors.) This latter set is seen to be a direct sum of two matrices of dimensions 1 and 2. It can be shown that the two-dimensional representation cannot further be reduced and is therefore called an irreducible representation. It is the irreducible representation of a group that is of paramount importance in physics. The sums of the diagonal elements of the matrices in an irreducible representation, or the traces of the matrices, are called characters χ. The relevant irreducible representation and characters of S_3 are rewritten as

$$\begin{matrix} E \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ \chi = 2 \end{matrix}, \quad \begin{matrix} A \\ \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \\ \chi = 0 \end{matrix}, \quad \begin{matrix} B \\ \begin{bmatrix} 1/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{bmatrix} \\ \chi = 0 \end{matrix},$$
$$\begin{matrix} C \\ \begin{bmatrix} 1/2 & -\sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{bmatrix} \\ \chi = 0 \end{matrix}, \quad \begin{matrix} D \\ \begin{bmatrix} -1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{bmatrix} \\ \chi = -1 \end{matrix}, \tag{5}$$
$$\begin{matrix} F \\ \begin{bmatrix} -1/2 & \sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{bmatrix} \\ \chi = -1 \end{matrix}$$

It is interesting to note that the representation can be a homomorphism and need not necessarily be an isomorphism. An important lemma, called Schur's lemma, says that no matrix other than a multiple of the identity matrix commutes with all the matrices in an irreducible representation. This therefore gives a red-litmus-blue type of test for checking whether a given representation is reducible or irreducible. One of the interesting representations is known as the regular representation. From the multiplication table for the S_3 group of order 6 (Table I) we know, for instance, that $DA = C$. If we consider this result as a linear combination of all the elements we get

$$DA = OE + OA + OB + 1C + OD + OF.$$

If we express the other operations in the Cayley table (i.e., DB, DC, DD, DF, and DE) we can write the result in matrix form. The transpose of this matrix is the representation matrix of the element D in the regular representation

$$\mathfrak{D}(D) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{6}$$

The matrices representing other elements of S_3 are obtained similarly. This is, however, a reducible representation. One matrix, by a similarity transformation, can reduce all the six matrices into block-diagonal form as a direct sum of lower-dimensional matrices. It can be shown that in the reduction every irreducible representation occurs as many times as its dimensionality. This regular representation decomposes into two one-dimensional and two two-dimensional representation matrices, because the S_3 group happens to have two one-dimensional and one two-dimensional representation.

We take the symmetric group S_4 (permutations on four objects) as an example to illustrate important theorems on representations and characters. The multiplication table for this group of order 24 is given in Table II. An interesting

TABLE II Cayley Multiplication Table for S_4

	E	*A*	*B*	*C*	*D*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*	*X*	
E	*E*	*A*	*B*	*C*	*D*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*	*X*	*E*
A	*A*	*B*	*E*	*D*	*F*	*C*	*R*	*U*	*O*	*K*	*N*	*I*	*V*	*J*	*L*	*H*	*T*	*X*	*Q*	*S*	*P*	*W*	*M*	*G*	*A*
B	*B*	*E*	*A*	*F*	*C*	*D*	*X*	*P*	*L*	*N*	*J*	*O*	*W*	*K*	*I*	*U*	*S*	*G*	*T*	*Q*	*H*	*M*	*V*	*R*	*B*
C	*C*	*F*	*D*	*E*	*B*	*A*	*K*	*O*	*U*	*R*	*G*	*P*	*T*	*X*	*H*	*L*	*V*	*J*	*W*	*M*	*I*	*Q*	*S*	*N*	*C*
D	*D*	*C*	*F*	*A*	*E*	*B*	*N*	*L*	*P*	*X*	*R*	*H*	*S*	*G*	*U*	*I*	*W*	*K*	*M*	*V*	*O*	*T*	*Q*	*J*	*D*
F	*F*	*D*	*C*	*B*	*A*	*E*	*J*	*I*	*H*	*G*	*X*	*U*	*Q*	*R*	*P*	*O*	*M*	*N*	*V*	*W*	*L*	*S*	*T*	*K*	*F*
G	*G*	*U*	*S*	*K*	*W*	*I*	*E*	*J*	*F*	*H*	*C*	*T*	*P*	*Q*	*R*	*M*	*N*	*O*	*B*	*L*	*A*	*X*	*D*	*V*	*G*
H	*H*	*T*	*X*	*V*	*L*	*J*	*I*	*E*	*G*	*F*	*S*	*D*	*R*	*P*	*Q*	*N*	*O*	*M*	*K*	*A*	*W*	*C*	*U*	*B*	*H*
I	*I*	*W*	*K*	*S*	*U*	*G*	*H*	*F*	*J*	*E*	*V*	*A*	*N*	*O*	*M*	*R*	*P*	*Q*	*X*	*D*	*T*	*B*	*L*	*C*	*I*
J	*J*	*L*	*V*	*X*	*T*	*H*	*F*	*G*	*E*	*I*	*B*	*W*	*O*	*M*	*N*	*Q*	*R*	*P*	*C*	*U*	*D*	*K*	*A*	*S*	*J*
K	*K*	*I*	*W*	*G*	*S*	*U*	*C*	*R*	*A*	*O*	*E*	*M*	*L*	*V*	*J*	*T*	*X*	*H*	*D*	*P*	*F*	*N*	*B*	*Q*	*K*
L	*L*	*V*	*J*	*T*	*H*	*X*	*P*	*D*	*N*	*B*	*M*	*E*	*K*	*I*	*W*	*G*	*U*	*S*	*R*	*C*	*Q*	*A*	*O*	*F*	*L*
M	*M*	*N*	*O*	*P*	*R*	*Q*	*T*	*S*	*V*	*W*	*L*	*K*	*E*	*A*	*B*	*C*	*F*	*D*	*H*	*G*	*X*	*I*	*J*	*U*	*M*
N	*N*	*O*	*M*	*R*	*Q*	*P*	*D*	*X*	*B*	*L*	*A*	*V*	*I*	*W*	*K*	*S*	*G*	*U*	*F*	*H*	*C*	*J*	*E*	*T*	*N*
O	*O*	*M*	*N*	*Q*	*P*	*R*	*U*	*C*	*K*	*A*	*W*	*B*	*J*	*L*	*V*	*X*	*H*	*T*	*G*	*F*	*S*	*E*	*I*	*D*	*O*
P	*P*	*Q*	*R*	*M*	*O*	*N*	*L*	*B*	*X*	*D*	*T*	*C*	*G*	*U*	*S*	*K*	*I*	*W*	*J*	*E*	*V*	*F*	*H*	*A*	*P*
Q	*Q*	*R*	*P*	*O*	*N*	*M*	*W*	*V*	*S*	*T*	*U*	*X*	*F*	*D*	*C*	*B*	*E*	*A*	*I*	*J*	*K*	*H*	*G*	*L*	*Q*
R	*R*	*P*	*Q*	*N*	*M*	*O*	*A*	*K*	*C*	*U*	*D*	*S*	*H*	*T*	*X*	*V*	*J*	*L*	*E*	*I*	*B*	*G*	*F*	*W*	*R*
S	*S*	*G*	*U*	*I*	*K*	*W*	*V*	*M*	*T*	*Q*	*H*	*R*	*D*	*C*	*F*	*A*	*B*	*E*	*L*	*N*	*J*	*P*	*X*	*O*	*S*
T	*T*	*X*	*H*	*L*	*J*	*V*	*M*	*W*	*Q*	*S*	*P*	*G*	*C*	*F*	*D*	*E*	*A*	*B*	*O*	*K*	*N*	*U*	*R*	*I*	*T*
U	*U*	*S*	*G*	*W*	*I*	*K*	*O*	*A*	*R*	*C*	*Q*	*F*	*X*	*H*	*T*	*J*	*L*	*V*	*N*	*B*	*M*	*D*	*P*	*E*	*U*
V	*V*	*J*	*L*	*H*	*X*	*T*	*S*	*Q*	*W*	*M*	*I*	*N*	*A*	*B*	*E*	*D*	*C*	*F*	*U*	*R*	*G*	*O*	*K*	*P*	*V*
W	*W*	*K*	*I*	*U*	*G*	*S*	*Q*	*T*	*M*	*V*	*O*	*J*	*B*	*E*	*A*	*F*	*D*	*O*	*P*	*X*	*R*	*L*	*N*	*H*	*W*
X	*X*	*H*	*T*	*J*	*V*	*L*	*B*	*N*	*D*	*P*	*F*	*Q*	*U*	*S*	*G*	*W*	*K*	*I*	*A*	*O*	*E*	*R*	*C*	*M*	*X*
	E	*A*	*B*	*C*	*D*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*	*X*	

theorem in permutation groups is that the "number of q" classes of the symmetric group S_n (of order $n!$) is equal to the number of distinct partitions in n. We thus have

$$
\begin{aligned}
&K_1 = (E) \quad 1+1+1+1\,(1^4), && g_1 = 1,\\
&K_2 = (C,\ D,\ F,\ G,\ H,\ Q) \quad 1+1+2+0 \quad (1^2 2)\\
&\qquad\qquad && g_2 = 6,\\
&K_3 = (K,\ L,\ M) \quad 2+2+0+0 \quad (2^2)\\
&\qquad\qquad && g_3 = 3, \qquad (7)\\
&K_4 = (A, B, I, J, N, O, V, W) \quad 1+3+0+0 \quad (31)\\
&\qquad\qquad && g_4 = 8,\\
&K_5 = (P, R, S, T, U, X) \quad 4+0+0+0 \quad (4^1)\\
&\qquad\qquad && g_5 = 6.
\end{aligned}
$$

Here we have denoted the classes by K_i, the number of elements in each class g_i. The partitions of the number 4 associated with the different classes are shown as (1^4), etc., which can be taken as an alternative description of classes. According to representation theory:

1. The number of inequivalent irreducible representations is equal to the number of classes. S_4 thus has five representations.
2. The sum of squares of the dimensions of the irreducible representations equals the order of the group:

$$n_1^2 + n_2^2 + n_3^2 + n_4^2 + n_5^2 = g = 24 \quad \text{for the } S_4 \text{ group.}$$

$D^{(2)}$: $\rightarrow$

$$D(A) = \begin{bmatrix} -i/2 & i/\sqrt{2} & -i/2 \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ i/2 & i/\sqrt{2} & i/2 \end{bmatrix}, \quad \chi = 0 \qquad D(F) = \begin{bmatrix} -1/2 & -1/\sqrt{2} & -1/2 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ -1/2 & 1/\sqrt{2} & -1/2 \end{bmatrix}, \quad \chi = -1 \qquad D(J) = \begin{bmatrix} -i/2 & 1/\sqrt{2} & i/2 \\ -i/\sqrt{2} & 0 & i/\sqrt{2} \\ -i/2 & -1/\sqrt{2} & i/2 \end{bmatrix}, \quad \chi = 0$$

$$D(B) = \begin{bmatrix} i/2 & 1/\sqrt{2} & -i/2 \\ -i/\sqrt{2} & 0 & -i/\sqrt{2} \\ i/2 & -1/\sqrt{2} & -i/2 \end{bmatrix}, \quad \chi = 0 \qquad D(G) = \begin{bmatrix} 0 & 0 & -i \\ 0 & -1 & 0 \\ i & 0 & 0 \end{bmatrix}, \quad \chi = -1 \qquad D(K) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad \chi = -1$$

$$D(C) = \begin{bmatrix} 0 & 0 & i \\ 0 & -1 & 0 \\ -i & 0 & 0 \end{bmatrix}, \quad \chi = -1 \qquad D(H) = \begin{bmatrix} -1/2 & i/\sqrt{2} & 1/2 \\ -i/\sqrt{2} & 0 & -i/\sqrt{2} \\ 1/2 & i/\sqrt{2} & -1/2 \end{bmatrix}, \quad \chi = -1 \qquad D(L) = \begin{bmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \quad \chi = -1$$

$$D(D) = \begin{bmatrix} -1/2 & -i/\sqrt{2} & 1/2 \\ i/\sqrt{2} & 0 & i/\sqrt{2} \\ 1/2 & -i/\sqrt{2} & -1/2 \end{bmatrix}, \quad \chi = -1 \qquad D(I) = \begin{bmatrix} i/2 & -i/\sqrt{2} & i/2 \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ -i/2 & -i/\sqrt{2} & -i/2 \end{bmatrix}, \quad \chi = 0 \qquad D(M) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \chi = -1$$

$$D(N) = \begin{bmatrix} i/2 & i/\sqrt{2} & i/2 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ -i/2 & i/\sqrt{2} & -i/2 \end{bmatrix}, \quad \chi = 0 \qquad D(R) = \begin{bmatrix} 1/2 & -i/\sqrt{2} & -1/2 \\ -i/\sqrt{2} & 0 & -i/\sqrt{2} \\ -1/2 & -i/\sqrt{2} & 1/2 \end{bmatrix}, \quad \chi = 1 \qquad D(V) = \begin{bmatrix} -i/2 & -i/\sqrt{2} & -i/2 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ i/2 & -i/\sqrt{2} & i/2 \end{bmatrix}, \quad \chi = 0$$

$$D(O) = \begin{bmatrix} i/2 & -1/\sqrt{2} & -i/2 \\ i/\sqrt{2} & 0 & i/\sqrt{2} \\ i/2 & 1/\sqrt{2} & -i/2 \end{bmatrix}, \quad \chi = 0 \qquad D(S) = \begin{bmatrix} 1/2 & i/\sqrt{2} & -1/2 \\ 1/\sqrt{2} & 0 & i/\sqrt{2} \\ -1/2 & i/\sqrt{2} & 1/2 \end{bmatrix}, \quad \chi = 1 \qquad D(W) = \begin{bmatrix} -i/2 & -1/\sqrt{2} & i/2 \\ -i/\sqrt{2} & 0 & -i/\sqrt{2} \\ -i/2 & 1/\sqrt{2} & i/2 \end{bmatrix}, \quad \chi = 0$$

$$D(P) = \begin{bmatrix} -i & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & i \end{bmatrix}, \quad \chi = 1 \qquad D(T) = \begin{bmatrix} i & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -i \end{bmatrix}, \quad \chi = 1 \qquad D(X) = \begin{bmatrix} 1/2 & -1/\sqrt{2} & 1/2 \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/2 & 1/\sqrt{2} & 1/2 \end{bmatrix}, \quad \chi = 1$$

$$D(Q) = \begin{bmatrix} -1/2 & +1/\sqrt{2} & -1/2 \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ -1/2 & -1/\sqrt{2} & -1/2 \end{bmatrix}, \quad \chi = -1 \qquad D(U) = \begin{bmatrix} 1/2 & 1/\sqrt{2} & 1/2 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1/2 & -1/\sqrt{2} & 1/2 \end{bmatrix}, \quad \chi = 1 \qquad \mathscr{D}(E) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad \chi = 3$$

This equation has the solution $1^2 + 1^2 + 2^2 + 3^2 + 3^2$.

Thus S_4 has two one-dimensional, two three-dimensional, and one two-dimensional representations. All finite groups have a trivial one-dimensional representation where each element is represented by 1. One set of the three-dimensional representation matrices of S_4 is given on the previous page. Swamy and Samuel give all the representation matrices in their work. It is important to note that, while the inequivalent representations will be finite in number, there will be an infinite number of equivalent representations, each differing from the other only by a similarity transformation. We use the notation $D^{(j)}$ for the jth irreducible representation, $D^{(j)}(R)$ for the matrix representing the general group element R in the Jth representation, and $D^{(j)}_{mm'}(R)$ for the mm'th matrix element of that particular matrix, $\chi^{(j)}(R)$ is the character of the element R in the Jth irreducible representation.

All the elements of one class have the same character in a given irreducible representation. Table III gives the classwise character table for S_4. Two important theorems in representation theory are summarized in the formulas

$$\sum_R D^{(\mu)}_{il}(R) D^{(\nu)^*}_{jm}(R) = (g/n_\mu)\,\delta_{\mu\nu}\,\delta_{ij}\,\delta_{lm},$$
$$\sum_R \chi^{\mu}(R)\chi^{(\nu)^*}(R) = g\,\delta_{\mu\nu}, \tag{8}$$

or, equivalently,

$$\sum_i^k \chi_i^{(\mu)} \chi_i^{(\nu)^*} g_i = g\,\delta_{\mu\nu}.$$

Here μ and ν label the different irreducible representations, R as a general group element, n_μ the dimensionality of the μth representation, g_i the number of elements in the ith class, and g the total number of elements or the order of the group. In other words, the above theorems go to show that, treated as vectors in the space of group elements, these are orthogonal. Similarly, characters $\chi^{(j)}(R)$ of the irreducible representations form an orthogonal vector system in the same space of group elements. It is easy to verify the above numerically with the help of the representation matrices for S_4 and its character table (Table III). The third relation above shows that normalized characters $\sqrt{g_i/g}\,\chi_i^{(j)}$ form an orthonormal vector system in the k-dimensional space of classes. Various prescriptions exist for calculating the characters of a finite group, and it is much easier to determine characters in many instances than the representation matrices themselves. Frobenius developed a systematic algebraic procedure for determining the characters of any permutation group. If we know the characters of a permutation group we can infer therefrom the characters of all finite groups because of a theorem in group theory which says that any finite group is isomorphic to a subgroup of some permutation or symmetric group. Isomorphic groups have the same character table. One of the interesting prescriptions for obtaining characters is the class product relation

$$k_i k_j = \sum_t C_{ijl} k_l, \tag{9}$$

where the k_i are classes and the summation is done over all the classes. From this can be derived

$$g_i g_j \chi_i^{(\mu)} \chi_j^{(\mu)} = n_\mu \sum_l C_{ijl} g_l \chi_l^{(\mu)}, \tag{10}$$

where the C coefficients are identical to the ones in Eq. (9). To illustrate the first identity let us take classes k_2 and k_3 of S_4:

$$\begin{aligned} k_3 k_2 &= (K,\ L,\ M)(C,\ D,\ F,\ G,\ H,\ Q) \\ &= (G,\ S,\ U,\ C,\ R,\ X), (T,\ H,\ X,\ P,\ D,\ U), \\ &\quad (P,\ R,\ Q,\ T,\ S,\ F) \\ &= k_2 + 2k_5. \end{aligned}$$

This gives for the values of the C coefficients

$$\begin{aligned} &C_{321} = 0, \quad C_{322} = 1, \quad C_{323} = 0, \\ &C_{324} = 0, \quad C_{325} = 2. \end{aligned} \tag{11}$$

With these numbers determined it is a trivial exercise to check (10) from the character table for S_4. Frobenius also gave a prescription for determining the characters of a group from the known characters of its subgroups. This is discussed lucidly in the work of Littlewood.

TABLE III Character Table of S_4

	K_1 $g_1 = 1$	K_2 $g_2 = 6$	K_3 $g_3 = 3$	K_4 $g_4 = 8$	K_5 $g_5 = 6$
$D^{(1)}$	1	−1	1	1	−1
$D^{(2)}$	3	−1	−1	0	1
$D^{(3)}$	2	0	2	−1	0
$D^{(4)}$	3	1	−1	0	−1
$D^{(5)}$	1	1	1	1	1

B. Construction of Representations

The construction of the representation of a finite group depends on the choice of a suitable set of basis functions, and the dimensionality of the representation is the number of basis functions chosen. If $\psi_1, \psi_2, \ldots, \psi_n$ are a set of n

TABLE IV $O_R x, O_R y, O_R z$

	EX_i	$F(132)$ C_3X_i	$D(123)$ C_3X_i	$A(23)$ $\sigma_v X_i$	$C(12)$ $\sigma_{v'} X_i$	$B(13)$ $\sigma_{v''} X_i$
$X_1 = X$	X	$-(1/2)x - (\sqrt{3}/2)y$	$-(1/2)x + (\sqrt{3}/2)y$	$-X$	$(1/2)x - (\sqrt{3}/2)y$	$(1/2)x + (\sqrt{3}/2)y$
$X_2 = Y$	Y	$(\sqrt{3}/2)x - (1/2)y$	$-(\sqrt{3}/2)x - (1/2)y$	Y	$-(\sqrt{3}/2)x - (1/2)y$	$(\sqrt{3}/2)x - (1/2)y$
$X_3 = Z$	Z	Z	Z	Z	Z	Z

basis functions, the fundamental formula of representation theory then is expressed as

$$\mathbf{O}_R \psi_\nu = \sum_{\mu=1}^{n} D_{\mu\nu}(R)\psi_\mu. \tag{12}$$

This means that the result of a group operation $\mathbf{O}_R$ (element of the group) on one of the basis functions is a linear combination of the basis functions and the matrix of the coefficients in the linear combination essentially determines the representation. We will illustrate this for the C_{3v} group. In Fig. 1 let us choose a rectangular coordinate system with origin at the centroid O and the Y axis through the median AA' with the X axis perpendicular to it in its plane. The Z axis will be pointing up from the origin perpendicular to the plane of the figure. The vertices will then have the coordinates $A = (0,\ 1,\ 0)$, $B(\sqrt{3}/2,\ -1/2,\ 0)$, and $C(\sqrt{3}/2,\ -1/2,\ 0)$. In Table IV we gather the results of making all the six operations of the group C_{3v}, on x, y, and z. From formula (12) and the table we see that if we choose z as one basis function, we have a one-dimensional representation with the number 1 representing all the elements. In the character table (Table V) this is shown as the A_1 representation, which means that the basis function is symmetric with respect to rotation about the cyclic axis. If we choose a pair of functions $x = \psi_1$, $y = \psi_2$ in that order we obtain the two-dimensional (E representation) irreducible representation

$$\begin{aligned}
&\mathfrak{D}(R) && \mathfrak{D}(A)\\
&\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, && \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix},\\
&\chi = 2 && \chi = 0\\
&\mathfrak{D}(B) && \mathfrak{D}(C)\\
&\begin{bmatrix} 1/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{bmatrix}, && \begin{bmatrix} 1/2 & -\sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{bmatrix},\\
&\chi = 0 &&\\
&\mathfrak{D}(D) && \mathfrak{D}(F)\\
&\begin{bmatrix} -1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{bmatrix}, && \begin{bmatrix} -1/2 & \sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{bmatrix}.\\
&\chi = -1 && \chi = -1
\end{aligned} \tag{13}$$

The characters are shown underneath each matrix, and Table V gives all the characters classwise. x, y are written against the E representation since these transform accordingly. x is said to belong to the first row of the representation $D^{(E)}$ and y to its second row. Since the A_2 representation is also one-dimensional, the matrix elements (numbers which are also the characters) are easily obtained by applying the orthogonality relations either to the matrix elements of different irreducible representations or to the characters. It is to be noted that the choice of x and y for the two-dimensional representation is fortuitous inasmuch as it straightaway generated the representation. However, there is a general prescription for obtaining the basis functions starting with any arbitrary function by means of projection operators. The formula for this projection operator is

$$p_i^{(\mu)} = \frac{n_\mu}{g} \sum_R D_{ii}^{(\mu)*}(R)\mathbf{O}_R, \tag{14}$$

$$p_i^2 = p_i, \qquad p_i p_j = 0, \tag{14a}$$

and for the two-dimensional representation these are explicitly

$$\begin{aligned}
p_1 &= \tfrac{1}{3}\{E - \tfrac{1}{2}D - \tfrac{1}{2}F - A + \tfrac{1}{2}B + \tfrac{1}{2}C\},\\
p_2 &= \tfrac{1}{3}\{E - \tfrac{1}{2}D - \tfrac{1}{2}F + A - \tfrac{1}{2}B - \tfrac{1}{2}C\}.
\end{aligned} \tag{15}$$

If we choose the arbitrary function $\phi(x, y, z) = x + y + z$, then it is easily seen that

$$\begin{aligned} p_1\Phi &= x\\ p_2\Phi &= y \end{aligned} \qquad \text{or} \qquad \psi_i^{(\nu)} = p_i^{(\nu)}\Phi. \tag{16}$$

We will conclude this discussion of representation theory by introducing the concept of direct product of two groups. Let a group G_1 be of order g_1 with elements A_1, $A_2, \ldots, A_{g_i}$ and another group of G_2 of order g_2 with elements $B_1, B_2, \ldots, B_{g_2}$. A direct product group $G_1 \times G_2$

TABLE V Character Table of S_3 or C_{3v}

	$K_1(1^3)E$ $g_1 = 1$	$K_2 21(A, B, C)$ $(g_2 = 3)$	$K_3(3)(D, F)$ $g_3 = 2$	
$D^{(A_1)}$	1	1	1	z
$D^{(A_2)}$	1	-1	1	
$D^{(E)}$	2	0	-1	(x, y)

is defined if the elements of G_1 commute with the elements of G_2. This group consists of the ordered pair of elements A_1B_1, $A_1B_2, \ldots$, and will be of the order g_1g_2. The multiplication rule for the elements of the direct product follows simply from the multiplication rules for each group. Thus if in group G_1, $A_1A_2 = A_5$ and in group G_2, $B_1B_2 = B_7$, then in the direct product group $(A_1B_1) \times (A_2B_2)$ will be equal to $(A_1A_2) \times (B_1B_2)$ the element corresponding to A_5B_7. For example, C_{3h}, is a direct product group of the cyclic group C_3 (with elements E, C_3, and C_3^2) and the group C_h consisting of the identity element and reflection in a horizontal plane. The rotation–reflection group in Lie groups is a direct product of the rotation group and the inversion group (parity operation). The direct products of some Lie groups decompose into a direct sum of irreducible components. A well-known theorem, called the Clebsch–Gordan theorem, prescribes how to construct basis functions for their irreducible representations (IRs) from the basis functions of the IRs of the factors making up the direct product. The characters of the direct product groups can be calculated easily from the characters of the two "factors" according to the formula

$$\chi^{(\nu)}(A_1)\chi^{\nu}(B_2) = \chi^{(\mu \times \nu)}(A_1B_2). \tag{17}$$

III. CONTINUOUS LIE GROUPS

The theory of continuous or topological groups is due to the Norwegian mathematician Sophus Lie and hence these are familiarly known as Lie groups. In these groups the group element is a continuous invertible transformation of a set of variables and the transformations satisfy the group postulates. For instance, the fundamental property of a group being closed with respect to multiplication here means that the transformation of a transformation is another transformation, and the identity element means a mapping of the variables into themselves. Translations and rotations of a coordinate system in three- or four-dimensional space are examples of these transformations which form a group.

A continuous transformation of a set of variables is best described in terms of parameters which can take on continuous values in a finite (compact) or infinite (noncompact) range, a certain equilibrium value of the parameters corresponding to the identity transformation. If $x_1, x_2, x_3, \ldots, x_n$ are the variables and $a_1, a_2, \ldots, a_k$ are the parameters, we can write the transformation as

$$\begin{aligned} x_i' &= f_i(x_i, a_r), \qquad i = 1, \ldots, n, \\ &\qquad\qquad\quad r = 1, \ldots, k \\ x_i' &= x_i, \qquad\quad \text{(identity)} \\ &\qquad\qquad\quad \text{for} \quad a_1 = a_1^0, \quad \text{etc.}, \end{aligned} \tag{18}$$

and further, for any arbitrary functions of the variables we have

$$\phi(x_i') = \phi(f_i). \tag{19}$$

While the precise nature of the finite transformations defines the Lie group, Lie proved that these transformations can be built up from the identity (since they are continuous) by means of infinitesimal generators X_k, there being in general as many generators as essential parameters. According to Lie,

$$\mathbf{X}_\alpha = \sum_{i=1}^{n} \left.\frac{\partial x_\iota'}{\partial a_\alpha}\right|_{a_\alpha 0} \frac{\partial}{\partial x_i}, \tag{20}$$

$$X_i' = \{e^{\varepsilon_1\mathbf{X}_1 + \varepsilon_2\mathbf{X}_2 + \cdots + \varepsilon_n\mathbf{X}_n}\}X_i, \tag{21}$$

and

$$\phi(x_i') = (e^{\Sigma_i \varepsilon_i \mathbf{X}_i})\phi(x_i). \tag{22}$$

The interesting thing about the Lie group is that it is uniquely defined in terms of certain structure relations that exist among these derived infinitesimal generators. These are

$$[\mathbf{X}_\chi, \mathbf{X}_\mu] = \sum_\lambda C_{\chi\mu}^\lambda \mathbf{X}_\lambda, \qquad [\mathbf{A}, \mathbf{B}] \equiv \mathbf{AB} - \mathbf{BA}, \tag{23}$$

where $C_{\chi\mu}^\lambda$ are called structure constants. Knowing the structure constants is synonymous with knowing the group of finite transformations f_i. The above relations are usually referred to as the Lie algebra of the generators, and often a Lie group is defined in terms of its Lie algebra. Sometimes a subset of the operators X_α itself satisfies a relation of the above type, and in this case we say these smaller number of generators constitutes a subgroup of the Lie group. In quantum mechanics and in particle physics there is a lot of interest in a certain category of Lie groups known as semisimple groups. The French mathematician E. Cartan, to whom we owe several developments of the original Lie theory, established the criterion for a Lie group to be semisimple. If we define

$$g_{\mu\rho} = \sum_{\alpha\beta} C_{\mu\alpha}^\beta C_{\nu\beta}^\alpha, \tag{24}$$

then $\|g_{\mu\rho}\|$ is the determinant of the matrix whose ordered coefficients are $g_{\mu\nu}$, and nonvanishing of this determinant is the condition for the group to be semisimple. Casimir has shown that in the case of such groups there exists a certain function of the infinitesimal generators, called the Casimir operator, that commutes with every generator

$$\begin{aligned} [\mathbf{C}, \mathbf{X}_\alpha] &= 0 \qquad \text{for all} \quad \alpha, \\ \mathbf{C} &= \sum_{\mu\nu} g^{\mu\nu}\mathbf{X}_\mu\mathbf{X}_\mu, \end{aligned} \tag{25}$$

where $g^{\mu\nu}$ is the matrix inverses to $g_{\mu\nu}$. Such an operator which commutes with all the generators is called an invariant of the Lie group, and this plays an important

role in understanding the structure of the group and also in constructing its irreducible representations.

A. Three-Dimensional Rotation Group

As an application of the preceding ideas we will discuss the three-dimensional rotation group SO(3) which describes one of the fascinating symmetries in physics. If a sphere is rotated in space about its radius it is brought into coincidence with itself and this is known as spherical symmetry. Spherical symmetry (rotational invariance) has experimentally interesting consequences in classical as well as quantum physics. Rotational invariance implies in general conservation of angular momentum. In classical physics, for instance, the latter conservation principle explains why a figure skater draws her outstretched arms closer to her body to increase her speed of gyration, or why a cat always lands on its feet when it falls! Three-dimensional rotations form a Lie group called the SO(3) group.

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos t_2 \cos t_3 & -\cos t_2 \sin t_3 & \sin t_2 \\ \cos t_1 \sin t_3 + \sin t_1 \sin t_2 \cos t_3 & \cos t_1 \cos t_3 - \sin t_1 \sin t_2 \sin t_3 & \sin t_1 \cos t_2 \\ \sin t_1 \sin t_3 - \cos t_1 \sin t_2 \cos t_3 & \sin t_1 \cos t_3 + \cos t_1 \sin t_2 \sin t_3 & \cos t_1 \cos t_2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{26}$$

$$x'^2 + y'^2 + z'^2 = x^2 + y^2 + z^2.$$

The way to describe rotations of a coordinate system mathematically in three-dimensional space was shown by Euler, centuries ago, who introduced three real parameters called the Eulerian angles. An alternative and equivalent description is given in Eq. (26). Here t_i are the parameters of the Lie group and these are naturally functions of the Eulerian angles. Following the Lie prescription, the infinitesimal generators are seen to be

$$\begin{aligned} \mathbf{X}_{t_1} \equiv \mathbf{X}_1 &= z\frac{\partial}{\partial y} - y\frac{\partial}{\partial z}, \\ \mathbf{X}_2 &= x\frac{\partial}{\partial z} - z\frac{\partial}{\partial x}, \\ \mathbf{X}_3 &= y\frac{\partial}{\partial x} - x\frac{\partial}{\partial y}, \end{aligned} \tag{27}$$

and the associated Lie algebra is

$$[\mathbf{X}_1,\ \mathbf{X}_2] = \mathbf{X}_3 \qquad \text{and cyclically.} \tag{27a}$$

This relation is written somewhat cryptically as

$$[\mathbf{X}_i,\ \mathbf{X}_j] = \sum_k \varepsilon_{ijk}\mathbf{X}_k, \tag{27b}$$

where the Levi–Civita symbol ε_{ijk} assumes the value 0 whenever any two indices are equal, equals $+1$ when ijk is an even permutation of 123, and equals -1 for an odd permutation. The finite transformation (26) can be obtained as

$$x_i' = \left(e^{\sum_i \varepsilon_i \mathbf{X}_i}\right) X_i = \left(e^{t_3 X_3} e^{t_2 X_2} e^{t_1 X_1}\right) X_i. \tag{28}$$

This is a semisimple group with $g_{\mu\nu}$ being a diagonal unit matrix and the Casimir operator, the invariant of the group, is simply the sum of the squares of the generators,

$$\mathbf{C} = \mathbf{X}_1^2 + \mathbf{X}_2^2 + \mathbf{X}_3^3. \tag{29}$$

In quantum mechanics this operator is related to the square of the angular momentum, which is one of the experimentally observable quantities.

The spherical harmonics Y_l^m defined on the surface of the unit sphere and originally discovered in the context of solving the Laplace equation are eigenfunctions of this Casimir operator, that is, they satisfy the equation

$$\mathbf{C}Y_l^m(\theta,\ \phi) = -l(l+1)Y_l^m(\theta,\ \phi). \tag{30}$$

Here l is any positive integer and for a given l m can assume any integral value between $-l$ and $+l$, a total of $2l+1$ values. Thus for a given l there are $2l+1$ spherical harmonics and these are the basic functions which generate the odd $(2l+1)$-dimensional irreducible representations of the three-dimensional rotation group. If a dynamical system in quantum mechanics has a certain symmetry, or more precisely if the Hamiltonian of the dynamical system is invariant to a certain set of group operations, then its solutions generate the representations of that group. In particular, if the dynamical system has spherical symmetry, which means that the Hamiltonian has rotational invariance, the spherical harmonics must be a factor in its solution. The Laplace operator is indeed rotationally invariant, and its solutions are either

$$r^l Y_l^m \qquad \text{or} \qquad r^{-l-l} Y_l^m. \tag{31}$$

B. SU(2) Covering Group of the Rotation Group

The even-dimensional representations of SO(3) were first discovered by Weyl. He showed that the covering group of this rotation group is SU(2), which is the unitary unimodular group in two dimensions. This group played a crucial role in characterizing the spin properties of particles such as electrons and protons (fermions), and recently it has proved to be one of the cornerstones of a theory that led to the discovery of the $W^{\pm}$ and Z^0 bosons.

A general element of the SU(2) group can be written as a 2×2 unitary matrix of determinant $+1$.

$$U = \begin{bmatrix} \alpha & \beta \\ -\beta^* & \alpha^* \end{bmatrix}, \qquad \alpha\alpha^* + \beta\beta^* = 1. \tag{32}$$

Here the asterisk stands for the complex conjugate. This is called a covering group of SO(3) because it not only induces three-dimensional rotations, it gives the odd- as well as even-dimensional representations of the latter. For instance, a similarity transformation with U of the matrix

$$\vec{\sigma} \cdot \vec{\mathbf{r}} = \begin{bmatrix} z & x - iy \\ x + iy & -z \end{bmatrix} \tag{33}$$

gives

$$\vec{\sigma} \cdot \vec{\mathbf{r}}' = \mathbf{U}^+ \vec{\sigma} \cdot \vec{\mathrm{r}}\, \mathbf{U}. \tag{34}$$

$r'(x',\ y',\ z')$ is related to $r(x,\ y,\ z)$ through α, β and their complex conjugates. From the theory of determinants we know that $\vec{\sigma} \cdot \vec{\mathrm{r}}'$ and $\vec{\sigma} \cdot \vec{\mathrm{r}}$ have the same determinant or

$$x^2 + y^2 + z^2 = x'^2 + y'^2 + z'^2.$$

Wigner has given the explicit connection between the complex α, β and of SU(2) and the Eulerian angles α, β, γ which characterize the three-dimensional rotations. It is straightforward algebra to establish the relation between the SU(2) elements and the t_i of (26). The Lie algebra satisfied by the generators of SU(2) is isomorphic to that of the SO(3) group. The Pauli spin operators that describe the magnetic electron in atomic physics, the Cayley–Klein operators that describe rigid body spin in classical mechanics, and quoternions in vector space theory all have algebras similar to the SU(2) algebra. The basis functions which generate the irreducible representations of SU(2) are known as monomials or tensors. A typical even-dimensional representation of SU(2) is given in Eq. (35). Some of the properties of SU(2) groups will be discussed elsewhere in this article in the context of applications in particle physics.

$$D^{1/2}_{mm'}(\alpha\beta\gamma) = \begin{bmatrix} e^{-i(\alpha/2)}(\cos\beta/2)e^{-i(\gamma/2)} & -i^{-i(\alpha/2)}(\sin\beta/2)e^{i(\gamma/2)} \\ e^{i(\alpha/2)}(\sin\beta/2)e^{-i(\gamma/2)} & e^{i(\alpha/2)}(\cos\beta/2)e^{i(\gamma/2)} \end{bmatrix}. \tag{35}$$

C. Four-Dimensional Rotation Group and Homogeneous Lorentz Group

The extension of the rotation group to four dimensions is straightforward but not trivial. The SO(4) group elements are the linear transformations that ensure

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 = x_1'^2 + x_2'^2 + x_3'^2 + x_4'^2, \tag{36}$$

where each element of the group transforms x_i into x_i':

$$x'_\mu = \sum_\lambda b_{\mu\lambda} x_\lambda. \tag{37}$$

The transformation matrix $b_{\mu\lambda}$, an element of the SO(4) group, can be expressed in terms of six real parameters $\alpha, \beta, \gamma, \alpha', \beta', \gamma'$. Forsyth has shown that the matrix elements $b_{\mu\lambda}$ are somewhat complicated trigonometric functions of these parameters. This matrix, as well as the element of SO(3), is an orthogonal matrix. The Lie algebra of the six infinitesimal generators is best given in terms of two sets of operators $X_1 = L_1$, $X_2 = L_2$, $X_3 = L_3$, and $X_4 = A_1$, $X_5 = A_2$, $X_6 = A_3$:

$$\begin{aligned} [\mathbf{L}_i, \mathbf{L}_j] &= \sum_k \varepsilon_{ijk}\, \mathbf{L}_k, \\ [\mathbf{A}_i, \mathbf{A}_j] &= \sum_k \varepsilon_{ijk}\, \mathbf{L}_k, \\ [\mathbf{L}_i, \mathbf{A}_j] &= \sum_k \varepsilon_{ijk}\, \mathbf{A}_k. \end{aligned} \tag{38}$$

These relations identify the SO(4) group and any six operators, which obey the same commutation relations as the X_i generate a group isomorphic to SO(4). It is interesting to note that the L_j themselves form a closed Lie algebra isomorphic to SO(3). This group is then a subgroup of SO(4). This group has two invariant operators,

$$\begin{aligned} \mathbf{F} &= \sum \mathbf{L}_i^2 + \mathbf{A}_i^2, \\ G &= \mathbf{L}_1\mathbf{A}_1 + \mathbf{L}_2\mathbf{A}_2 + \mathbf{L}_3\mathbf{A}_3. \end{aligned} \tag{39}$$

This group plays an important role in understanding the bound states of the nonrelativistic hydrogen atom as well as in particle physics.

If instead of Eq. (36) we have transformations which leave

$$X_1^2 + X_2^2 + X_3^2 - X_4^2 \tag{40}$$

invariant, then these form the elements of the homogeneous Lorentz group, of fundamental importance in the special theory of relativity and elementary particle physics. We have explicitly

$$X'_\mu = \sum_\lambda a_{\mu\lambda} x_\lambda. \tag{41}$$

The $a_{\mu\lambda}$ are once again expressed in terms of six real parameters. How algebraically complicated these functions, elements of $a_{\mu\lambda}$, are can be judged from the typical element

$$\begin{aligned} a_{12} = \sin\gamma\cos\gamma\{&\cosh\alpha'\cos^2\beta \\ &- \cos\alpha\cosh^2\beta' + \cos\alpha\sin^2\beta \\ &+ \cosh\alpha'\sinh^2\beta'\} - \sin a\sin\beta\cosh\beta' \\ &+ \sinh\alpha'\sinh\beta'\cos\beta. \end{aligned} \tag{42}$$

For instance, when $\alpha = 1$, $\beta = 0.7$, $\gamma = 0.35$, $\alpha' = 0.5$, $\beta' = 1.4$, and $\gamma' = 0.1$, the $a_{\mu\lambda}$ becomes a numerical matrix, the element of the homogenous Lorentz group, as given in Eq. (43). The

$$a_{\mu\lambda} = \begin{bmatrix} 0.5930545 & 0.3897004 & -0.9044001 & 0.5670270 \\ 1.2037232 & -1.2987247 & 1.5783052 & -2.1509726 \\ 0.5232450 & -2.2304747 & 0.7592345 & -2.1966429 \\ -1.0365561 & 2.4111345 & -1.6986536 & 3.2822924 \end{bmatrix} \tag{43}$$

infinitesimal generators of this group and the associated Lie algebra are

$$\begin{aligned} \mathbf{X}_1 &= y\frac{\partial}{\partial z} - z\frac{\partial}{\partial y} \equiv \mathbf{L}_1, \\ \mathbf{X}_2 &= z\frac{\partial}{\partial x} - x\frac{\partial}{\partial z} \equiv \mathbf{L}_2, \\ \mathbf{X}_3 &= x\frac{\partial}{\partial y} - y\frac{\partial}{\partial z} \equiv \mathbf{L}_3, \\ \mathbf{X}_4 &= t\frac{\partial}{\partial x} + x\frac{\partial}{\partial t} \equiv \mathbf{A}_1, \\ \mathbf{X}_5 &= t\frac{\partial}{\partial y} + y\frac{\partial}{\partial t} \equiv \mathbf{A}_2, \\ \mathbf{X}_6 &= t\frac{\partial}{\partial z} + z\frac{\partial}{\partial t} \equiv \mathbf{A}_3, \\ &(t \equiv x_4); \\ [\mathbf{L}_i, \mathbf{L}_j] &= -\sum_k \varepsilon_{ijk}\mathbf{L}_k, \\ [\mathbf{A}_i, \mathbf{A}_j] &= \sum_k \varepsilon_{ijk}\mathbf{L}_k, \\ [\mathbf{L}_i, \mathbf{A}_j] &= -\sum_k \varepsilon_{ijk}\mathbf{A}_k, \end{aligned} \tag{44}$$

When three of the parameters are set equal to zero we have the "Lorentz rotations" or the Lorentz transformation, where one frame of reference moves with respect to another frame with constant velocity $\vec{\mathrm{v}}$ in an arbitrary direction. In this case the general element reduces to Eq. (45a) below.

$$a_{\mu\lambda} = \begin{bmatrix} \cosh\alpha' & \sinh\alpha'\sinh\beta' & \sinh\alpha'\cosh\beta'\sinh\gamma' & \sinh\alpha'\cosh\beta'\cosh\gamma' \\ -\sinh\alpha'\sinh\beta' & \begin{Bmatrix}\cosh\alpha' \\ +(1-\cosh\alpha')\cosh^2\beta'\end{Bmatrix} & \begin{Bmatrix}(1-\cosh\alpha')\sinh\beta' \\ \cdot\cosh\beta'\sinh\gamma'\end{Bmatrix} & \begin{Bmatrix}(1-\cosh\alpha') \\ \cdot\sinh\beta'\cosh\beta'\cosh\gamma'\end{Bmatrix} \\ -\sinh\alpha'\cosh\beta'\sinh\gamma' & \begin{Bmatrix}(1-\cosh\alpha') \\ \cdot\sinh\beta'\cosh\beta'\sinh\gamma'\end{Bmatrix} & \begin{Bmatrix}(1-\cosh\alpha')\cosh^2\beta' \\ \cdot\sinh^2\gamma'+1\end{Bmatrix} & \begin{Bmatrix}(1-\cosh\alpha') \\ \cdot\cosh^2\beta'\sinh\gamma'\cosh\gamma'\end{Bmatrix} \\ \sinh\alpha'\cosh\beta'\cosh\gamma' & \begin{Bmatrix}(\cosh\alpha'-1) \\ \cdot\sinh\beta'\cosh\beta'\cosh\gamma'\end{Bmatrix} & \begin{Bmatrix}(\cosh\alpha'-1) \\ \cdot\cosh^2\beta'\sinh\gamma'\cosh\gamma'\end{Bmatrix} & \begin{Bmatrix}(\cosh\alpha'-1) \\ \cdot\cosh^2\beta'\cosh^2\gamma'+1\end{Bmatrix} \end{bmatrix}. \tag{45a}$$

In (45a),

$$\begin{bmatrix} a\cdot \\ \cdot b \end{bmatrix} \equiv ab. \tag{45b}$$

The above matrix elements are related to the velocity components of the moving frame (with the speed of light $c = 1$) as

$$\begin{aligned} v_x &= \tanh\alpha'\cosh\beta'\cosh\gamma', \\ v_y &= i\tanh\alpha'\cosh\beta'\cosh\gamma', \\ v_z &= i\tanh\alpha'\sinh\beta', \\ v &= \sqrt{v_x^2 + v_y^2 + v_z^2} = \tanh\alpha'. \end{aligned} \tag{46}$$

This group also has two invariant operators, somewhat complicated functions of the generators. The representations of the homogeneous Lorentz group can be obtained from its covering group which was shown by Bargmann to be SL(2, C), an element of which is a 2×2 complex matrix with determinant $+1$. For example, the element of the covering group corresponding to the element of the homogeneous Lorentz group in Eq. (43) is

$$\begin{bmatrix} -1.222144 + i(0.2411005) & -0.6624081 - i(1.0352785) \\ 0.1311536 - i(0.9976158) & 1.9325431 - i(0.4832207) \end{bmatrix}. \tag{47}$$

It is important to mention that the full group of relativistic quantum mechanics is the inhomogeneous Lorentz group or the Poincaré group which has, in addition, elements corresponding to translations in space–time.

IV. APPLICATIONS IN ATOMIC PHYSICS

A. Symmetry of the Hydrogen Atom

The nonrelativistic quantum mechanical Hamiltonian describing the motion of an electron in the hydrogen atom is

$$H = \frac{-1}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) - \frac{e^2}{r},$$
$$r = \sqrt{x^2 + y^2 + z^2}, \tag{48}$$

where e and m are the charge and mass of the electron and r the distance from the nucleus. The electron is in a central field, the potential energy e^2/r being invariant to rotations in three-dimensional space. This Hamiltonian thus has SO(3) symmetry and its bound-state solutions, which generate the irreducible representations of the three-dimensional rotation groups, are expressed in terms of three quantum numbers n, l, m:

$$U_{nlm}(\vec{\mathrm{r}}) = Y_l^m(\theta,\ \phi)R_{nl}(r). \tag{49}$$

As can be expected from our discussion of the rotation group, the spherical harmonics Y_l^m are a factor in the solution. $D^{(l)}_{mm'}$ are the matrices of the irreducible representation of SO(3) of dimension $2l+1$. U_{nlm} is the general form of a central field wave function. L^2 and L_z are the angular momentum operators which commute with the Hamiltonian and U_{nlm} is an eigenfunction of both with eigenvalues $l(l+1)$ and m in units of $\hbar = l$. The L^2 is just the Casimir operator of the SO(3) group apart from a constant and L_z is essentially one of the infinitesimal generators of the SO(3) group.

Schrödinger, who solved the quantum mechanical equation for the hydrogen atom, pointed out that the energy (eigenvalue of H) depends on only one quantum number n, the so-called principal quantum number, and is independent of both l and m. Since there are n^2 number of different states (combinations of l and m for a given n), all these have the same energy and this is known as degeneracy. The independence of the energy on the orbital angular momentum quantum number l is known as accidental degeneracy and was difficult to understand for quite some time. As a result of the work of Pauli, Fock, and Bargmann we now know that this degeneracy exists because the hydrogen atom has a higher symmetry than SO(3). The Hamiltonian is invariant to rotations in four-dimensional space, or the group of this Hamiltonian is SO(4), the four-dimensional rotation group. Fock showed that the Schrödinger equation in momentum space can be written as a four-dimensional Laplacian equation,

$$\left(\frac{\partial^2}{\partial p_1^2} + \frac{\partial^2}{\partial p_2^2} + \frac{\partial^2}{\partial p_3^2} + \frac{\partial^2}{\partial p_4^2}\right)\Phi = 0, \tag{50}$$

and that its solutions Φ can be written in terms of spherical harmonics defined on the three-dimensional surface of a sphere in four-dimensional space. We recall that there are six generators of SO(4) which satisfy the Lie algebra appropriate to this group. Pauli demonstrated that there exists another vector operator, besides the usual angular momentum operator $\vec{\mathrm{L}}$, which commutes with this Hamiltonian. He generalized the classical Runge–Lenz vector to make it a quantum mechanical invariant. This vector operator, which commutes with the Hamiltonian, is

$$\vec{\mathbf{A}} = \frac{1}{\sqrt{-8m\mathbf{H}}}(\vec{\mathrm{p}} \times \vec{\mathrm{L}} - \vec{\mathrm{L}} \times \vec{\mathrm{p}}) - \frac{m}{-2H}\frac{\vec{\mathrm{r}}}{r}. \tag{51}$$

The three components of $\vec{\mathrm{A}}$ and the three components of $\vec{\mathrm{L}}$ make up the six generators of SO(4) and the quantum mechanical commutation relations among these operators are the same as the Lie algebra of the infinitesimal generators of SO(4). This group has two invariants and one of them, $\vec{\mathrm{L}} \cdot \vec{\mathrm{A}} = \vec{\mathrm{A}} \cdot \vec{\mathrm{L}}$, vanishes in this case. The other invariant is simply the sum of the squares of the six components of $\vec{\mathrm{A}}$ and $\vec{\mathrm{L}}$. It is interesting to note that linear combinations of $\vec{\mathrm{A}}$ and $\vec{\mathrm{L}}$, $\vec{\mathrm{j}}_1$, and $\vec{\mathrm{j}}_2$ commute with each other and each satisfies an SU(2) Lie algebra:

$$\vec{\mathrm{j}}_1 = \tfrac{1}{2}(\vec{\mathrm{L}} + \vec{\mathrm{A}}), \qquad \vec{\mathrm{j}}_2 = \tfrac{1}{2}(\vec{\mathrm{L}} - \vec{\mathrm{A}}),$$
$$[j_{1i},\, j_{1j}] = \sum_k \varepsilon_{ijk} j_{1k},$$
$$[j_{2i},\, j_{2j}] = \sum_k \varepsilon_{ijk} j_{2k}, \tag{52}$$
$$[j_{1i},\, j_{2j}] = 0 \qquad \text{for all} \quad i,\ j.$$

This means that SO(4) is the direct product SU(2)$\otimes$ SU(2). Bargmann showed that this is related to the solution of the Hamiltonian in parabolic coordinates (of importance in the Stark effect resulting from putting the hydrogen atom in an electric field), just as the existence of $\vec{\mathrm{L}}$ as an invariant is related to its solutions in spherical polar coordinates. Furthermore, $\vec{\mathrm{j}}_1 + \vec{\mathrm{j}}_2 = \vec{\mathrm{L}}$, which means that the two solutions are connected through a Clebsch–Gordan-type relation.

B. Isotropic Harmonic Oscillator

In atomic physics it is possible to have quantum mechanical Hamiltonians because the forces between particles are well known. The Coulomb field is an outstanding example. Another such example is the three-dimensional isotropic harmonic oscillator, and the Schrödinger equation with this potential has exact solutions which are helpful in understanding its symmetry properties. In sharp contrast, it is difficult to have a Hamiltonian in elementary particle physics because the forces between strongly interacting particles such as the neutron and proton are known only vaguely. One thus derives maximum benefit from knowledge of symmetries such as SO(4), SU(3), and SU(2) in understanding which atomic physics plays a big role.

To understand the symmetry of the isotropic harmonic oscillator it is necessary to express its Hamiltonian in terms

of coordinates and momenta (with units $\hbar = m = c = 1$)

$$H = \tfrac{1}{2}\left(p_x^2 + p_y^2 + p_z^2\right) + \tfrac{1}{2}\omega^2(x^2 + y^2 + z^2). \quad (53)$$

Here ω is the angular frequency which characterizes the oscillator. In the Schrödinger scheme of quantization the operator $\frac{1}{2}\left(p_x^2 + p_y^2 + p_z^2\right)$ is the Laplace operator $-\frac{1}{2}\nabla^2$ and the Schrödinger equation has exact solutions of the form

$$U_{nlm}(r^2) = Y_l^m(\theta,\ \phi)R_{nl}(r^2). \quad (54)$$

Since the Laplacian

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

and $r^2 = x^2 + y^2 + z^2$ both do not change on a rotation of the coordinate system in space, the Hamiltonian has rotational invariance and naturally the spherical harmonics Y_l^m are a factor in the eigenfunctions of H. However, the harmonic oscillator, like the hydrogen atom, has a higher symmetry than SO(3). Jauch, Hill, and Baker demonstrated that the group of the isotropic harmonic oscillator is SU(3), a three-dimensional generalization of SU(2), the unitary unimodular group in three dimensions. A typical element of this group is a 3×3 unitary matrix with determinant $+1$. To establish the group structure or symmetry of a Hamiltonian one should find, on the one hand, operators that commute with it and show, on the other, that these operators are the infinitesimal generators of that particular Lie group. In the 3×3 matrix there are nine complex elements or 18 real parameters. However, the conditions that need to be satisfied by the rows and columns of a unitary matrix, as well as the requirement that the determinant of the matrix should be $+1$, reduce the number to eight independent real parameters. These eight parameters lead to eight generators of the SU(3) group. Three of these are the familiar L_x, L_y, and L_z operators,

$$y\frac{\partial}{\partial z} - z\frac{\partial}{\partial y}, \qquad z\frac{\partial}{\partial x} - x\frac{\partial}{\partial z}, \qquad \text{and} \qquad x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x},$$

respectively, which are the operators that generate the SO(3) group. The other generators are the five independent components of the symmetric tensor

$$\mathbf{A}_{ij} = (1/2\omega)\left(\mathbf{p}_i\mathbf{p}_j + \omega^2\mathbf{x}_i\mathbf{x}_j\right), \quad (55)$$

the sum of whose diagonal elements is essentially the Hamiltonian. The Lie algebra of SU(3) can be conviently expressed in terms of linear combinations of these eight operators:

$$\begin{aligned}
\mathbf{H}_1 &= (1/2\sqrt{3})\mathbf{L}_z,\\
\mathbf{H}_2 &= (1/6)(\mathbf{A}_{11} + \mathbf{A}_{22} - 2\mathbf{A}_{33}),\\
\mathbf{E}_1 &= (1/2\sqrt{6})(\mathbf{A}_{11} - \mathbf{A}_{22} + 2i\mathbf{A}_{12}),\\
E_{-1} &= (1/2\sqrt{6})(\mathbf{A}_{11} - \mathbf{A}_{22} - 2i\mathbf{A}_{12}),\\
E_2 &= (1/4\sqrt{3})(\mathbf{L}_x + i\mathbf{L}_y - 2\mathbf{A}_{13} - 2i\mathbf{A}_{23}),\\
E_{-2} &= (1/4\sqrt{3})(\mathbf{L}_x - i\mathbf{L}_y - 2\mathbf{A}_{13} + 2i\mathbf{A}_{23}),\\
E_3 &= (1/4\sqrt{3})(\mathbf{L}_x + i\mathbf{L}_y + 2\mathbf{A}_{13} + 2i\mathbf{A}_{23}),\\
E_{-3} &= (1/4\sqrt{3})(\mathbf{L}_x - i\mathbf{L}_y + 2\mathbf{A}_{13} - 2i\mathbf{A}_{23}). \quad (56)
\end{aligned}$$

The structure relations can be written explicitly as

$$\begin{aligned}
&[H_i,\ H_j] = 0, && i, j = 1, 2, \ldots, l \ (l = 2 \text{ here}),\\
&[H_i,\ E_\alpha] = r_i(\alpha)E_\alpha, && \alpha = \pm1, \pm2, \ldots, \quad (57)\\
&[E_\alpha,\ E_{-\alpha}] = r_i(\alpha)H_i,\\
&[E_\alpha,\ E_\beta] = \sum_\gamma C_{\alpha\beta}^\gamma E_r, && \alpha = \beta = \pm1, \pm2, \ldots.
\end{aligned}$$

The number of mutually commuting generators of this group, called the rank of the group l, is 2 in this case, H_1 and H_2 being these generators $r_i(\alpha)$ is considered as the ith component of an l-dimensional "root vector" $\vec{\mathrm{r}}(\alpha)$. The different root vectors and their components are $\vec{\mathrm{r}}(1) = (1/\sqrt{3},\ 0)$ or $r_1(1) = (1/\sqrt{3})r_2(1) = 0$,

$$\begin{aligned}
\vec{\mathrm{r}}(-1) &= (-1/\sqrt{3},\ 0),\\
\vec{\mathrm{r}}(2) &= \left(\tfrac{1}{2}\sqrt{3},\ \tfrac{1}{2}\right),\\
\vec{\mathrm{r}}(-2) &= \left(-\tfrac{1}{2}\sqrt{3},\ -\tfrac{1}{2}\right), \quad (58)\\
\vec{\mathrm{r}}(3) &= \left(\tfrac{1}{2}\sqrt{3},\ -\tfrac{1}{2}\right),\\
\vec{\mathrm{r}}(-3) &= \left(-\tfrac{1}{2}\sqrt{3},\ \tfrac{1}{2}\right).
\end{aligned}$$

These are represented graphically in Fig. 2 in a root diagram with r_1 and r_2 as rectangular axes. The root vectors have the "orthonormal" property

$$\sum_\alpha r_i(\alpha)r_j(\alpha) = \delta_{ij}. \quad (59)$$

In the relation $[E_\alpha,\ E_\beta] = C_{\alpha\beta}^\gamma E_\gamma$ the structure constants $C_{\alpha\beta}^\gamma$ vanish whenever $\vec{\mathrm{r}}(\alpha)$ and $\vec{\mathrm{r}}(\beta)$is not a root vector. The root diagram is a concise way of showing the structure of the Lie algebra.

The well-known solutions of the isotropic harmonic oscillator can be used as basis functions for calculating the irreducible representations of the SU(3) group and enumeration of the degenerate states give us the dimensionality of the representations. This degeneracy is easily known from the expression for the energy eigenvalue of the Hamiltonian:

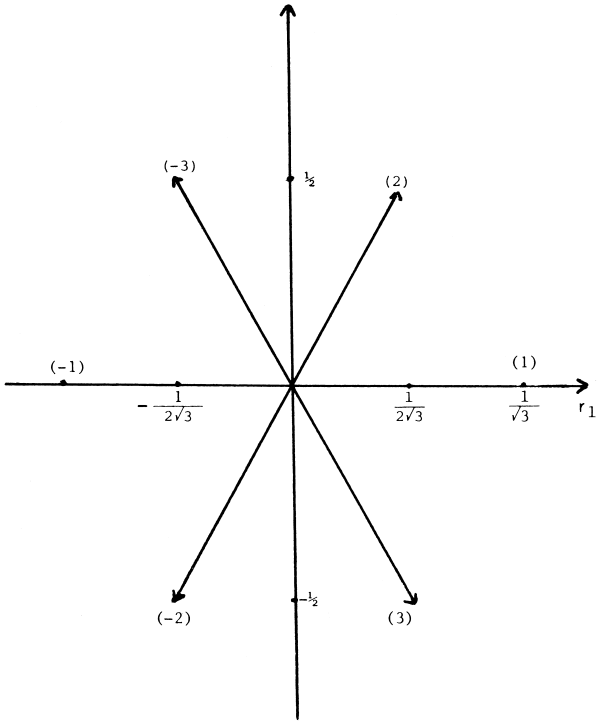


FIGURE 2 Root vector diagram. [Reprinted with permission from Swamy, N. V. V. J., and Samuel, M. A. (1979). "Group Theory Made Easy for Scientists and Engineers," Wiley-Interscience, New York. Copyright 1979 John Wiley and Sons.]

$$\begin{aligned}\mathbf{H}U_{nlm} &= \left(2(n-1)+l+\tfrac{3}{2}\right)\omega U_{nlm} \\ &\equiv \left(N+\tfrac{3}{2}\right)\omega U_{nlm}, \qquad N=0,\ 1,\ 2,\ 3,\ \ldots.\end{aligned} \tag{60}$$

Since n assumes an integral value from 1 onward and l likewise from 0 onward, the degeneracy arises from the different partitions of N between n and 1. We thus have, since for a given l there are $2l+1$ substates, 1, 3, 6, 10, etc., states having energies $\frac{2}{2}$, $\frac{5}{2}$, $\frac{7}{2}$, etc., in units of ω. Thus the dimensionality of the representations provided by the solutions of this Hamiltonian are 1, 3, 6, 10, etc. The three basis functions with $n=1$, $l=1$ (1p states) are simultaneous eigenfunctions of the commuting generators H_1 and H_2 with eigenvalues m_1 and m_2 and the latter are treated as components of a vector in l-dimensional space called a "weight vector." The weight vector for $l=1$ are shown in Fig. 3, the weight diagram. The explicit matrices of this representation are

$$H_1 = \frac{1}{2\sqrt{3}}\begin{bmatrix}1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & 0\end{bmatrix}, \quad H_2 = \frac{1}{\sqrt{6}}\begin{bmatrix}0 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & -2\end{bmatrix},$$

$$E_{-1} = \frac{1}{\sqrt{6}}\begin{bmatrix}0 & 0 & 0\\ 1 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}, \quad E_1 = \frac{1}{\sqrt{6}}\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix},$$

$$E_2 = \frac{1}{\sqrt{6}}\begin{bmatrix}0 & 0 & 1\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}, \quad E_{-2} = \frac{1}{\sqrt{6}}\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 1 & 0 & 0\end{bmatrix},$$

$$E_3 = \frac{1}{\sqrt{6}}\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}, \quad E_{-3} = \frac{1}{\sqrt{6}}\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 1\\ 0 & 0 & 0\end{bmatrix}. \tag{61}$$

The SU(3) group plays an important role in nuclear as well as particle physics.

C. Splitting of Atomic Levels

The energy levels of an electron in a spherically symmetric field (central field) in an atom are degenerate, more than one state having the same energy. Quantum mechanical perturbation theory shows that very often a perturbation splits the degenerate levels. For instance, in a hydrogen-like atom such as sodium the energy levels, taking the spin of the electron into account, are $2(2l+1)$ degenerate having 2S ($l=0$) levels, 6P ($l=1$) levels, and so on. If the magnetic field due to the spin current of the electron interacts with the magnetic field due to its orbital motion

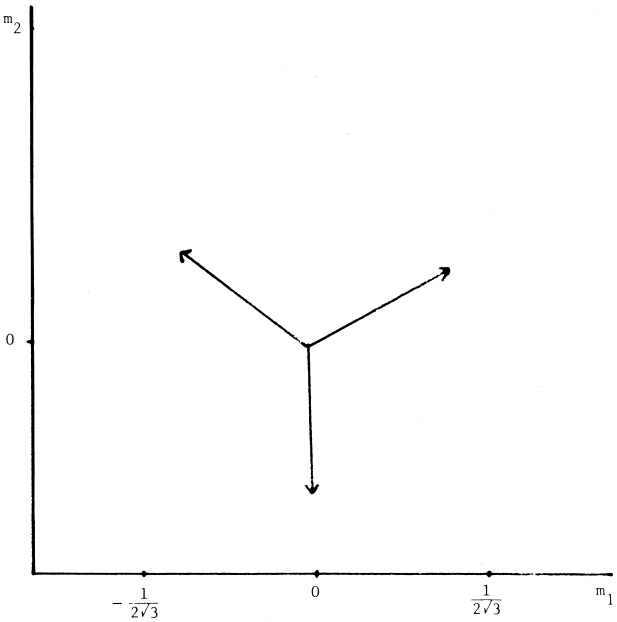


FIGURE 3 Weight diagram. [Reprinted with permission from Swamy, N. V. V. J., and Samuel, M. A. (1979). "Group Theory Made Easy for Scientists and Engineers," Wiley-Interscience, New York. Copyright 1979 John Wiley and Sons.]

around the nucleus, which is called spin–orbit coupling, then these degenerate levels are split into two groups, giving rise to what is known as fine structure of spectra lines. In the case of sodium a 5896-Å wavelength sodium D line is really made up of two lines of 5890 and 5896 Å wavelengths, the D_1 and D_2 lines. The unsplit line is due to an electric dipole transition between one of the six degenerate P states to an S state. The two lines appear because the six states split into two groups and there are transitions between levels in either group and the S state. This splitting can be predicted by group theory if the symmetries of the degenerate states and the symmetry of the perturbation are known, as was first pointed out by Bethe.

We will calculate and see how the levels of the atom are split when the perturbation arises, say from the atom being in a crystalline field of D_{3d} symmetry. The elements of D_{3d} form a subgroup of the full rotation–reflection group and hence the irreducible representations of the latter become reducible representations of the subgroup. The reduction theorem

$$a_\mu = \frac{1}{g}\sum_i g_i \chi_i^{(\mu)^*} \chi_i \tag{62}$$

gives a_μ the number of times the μth irreducible representation of the subgroup is contained in the reducible representation in which the elements of the ith class have the compound character χ_i (see Table VI). g_i is the number of elements in class K_i which have the character $\chi_i^{(\mu)}$ in the μth representation. The character table (Table VII) specifies g_i, K_i, and χ_i. The irreducible representations are classified as even or odd (g or μ) because of inversion (or the parity operation) as an element of the subgroup. In the rotation–reflection group $O(3)$ the rotation angles that correspond to elements in classes K_1, K_2, and K_3, are, respectively 0, $2\pi/3$, and $2\pi/2$, although the axes for the cyclic and dihedral rotations (i.e., rotations about axes perpendicular to the cyclic axis) are different. The characters are given by the formula

$$\chi_{(\alpha)}^{(l)} = \frac{\sin(1+\frac{1}{2})\alpha}{\sin(\alpha/2)} \qquad (\alpha = \text{angle of rotation}). \tag{63}$$

TABLE VI Compound Character Table of Subgroup D_{3d}

	E	$2C_3$	$3C_2$	i	$2S_6$	$3\sigma_d$
$\chi^{(l=6)}$	1	1	1	1	1	1
$\chi^{(1)}$	3	0	−1	−3	0	1
$\chi^{(2)}$	5	−1	1	5	−1	1
$\chi^{(3)}$	7	1	−1	−7	−1	1
$\chi^{(4)}$	9	0	1	9	0	1
$\chi^{(5)}$	11	−1	−1	−11	1	1

Applying this formula and remembering that the spherical harmonics Y_1^m have parity $(-)^l$, we obtain the compound characters displayed in Table VI. Since we know the compound characters of the subgroup and its characters $\chi_i^{(\mu)}$, calculation of a_μ following formula (62) is straightforward. For instance, the sevenfold-degenerate level $l = 3$ of the free atom (ignoring spin) is split into one level of symmetry A_{1u}, two levels of symmetry A_{2u}, and two twofold-degenerate levels each of symmetry E_μ. The splitting of the other levels is shown below. Thus group theory predicts the number of split levels and their symmetries, which implies that the transitions and spectral lines can be predicted without calculation. The magnitude of the splitting or the frequency of the spectral lines cannot, of course, be known from symmetry considerations alone. These types of symmetry considerations are of great importance in solid state. For instance, the presence in a crystal of an impurity atom occupying a site of cubic symmetry leads to the formation of a localized triply degenerate vibrational mode. If the crystal has C_{3v} point symmetry at that site the triply degenerate level will split into a doublet (two levels E type) and a singlet (a type). This is of experimental use in infrared studies of crystal defects:

$$\begin{aligned}
E_0 &\to A_{1g},\\
E_1 &\to A_{2u} + E_u,\\
E_2 &\to A_{1g} + 2E_g,\\
E_3 &\to A_{1u} + 2A_{2u} + 2E_u,\\
E_4 &\to 2A_{1g} + A_{2g} + 3E_g,\\
E_5 &\to A_{1u} + 2A_{2u} + 4E_u.
\end{aligned} \tag{64}$$

D. Selection Rules in Atomic Spectra

The spectrum radiated by an atom, hydrogen for example, consists in general of discrete lines, all of which do not have the same intensity of illumination. Also, in the spectra of different atoms it usually happens that some expected lines are not seen; these are called forbidden lines. It was Neils Bohr who pointed out that a spectral line is the result of a quantum mechanical transition between two states of energy E_i and E_f and that the frequency of the radiated line is given by the following Bohr frequency condition, which follows from the principle of conservation of energy:

$$\frac{hc}{\lambda} = h\nu = E_i - E_f. \tag{65}$$

Here ν and λ, respectively, are the frequency and wavelength of the spectral line, h Planck's constant, and c the speed of light. Since more than one atom in a gas makes

TABLE VII Character Table of D_{3d}

	$K_1(E)$ $g_1=1$	$K_2(C_3)$ $g_2=2$	$K_3(C_2)$ $g_3=3$	$K_4(i)$ $g_4=1$	$K_5(S_6=i\times C_3)$ $g_5=2$	$K_6(\sigma_d=i\times C_2)$ $g_6=3$
A_{1g} $\mu=1$	1	1	1	1	1	1
A_{2g} $\mu=2$	1	1	−1	1	1	−1
E_g $\mu=3$	2	−1	0	2 $=\chi_4^{(3)}$	−1	0
A_{1u} $\mu=4$	1	1	1	−1	−1	−1
A_{2u} $\mu=5$	1	1	−1	−1	−1	1
E_{2u} $\mu=6$	2	−1	0	−2	1	0

the same transition between quantized states, the intensity of a spectral line depends on the population of these atoms, which can be determined from the statistical distribution of the atoms at a given temperature and the quantum mechanical transition probability between these states. For instance, the transition probability for transition from a 2P state to a 1S state in a hydrogen atom is $6.25\times10^8\text{s}^{-1}$. The radiation is usually expressed as a sum of multipoles and the transition probability decreases, by several orders of magnitude, with increasing order of multipoles. That certain transitions are seen to happen experimentally and others not is related to selection rules which vary with the multipole order of the radiation. For a linearly polarized electric dipole radiation, the transition probability between states $|i\rangle$ and $|f\rangle$ is given by

$$A_{if} = (64\pi^4\nu^3/3hc^3)|\langle f|e\mathbf{z}|i\rangle|^2. \tag{66}$$

Here ν is the frequency of the radiation, $e\mathbf{z}$ the dipole operator, and $\langle f|e\mathbf{z}|i\rangle$ the quantum mechanical matrix element which depends on the nature of the initial state $|i\rangle$ and the final state $|f\rangle$. It is this matrix element that dictates whether the transition $|i\rangle\to|f\rangle$ is allowed or not. If the matrix element vanishes, then that transition is forbidden—at least in the electric dipole approximation. The selection rules then depend on this matrix element and it is the vanishing or nonvanishing of this matrix element that can be predicted by group theory without actual calculation.

Let us assume that the initial quantum state of the electron $|i\rangle$ in the hydrogen atom making the transition is the central field state $U_{nl_im_i}$, and that the final state it jumps into is $|1\rangle U_{nl_fm_f}$. The matrix element is explicitly

$$e\int U^*_{nl_fm_f}(e\mathbf{z})U_{nl_im_i}\,d\tau. \tag{67}$$

Now $U_{nl_im_i}$ is the basis function which belongs to the m_ith row of the $D^{(li)}$ irreducible representation of the rotation group. The operator $\mathbf{z}$ which is Y_1^0, apart from a constant, belongs to the $D^{(1)}$ representation and the product $zU_{nl_im_i}$ belongs to the direct product $D^{(1)}\times D^{(li)}$. More precisely, $zU_{nl_im_i}$ is a certain linear combination of the basis functions which generate the representations of the irreducible components of $D^{(1)}\times D^{(li)}$. $U_{nl_fm_i}$ belongs to the m_fth row of the $D^{(lf)}$ representation of SO(3). If $D^{(lf)}$ is not one of the terms in the direct sum into which $D^{(1)}\times D^{(li)}$ decomposes, the matrix element vanishes. According to the Clebsch–Gordan theorem we have

$$D^{(1)}\times D^{(l_i)} = D^{(l_i+1)}\oplus D^{(l_i)}\oplus D^{(l_i-1)}. \tag{68}$$

It is clear that unless $l_f=l_i+1, l_i, l_i-1$, the matrix element vanishes with the exception that when $l_i=0$, l_f can only be $+1$. Furthermore, the central field functions have definite parity. For a reflection operation $x\to-x$, $y\to-y$, $z\to-z$ these functions are even (do not change sign) if l is even, and vice versa. In other words, they are also basis functions of the inversion or parity group. Since z has odd parity, the matrix element vanishes unless $l_f=l_i+1$ or l_i-1. Since l_i cannot equal l_f, it is customary to say that parity should change in an electric dipole transition. Thus the electron can make a transition between a P state and an S state.

If an atom is not free but is in a crystalline field, say of C_{3v} symmetry, the matrix element can be analyzed as follows. The initial state function $\psi_\mu^{(i)}$ belongs to the μth row of the ith irreducible representation of C_{3v} and $\psi_\nu^{(f)}$ belongs to the νth row of the fth representation. From the character table of C_{3v} we see that z belongs to the one-dimensional representation $D^{(A_1)}$ of C_{3v}. Hence the matrix element will vanish unless $D^{(f)}$ occurs in the decomposition of the direct product $D^{(A_1)}\times D^{(i)}$. Without having to apply the systematic reduction formula one can easily see from the character table, for instance,

$$\begin{aligned} D^{(A_1)}\times D^{(A_1)} &= D^{(A_1)}\oplus OD^{(A_2)}\oplus OD^{(E)},\\ D^{(A_1)}\times D^{(A_2)} &= OD^{(A_1)}\oplus D^{(A_2)}\oplus OD^{(E)}. \end{aligned} \tag{69}$$

The coefficients of the sums of the right-hand side of the equation are determined by looking at the appropriate characters on both sides. Thus if $\psi_\mu^{(i)}$ refers to A_1, then $\psi_\nu^{(f)}$ also must have A_1 symmetry and likewise for A_2. Thus A_1-to-A_2 transition, or vice versa, is strictly forbidden in electric dipole radiation.

V. APPLICATIONS IN NUCLEAR PHYSICS

Nuclear physics is somewhat unique in that, because the fundamental interaction between two nucleons is only approximately known, there does not exist one model which can account for all the different phenomena observed experimentally. Thus, for instance, the shell model has successfully predicted the ground-state spins and the observed shell structure of many nuclei. The optical model has been very useful in analyzing neutron scattering. The Bohr–Mottelson model successfully explains the properties of deformed nuclei. It is therefore not surprising that more than one group has been used in understanding the different characteristics of nuclear structure and nuclear levels, and once again the pioneering work was done by Wigner in his theory of supermultiplets proposed in 1937. We will discuss here

1. The construction of the many-particle wavefunction following symmetry principles
2. A model based on the SU (3) group and then comment on recent applications.

In the L–S coupling scheme of central field wave functions the many-particle wave function is a simultaneous eigenfunction of $\vec{\mathrm{L}}^2$, L_z, $\vec{\mathrm{S}}^2$, and S_z where $\vec{\mathrm{L}}$ is a vector sum of orbital angular momenta of all the particles and $\vec{\mathrm{S}}$ the sum of spin operators:

$$\vec{\mathrm{L}} = \sum_i \vec{l}_i, \qquad \vec{\mathrm{s}} = \sum_i \vec{\mathrm{S}}_i. \tag{70}$$

The corresponding Hamiltonian is symmetrical in the interchange of particle coordinates and is free of any spin operators. The Pauli principle restricts the many-particle function to be antisymmetric in the interchange of the coordinates (space as well as spin coordinates) of any two particles because all the nucleons are fermions with spin $\frac{1}{2}$. This implies that, when written as a product of space and spin functions, the function is symmetrical in space coordinates and antisymmetric in spin coordinates and vice versa. To illustrate this let us take a three-nucleon system and consider the spin functions first. Denoting an up-spin state by α and a down-spin state by β and labeling the particles 1, 2, 3, we have both symmetric and antisymmetric functions which are built following the Clebsch–Gordan theorem. Two such functions, one symmetric and the other antisymmetric, in the interchange of 1 and 2 are given below. These are eigenfunctions of S^2 and S_z with eigenvalues $\frac{1}{2}$ and $\frac{1}{2}$:

$$\begin{aligned}\phi_3 \equiv \chi_{1/2}^{1/2} = \frac{1}{\sqrt{6}}\{&2\alpha(1)\alpha(2)\beta(3) - \alpha(2)\alpha(3)\beta(1) \\ &-\alpha(3)\alpha(1)\beta(2)\},\end{aligned} \tag{71}$$

$$\phi_4 \equiv \chi_{1/2}^{1/2} = \frac{1}{\sqrt{2}}\{\alpha(1)\beta(2)\alpha(3) - \alpha(2)\beta(1)\alpha(3)\}.$$

Since the Hamiltonian is symmetric in the coordinates of all particles, its eigenfunctions should be basis functions of the irreducible representations of permutation group, here the group S_3. Two of the representation matrices generated by the above set, together with the respective characters, are

$$\begin{array}{cc} \mathscr{D}(A) & \mathscr{D}(C) \\ \begin{bmatrix} -1/\sqrt{2} & \sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{bmatrix}, & \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \\ \chi = 0 & \chi = 0 \end{array} \tag{72}$$

We have seen that the most straightforward way of generating basis functions for irreducible representations, starting with any arbitrary function, is by means of projection operators. For the permutation group there is an elegant theory of Young tableaux for deriving these projection operators. Let us assume three boxes arranged in rows and columns as

$$\begin{array}{cc} \boxed{1} & \boxed{2} \\ \boxed{3} & \end{array} \qquad \begin{array}{cc} \boxed{1} & \boxed{3} \\ \boxed{2} & \end{array} \tag{73}$$

A standard tableau is an arrangement of the numbers 1, 2, 3 in the boxes such that they are in increasing order in rows as well as columns, as shown above. The number of such standard tableaux is the dimensionality of the representation, which in this case is two. Since each diagram is obtained from the other by interchanging rows and columns, these are said to be conjugate to each other. The projection operators corresponding to these two diagrams are

$$\begin{aligned} \mathbf{P}_1 &= \tfrac{1}{3}[E - (13)][E + (12)] \\ &= \tfrac{1}{3}(E + C - B - D), \\ \mathbf{P}_2 &= \tfrac{1}{3}[E - (12)][E + (13)] \\ &= \tfrac{1}{3}(E - C + B - F), \end{aligned} \tag{74}$$

where we used the notation of Section II. Operating with these on $\alpha(1)\alpha(2)\beta(3)$, for instance, generates the basis functions

$$\begin{aligned} &\tfrac{2}{3}\{\alpha(1)\alpha(2)\beta(3) - \alpha(2)\alpha(3)\beta(1)\}, \\ &\tfrac{1}{3}\{\alpha(2)\alpha(3)\beta(1) - \alpha(1)\alpha(3)\beta(2)\}. \end{aligned} \tag{75}$$

The elements A, C of the S_3 group are represented in this basis by the matrix

$$\begin{array}{cc} \mathscr{D}(A) & \mathscr{D}(C) \\ \begin{bmatrix} 0 & -\frac{1}{2} \\ -2 & 0 \end{bmatrix}, & \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}. \\ \chi = 0 & \chi = 0 \end{array} \tag{76}$$

Other matrices can be calculated in a similar manner. The characters show that this representation is equivalent to the one in Eq. (72). Thus linear combinations of functions of the type $\alpha(1)\alpha(2)\beta(3)$ give rise to basis functions identical to the ones in Eq. (71). The functions of the space coordinates can be treated likewise. On the one hand, since the single-particle states are described by central field functions (product of spherical harmonics and appropriate radial function) the three-particle function must be an eigenfunction of $\vec{\mathrm{L}}^2$ and L_z. This means that the latter should be a Clebsch–Gordan-type linear combination of products of single-particle functions. On the other hand, they should be basis functions for the irreducible representations of S_3, which means that these should be obtainable by means of projection operators associated with appropriate Young tableaux. Since the wavefunction, a product of space and spin functions, has to be antisymmetric in the interchange of all the coordinates (space and spin) of any two particles, we need spatially symmetric functions multiplying spin antisymmetric functions and vice versa. The space and spin functions then correspond to conjugate Young diagrams. More details about these can be known from the work of Swamy and Samuel.

A. The Elliott Model

In 1958 Elliott introduced a model of the nucleus wherein the particles move in a common harmonic oscillator potential and in addition have a mutual interaction of the quadrupole type. Before writing this Hamiltonian which mathematically describes the model, let us recall some of the features of the nonrelativistic isotropic harmonic oscillator in three dimensions discussed earlier. The simultaneous eigenfunctions of energy and angular momentum have the central field form

$$U_{nlm}(\vec{\mathrm{r}}) = Y_l^m(\theta, \phi) R_{nl}(r), \tag{77}$$

which means three quantum numbers define a single state. While the energy levels are equally spaced they are degenerate. The ground state is nondegenerate, but the higher excited states have degeneracies 3, 6, 10, etc., which is the underlying basis for the shell structure. These states are labeled $1s$, $1p$ $(2s, 1d)$, $(2p, 1f)$, and so on. Because of the presence of Y_l^m as a factor, these are basis functions for the irreducible representation of the rotation group (the orbital angular momentum group), the Hamiltonian being invariant to rotations in space. However, as we saw earlier, the bigger group is SU(3) with eight infinitesimal generates H_1, H_2, E_1, E_{-1}, E_2, E_{-2}, E_3, E_{-3} satisfying the Lie algebra appropriate to this group, and commuting with the Hamiltonian of the oscillator. The Casimiar operator $\mathbf{C}$ is simply the sum of squares of these eight operators and this commutes with every generator of the group. As Lipkin has shown, the Elliott Hamiltonian can be written

$$\mathbf{H} = H_0 + \lambda_1 V \mathbf{C} + \lambda_2 V \vec{\mathrm{L}}^2, \tag{78}$$

where λ_1 and λ_2 are constants and V is related to the strength of the quadrupole interaction between two particles. H_0 describes the independent particle motion in an oscillator potential. The additional terms are intended to remove some of the degeneracies associated with H_0; in other words, the SU(3) multiplets are split. Since the added terms commute with H_0 the eigenfunction of H is also simultaneously an eigenfunction of $\mathbf{C}$ as well as the angular momentum $\vec{\mathrm{L}}^2$. The required eigenfunctions are chosen to make sure that they are SU(3) multiplet states and within these multiplets they are also eigenfunctions of the angular momentum $\vec{\mathrm{L}}^2$. Since the energy levels corresponding to the eigenvalues of $\vec{\mathrm{L}}^2$ are in the nature of rotational energy level, each SU(3) multiplet constitutes a rotational band. Thus the rotational spectrum, which is assumed to be due to collective motions of the nuclear particles, is derivable from an SU(3)-type independent-particle shell model under the assumption of a quadrupole two-body interaction. Experiments have shown, for instance, that in ^{20}Ne (whose outer nucleons can be considered to be in the $2s$–$1d$ harmonic oscillator shell) there are excited levels, corresponding to a rotational spectrum, of energies 9.48 MeV(2^+), 10.24 MeV(2^+), 10.64 MeV(6^-), 11.99 MeV(8^+), and so on.

We conclude this section by remarking that currently much research activity is related to supersymmetries such as $U(6/4)$ and $U(6/12)$ starting with the interacting boson model initiated by Iachello and Arima. In the original interacting boson model the valence neutrons and protons, which are fermions individually, are paired into "s" and "d" bosons in even–even nuclei, similar to Cooper pairs in the theory of superconductivity. The $U(6/4)$ is extended

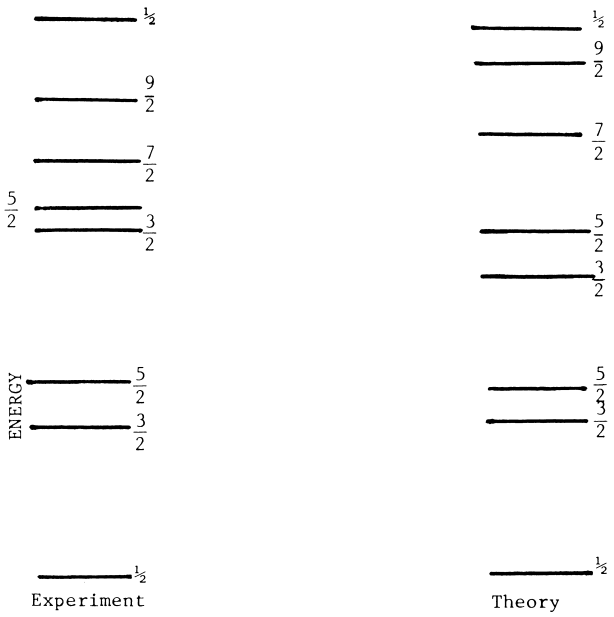


FIGURE 4 Experimental and theoretical levels of ^{195}Pt.

to $U(6/12)$ to include fermions and account for the spectra of odd–even nuclei as well. In Fig. 4 we show the success of this model in predicting the level structure of ^{195}Pt. The levels in this figure are not, however, drawn to the exact scale of numerical energies.

VI. APPLICATIONS IN MOLECULAR AND SOLID-STATE PHYSICS

A. Normal Modes of Vibration in Molecules

The application of group theory to the vibrations of symmetrical molecules was done by Wigner, who analyzed the normal modes of tetrahedral methane. We will illustrate his method by studying the vibrations of an OL_4-type molecule having a C_{4v} symmetry. Stephenson and Jones made an experimental study of the vibrational spectra of BrF_5 and concluded from its features that this molecule should have C_{4v} symmetry. The geometric arrangement of the five atoms in the OL_4 arrangement is shown in Fig. 5. In the case of BrF_5, the fifth fluorine atom is vertically above the bromine atom which is located at the pyramidal point O in the figure. The elements of this group of order $g = 8$ are the identity E (E), three cyclic rotations through angles $2\pi/4$, $4\pi/4$, and $6\pi/4$, respectively ($C_4 = M$, $C_4^2 = N$, $C_4^3 = P$), about the Z axis passing through the point O, reflections in vertical planes XZ and YZ (Q and S, respectively), and two reflections in vertical planes containing O and the diagonal L_1L_3 and L_2L_4 (T and U, respectively). Reflections, or improper rotations so-called, are treated as rotations through the angle 0° in this analysis. The group has five classes $K_1(E)$, $K_2(N_4)$, $K_3(M_4, P_4)$, $K_4(Q, S)$, and $K_5(T, U)$, and therefore has five irreducible representations. Two of these are one-dimensional A_1, A_2 symmetric with respect to the cyclic

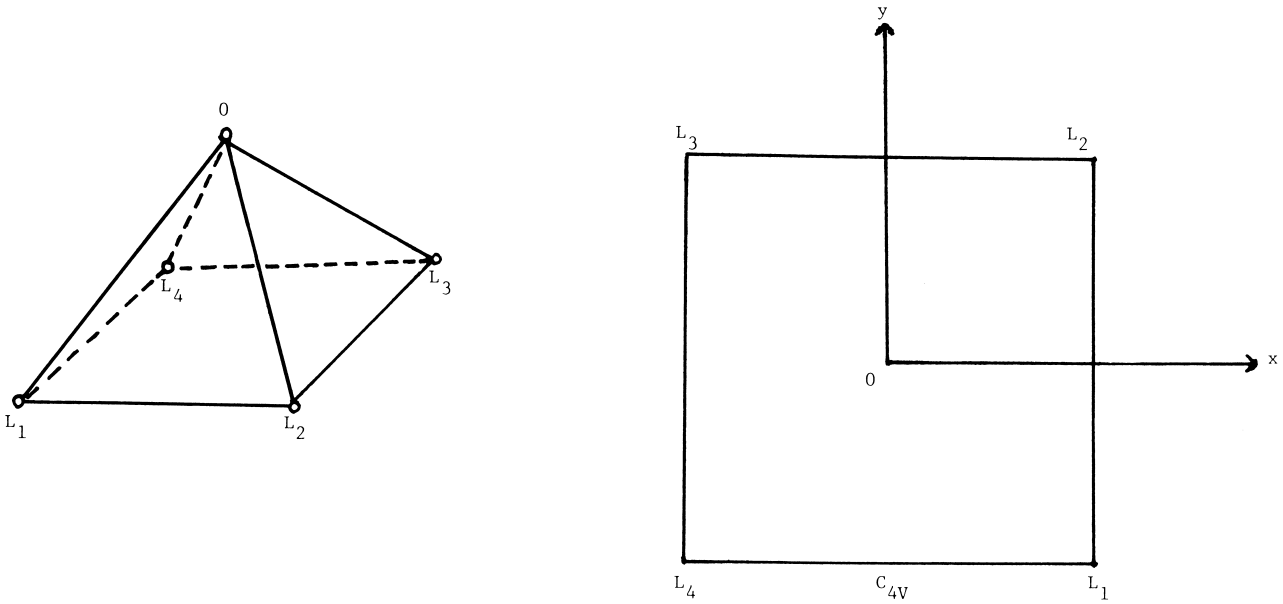


FIGURE 5 Arrangement of atoms in OL_4 molecule. [Reprinted with permission from Swamy, N. V. V. J., and Samuel, M. A. (1979). "Group Theory Made Easy for Scientists and Engineer," Wiley-Interscience, New York. Copyright 1979 John Wiley and Sons.]

axis, two antisymmetrical one-dimensional representations B_1 and B_2, and a degenerate two-dimensional representation. The subscripts 1 and 2 and A and B indicate whether it is symmetric or antisymmetric with respect to the vertical reflections. Wigner's fundamental demonstration has been that the allowed normal modes of any molecule with a certain symmetry (belonging to a group) should correspond to the symmetries of these irreducible representations and none else.

According to the theory of small vibrations in classical mechanics, a system of N particles connected by springs can have only $3N-6$ normal modes unless all the particles are in one line. In the case of BrF_5 this means 12 frequencies, and these modes should have the symmetries of the C_{4v} representations. We will apply the Wigner prescription to ascertain these symmetries. The first step is to calculate the compound character $\chi(R)$ of each symmetry operation following the rule

$$\begin{aligned}\chi(R) &= (\mu_R - 2)(1 + 2\cos\phi) \\ &\quad \text{for proper rotations,} \\ &= \mu_R(-1 + 2\cos\phi) \\ &\quad \text{for improper rotations.}\end{aligned} \tag{79}$$

Here μ_R indicates the number of atoms left unchanged by that particular group operation. These compound characters as well as the characters of the C_{4v} group (not classwise) are shown in Table VIII. Following the usual reduction formula relating characters to compound characters, the number of modes of each symmetry type is

$$N_i = \frac{1}{g}\sum_R \chi(R)\chi_i(R). \tag{80}$$

Calculation shows that there will be $3A_1$, $0A_2$, $1B_1$, $2B_2$, and three sets of degenerate E-type vibrations. For instance, for the A_1 type we have

$$\begin{aligned}N_{A_1} &= \tfrac{1}{8}[(12\times 1) + (1\times 0) + (1\times 0) + (1\times 0) \\ &\quad + (2\times 1) + (2\times 1) + (4\times 1) + (4\times 1)] \\ &= 3\end{aligned} \tag{81}$$

Since the E type are doubly degenerate, we notice that we have accounted for all 12 frequencies. Pictures of the modes of vibration of all important symmetry types are found in the classic work of Herzberg. The transformation properties of the components of the dipole moment vector (essentially x or y or z) and the components of the polarizability tensor (xz, yz, $x^2 - y^2$, etc.) determine whether these modes are Raman active or infrared active. From the character table we notice that both these sets of components transform as the A_1 and E representations, whereas only the components of the polarizability tensor have the B-type symmetries. Thus the A_1 and E modes are both Raman and infrared active, whereas B_1 and B_2 are only Raman active. The experimentally measured frequencies of Stephenson and Jones (expressed in units of cm^{-1}) are 365, 572, 683 (A type), 315, 481, 536 (B type), 244, 415, 626 (E type).

It is important to note that group theory predicts only the symmetries of the normal modes but not their numerical frequencies. One needs to solve the secular determinant to obtain these frequencies and even then the numerical values are dependent on assumed force constants between the atoms. In general the calculation is cumbersome, but considerable reduction in labor and time will result if "symmetry-adapted eigenfunctions (symmetry coordinates)" are used in the evaluation of the individual matrix elements. These latter set of functions are easily obtained by means of projection operators. Four of the representations are one-dimensional and we give

TABLE VIII Characters and Compound Characters of C_{4v}

	E	M	N	P	Q	S	T	U
A_1 z, x^2+y^2, z^2	1	1	1	1	1	1	1	1
A_2	1	1	1	1	−1	−1	−1	−1
B_1 x^2-y^2	1	−1	1	−1	1	1	−1	−1
B_2 xy	1	−1	1	−1	−1	−1	1	1
E $(x, y)(xz, yz)$	2	0	−2	0	0	0	0	0
ϕ	0°	90°	180°	270°	0°	0°	0°	0°
$\pm 1 + 2\cos\phi$	3	1	−1	1	1	1	1	1
μ_R	6	2	2	2	2	2	4	4
$\chi(R)$	12	0	0	0	2	2	4	4

a set of irreducible representation matrices for the two-dimensional E representation:

$$\underset{\mathscr{D}(E)}{\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}},\quad \underset{\mathscr{D}(M)}{\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}},\quad \underset{\mathscr{D}(N)}{\begin{bmatrix}-1 & 0\\ 0 & -1\end{bmatrix}},\quad \underset{\mathscr{D}(P)}{\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}},$$
$$\underset{\mathscr{D}(Q)}{\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}},\quad \underset{\mathscr{D}(S)}{\begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix}},\quad \underset{\mathscr{D}(T)}{\begin{bmatrix}0 & -1\\ -1 & 0\end{bmatrix}},\quad \underset{\mathscr{D}(V)}{\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}}. \tag{82}$$

Applying formula (14), we can readily obtain the projection operators

$$\begin{aligned}
P_{A_1} &= \tfrac{1}{8}\{E + M + N + P + Q + S + T + U\},\\
P_{B_1} &= \tfrac{1}{8}\{E - M + N - P + Q + S - T - U\},\\
P_{B_2} &= \tfrac{1}{8}\{E - M + N - P - Q - S + T + U\},\\
P_{E_1} &= \tfrac{1}{4}(E - N + Q - S),\\
P_{E_2} &= \tfrac{1}{4}(E - N - Q + S).
\end{aligned} \tag{83}$$

A convenient set of symmetry coordinates is easily obtained by simply applying these projection operators to the changes in the various bond lengths and bond angles of the molecule (Table VIII).

B. Brillouin Zones and Compatibility Relations

Because of the lattice structure, an electron moves in a periodic electric potential field in the solid and the Hamiltonian describing its motion has the form

$$\mathbf{H} = -\left(\frac{h^2}{2m}\right)\nabla^2 + V(\vec{\mathrm{r}}), \tag{84}$$

where $V(\vec{\mathrm{r}})$ satisfies the periodicity condition

$$V(\vec{\mathrm{r}} + \vec{\mathrm{a}}) = V(\vec{\mathrm{r}}). \tag{85}$$

Here $\vec{\mathrm{a}}$ describes the periodicity of the lattice in three dimensions. This Hamiltonian has translational invariance and its eigenfunctions are basis functions for the irreducible representations of the group of translations in space. Bloch showed that group theory requires the stationary-state solution $\psi_{\vec{\mathrm{k}}}$ of the Schrödinger equation,

$$\mathbf{H}\psi_{\vec{\mathrm{k}}}(\vec{\mathrm{r}}) = E\vec{\mathrm{k}}\psi_{\vec{\mathrm{k}}}(\vec{\mathrm{r}}), \tag{86}$$

be of the form

$$\psi_{\vec{\mathrm{k}}}(\vec{\mathrm{r}}) = e^{i\vec{\mathrm{k}}\cdot\vec{\mathrm{r}}}U_{\vec{\mathrm{k}}}(\vec{\mathrm{r}}), \tag{87}$$

where $U_{\vec{\mathrm{k}}}$ is a periodic function with the periodicity of the lattice. This translational symmetry of the crystal is in addition to its point group symmetry of the lattice, whose elements are those rotations and reflections which always leave one point fixed. The point group is the local symmetry at the site of an atom. In an earlier section we discussed how the symmetry of the crystalline field splits the degenerate energy levels of a free atom when this is in a crystal. The complete set of symmetry operations carrying a crystal into itself, including translations and the point group operations, is known as the space group.

The eigenvalues $E(\vec{\mathrm{k}})$ are such that there are planes of energy discontinuity in $\vec{\mathrm{k}}$ space which are the boundaries of a volume, the Brillouin zone. The translational symmetry enables one to consider a single unit cell in $\vec{\mathrm{k}}$ space known as the first Brillouin zone or the reduced zone with the corresponding reduced wave vector. Bouckaert, Smoluchowski, and Wigner studied the symmetry properties of the Brillouin zone and derived "compatibility relations" between adjoining points, lines, and planes of symmetry. These relations are of fundamental importance in the analysis of solid-state experiments. For instance, energy-band calculations become considerably simplified for states along symmetry lines and at symmetry points in the Brillouin zone. Selection rules for the absorption of polarized electromagnetic radiation depend on the symmetry associated with a given point in the reduced zone. In the analysis of the vibronic spectra of doped crystals, vibronic selection rules need to be determined for phonons at various points in the Brillouin zone, and compatibility relations are indispensable for doing this. We now discuss these compatibility relations.

The Bloch function is a product of two factors, the phase function $e^{i\vec{\mathrm{k}}\cdot\vec{\mathrm{r}}}$ which determines what is known as the "star of $\vec{\mathrm{k}}$" and the periodic function $U_{\vec{\mathrm{k}}}$ which determines the "small representations" of the "group of the wave vector $\vec{\mathrm{k}}$." The "star of $\vec{\mathrm{k}}$" is the figure one obtains when a given wave vector $\vec{\mathrm{k}}$ is subjected to all the symmetry operations of the point group. If $\vec{\mathrm{k}}$ terminates on a zone boundary, two points separated by a reciprocal lattice vector are considered identical. Those elements of the point group that leave a $\vec{\mathrm{k}}$ invariant constitute a subgroup called "group of the wave vector." Suppose an irreducible representation Γ_j of the point group is decomposed in terms of the irreducible representations of its subgroup. If Γ_j, a given irreducible representation of the latter, occurs in this decomposition, it is said to be compatible with Γ_j. In band theory, compatibility relates states that can exist together in a single band.

To illustrate the star of $\vec{\mathrm{k}}$ let us assume a two-dimensional square zone, the length of whose side equals a reciprocal lattice vector. Let OE in Fig. 6 be the position of our k vector and let this correspond to the identity element of the group C_{4v} of the square. The other symmetry operations, C_4, C_4^2, C_4^3, σ_x, σ_y, σ_α, and σ_α, take the vector into the positions shown in the figure, which does resemble a

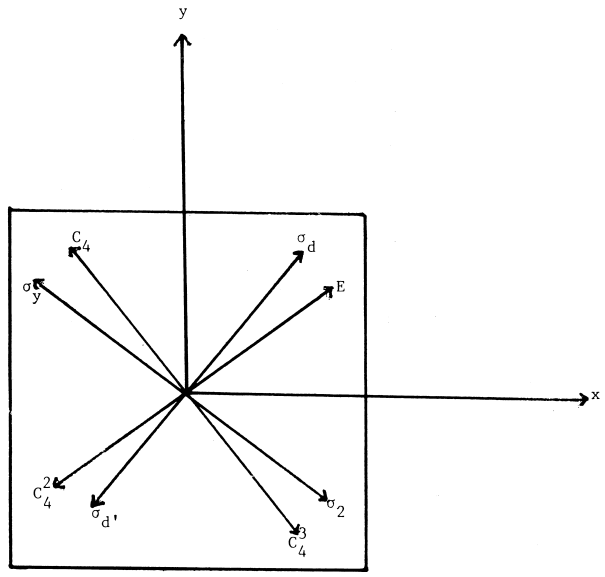


FIGURE 6 Star of $\vec{k}$. [Reprinted with permission from Swamy, N. V. V. J., and Samuel, M. A. (1979) "Group Theory Made Easy for Scientists and Engineers," Wiley-Interscience, New York. Copyright 1979 John Wiley and Sons.]

star. Since only the identity element leaves OE alone, the group of the wave vector in this case is just this trivial one (identity) element. Other cases of the star of $\vec{k}$, can be found in Wigner's original paper or the book of Tinkham. To understand the compatibility relations let us consider a Brillouin zone with the symmetry of the simple cubic lattice as in Fig. 7. The wave vector $\vec{k}=0$ (Γ) has a full symmetry O_h, as does the wave vector ΓR in the figure. To establish the compatibility relations between points Γ and Λ consider a k vector which lies at an intermediate position along the body diagonal (line joining Γ and Λ).

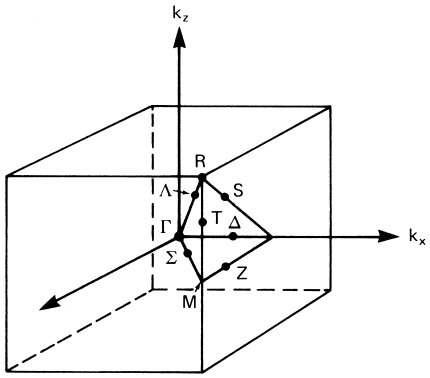


FIGURE 7 Brillouin zone with the symmetry of a cubic lattice. [Reprinted with permission from Swamy, N. V. V. J., and Samuel, M. A. (1979). "Group Theory Made Easy for Scientists and Engineers," Wiley-Interscience, New York. Copyright 1979 John Wiley and Sons.]

TABLE IX Characters of the Symmetry Elements of O_h

	E	$8C_3$	$6C_4$	$3C_4^2$	$6C_2$	i	$8S_6$	$6S_4$	$3\sigma_v$	$6\sigma_d$
O_h										
A_{1g}	1	1	1	1	1	1	1	1	1	1
E_g	2	−1	0	2	0	2	−1	0	2	0
T_{1g}	3	0	1	−1	−1	3	0	1	−1	−1
A_{1u}	1	1	1	1	1	−1	−1	−1	−1	−1
E_u	2	−1	0	2	0	−2	1	0	−2	0
T_{1u}	3	0	1	−1	−1	−3	0	−1	1	1
C_{3v}										
A_1	1	1								1
A_2	1	1								−1
E	2	−1								0

This vector is invariant to three rotations C_3, C_3^2, C_3^3 (identity) about this line as the cyclic axis, and reflections in the diagonal planes of the cube passing through this line. These are the six symmetry operations that are elements of O_h and which also leave the wave vector $\Gamma\Lambda$ invariant. It is easy to see that these six operations are elements of the group C_{3v}. The characters of these elements in the irreducible representations of C_{3v} and the characters of the elements of O_h in some of its irreducible representations are given in Table IX. With the use of these characters of the reducible representations in the well-known reduction formula, a trivial calculation shows that

$$\begin{aligned} A_{1g} &= A_1, \qquad E_g = E, \qquad T_{1g} = A_2 + E, \\ A_{1u} &= A_2, \qquad E_u = E, \qquad T_{1u} = A_1 + E. \end{aligned} \tag{88}$$

We thus conclude that A_{1g} is compatible with A_1, E_g with E, A_{1u} with A_2, E_u with E, T_{1g} with A_2, E and T_{1u} with A_1 and E. There is then continuity between the symmetry orbitals at Λ and at Γ.

VII. APPLICATION OF LIE GROUPS IN ELECTRICAL ENGINEERING

The application of the underlying concepts of Lie groups in frequency modulation (FM) in electrical engineering is best described by means of an example. Figure 8 is a block diagram of a Wien bridge $RC(t)$ variable frequency oscillator (VFO) network, where μ is the appropriate amplifier circuit. In the mathematical analysis by Gardner of the modulation rate distortion originating in the VFO, the modulated signal output V_{FM} satisfies a somewhat complicated linear differential equation

$$V_{\mathrm{FM}} + 3\frac{\dot{c}}{c}\dot{V}_{\mathrm{FM}} + \left[\left(\frac{\dot{c}}{c} + \frac{\dot{c}^2}{c^2}\right) + \frac{1}{(Rc)^2}\right]V_{\mathrm{FM}} = 0. \tag{89}$$

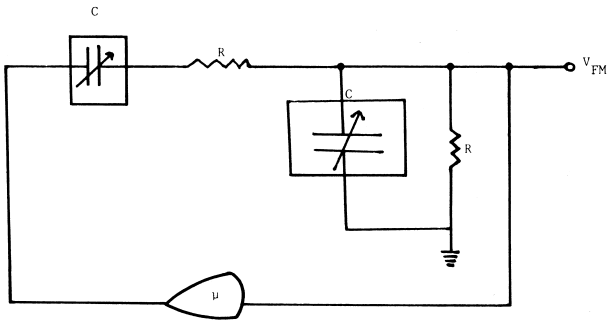


FIGURE 8 Block diagram of Wien bridge oscillator.

Here the dot above a variable means derivative with respect to time. This second-order equation can be transformed into an equivalent matrix differential equation involving coupled first-order quantities by the substitution

$$\begin{aligned} \dot{V}_1 &= -(\dot{c}/c)V_1 = (1/Rc)V_2, \\ \dot{V}_2 &= (1/Rc)V_1 - (\dot{c}/c)V_2, \qquad V_2 = V_{\mathrm{FM}}. \end{aligned} \tag{90}$$

Thus

$$\begin{bmatrix} \dot{V}_1 \\ \dot{V}_2 \end{bmatrix} = \begin{bmatrix} -\dot{c}/c & -1/Rc \\ 1/Rc & -(\dot{c}/c \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}, \tag{91}$$

which can be written in operator form $d\mathbf{V}/dt = \mathbf{A}(t)\mathbf{V}$. Here $\mathbf{V}$ is a column vector and $\mathbf{A}$ a 2×2 matrix representing the modulator element. This is sought to be solved with the initial condition $\mathbf{V}(0) = \mathbb{I}$. This can be treated as a special case of an abstract operator equation,

$$\frac{d\mathbf{U}}{dt} = \mathbf{A}(t)\mathbf{U}, \qquad \mathbf{U}(0) = I, \tag{92}$$

I being the identity operator. Magnus, and Wei and Norman developed a technique for solving this equation in the special case where $A(t)$ is expandable in terms of the infinitesimal generators $\mathbf{X}_i$ of a certain Lie group. In particular, let $\mathbf{A}(t)$ be written as

$$\mathbf{A}(t) = \sum_{i=1}^{m} a_i(t)\mathbf{X}_i, \qquad m \text{ finite} \tag{93}$$

where the time-independent operators $\mathbf{X}_i$ are the generators of a Lie group. The Lie algebra of the $\mathbf{X}_i$ is given by

$$[\mathbf{X}_\chi, \mathbf{X}_\mu] = \sum_\lambda C^\lambda_{\chi\mu}\mathbf{X}_\lambda \tag{94}$$

$C^\lambda_{\chi\mu}$ being the structure constants defining the group. In the so-called global representation of the solution Wei and Norman show that this is a product of exponentials,

$$\mathbf{U}(t) = \prod_i \exp[g_i(t)\mathbf{X}_i] \tag{95}$$

where the g_i are related to $a_i(t)$ through a set of simple first-order differential equations, and thus the question of solving (92) reduces to solving the latter more simple set of equations. As an illustration, let us assume that $\mathbf{A}(t)$ is the matrix

$$\begin{aligned} \mathbf{A}(t) &= \begin{bmatrix} i(1-\sin t) & e^{2it}\left(\frac{1}{2} - i\cos t\right) \\ -e^{-2it}\left(\frac{1}{2} + i\cos t\right) & -i(1-\sin t) \end{bmatrix}, \\ &= \sum_k a_k(t)\mathbf{X}_k. \end{aligned} \tag{96}$$

It turns out that this particular operator $\mathbf{A}$ can be expanded in terms of the generators of a Lie group isomorphic to the SO(3) group familiar in quantum mechanics, with the defining Lie algebra

$$[\mathbf{X}_i, \mathbf{X}_j] = \varepsilon_{ijk}\mathbf{X}_k. \tag{97}$$

ε_{ijk} is the well-known Levi–Civita tensor symbol, whose components can only assume values $1, -1$, or 0, and Einstein summation convention over repeated indices is implied. In the 2×2 matrix realization the generators are explicitly

$$\begin{aligned} \mathbf{X}_1 &= \begin{bmatrix} i/2 & 0 \\ 0 & -i/2 \end{bmatrix}, \qquad \mathbf{X}_2 = \begin{bmatrix} 0 & 1/2 \\ -1/2 & 0 \end{bmatrix}, \\ \mathbf{X}_3 &= \begin{bmatrix} 0 & i/2 \\ i/2 & 0 \end{bmatrix}. \end{aligned} \tag{98}$$

If we now write

$$U(t) = e^{\alpha\mathbf{X}_1}e^{\beta\mathbf{X}_2}e^{\gamma\mathbf{X}_3} \tag{99}$$

and use the Baker–Hausdorff expansion

$$\begin{aligned} &\exp(\mathbf{X}_1)\mathbf{X}_2\exp(-\mathbf{X}_1) \\ &\quad = \mathbf{X}_2 + [\mathbf{X}_1, \mathbf{X}_2] + \left(\frac{1}{2!}\right)[\mathbf{X}_1[\mathbf{X}_1, \mathbf{X}_2]] + \cdots \end{aligned} \tag{100}$$

we arrive at the defining differential equations for α, β, γ:

$$\begin{aligned} a_1 &= 2 - s\sin(t) = \dot{\alpha} + \dot{\gamma}\sin\beta, \\ a_2 &= \cos(2t) + 2\sin(2t)\cos(t) \\ &= \dot{\beta}\cos\alpha - \dot{\gamma}\sin\alpha\cos\beta, \\ a_3 &= \sin(2t) - 2\cos(2t)\cos(t) \\ &= \dot{\beta}\sin\alpha + \dot{\gamma}\cos\alpha\cos\beta. \end{aligned} \tag{101}$$

$$U(t) = \begin{bmatrix} U_{11} & U_{12} \\ (e^{-it}[\cos t \cos(t/2) - i \sin t \sin(t/2)]) & (e^{it}[\cos t \sin(t/2) - i \sin t \sin(t/2)]) \\ U_{21} & U_{22} \\ -(e^{-it}[\cos t \sin(t/2) + i \sin t \cos(t/2)]) & (e^{-it}[\cos t \cos(t/2) + i \sin t \sin(t/2)]) \end{bmatrix} \tag{102}$$

Solving these leads to the final solution given in Eq. (102) above. If, on the other hand, the given U is treated as a column vector, then either column of the above matrix satisfies the equation with the initial values $\binom{1}{0}$ and $\binom{0}{1}$, respectively.

It is interesting to note that, unlike in quantum mechanics or elementary particle physics, the properties of the Lie group itself do not seem to be amenable to a direct physical interpretation bearing on the behavior of the VFO, although the structure constants that define the group are involved in the differential relations satisfied by the g_i's in the exponential solutions.

VIII. APPLICATIONS IN PARTICLE PHYSICS

Group theory has been important in particle physics since the early 1960s. The hadrons were first classified into charge multiplets (isospin). Then approximate SU(3) symmetry was used to classify the hadrons into larger multiplets. (This is the so-called flavor SU(3).) More recently, in quantum chromodynamics (QCD), exact SU(3) color symmetry is necessary to describe the strong interactions.

A. Unitary Irreducible Representations of SU(2)

Consider the transformation

$$\begin{aligned} U(a,\ b)\text{:} \begin{bmatrix} u' \\ v' \end{bmatrix} &= U(a,\ b) \begin{bmatrix} u \\ v \end{bmatrix} \\ &= \begin{bmatrix} a^* & -b \\ b^* & a \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}, \end{aligned} \tag{103}$$

with $|a|^2 + |b|^2 = 1$. U is thus a unitary matrix, $U^{-1} = U^+$, and is an element of the SU(2) group. To obtain the unitary irreducible representations consider the basis polynomials (functions of u and v)

$$f_j^m = \frac{u^{J+m} v^{J-m}}{\sqrt{(J+m)!(J-m)!}}, \qquad m = -J, \ldots, +J,$$

and

$$\begin{aligned} U(a,\ b) f_J^m &= f_J'^m = f_J^m \left[U^{-1} \begin{bmatrix} u \\ v \end{bmatrix}\right] \\ &= \frac{(au + bv)^{J+m}(-b^* u + av)^{J-m}}{\sqrt{(J+m)!(J-m)!}}. \end{aligned} \tag{104}$$

As shown in Hamermesh's (1959) text, a straightforward binomial expansion leads to the relation

$$U(a,\ b) f_J^m = \sum_{m'} f_J^{m'} D_{m'n}^J(a,\ b),$$

where the representation matrix $D_{m'm}^J$ is explicitly

$$\begin{aligned} &\sum_k \frac{\sqrt{(J+m)!(J-m)!(J+m')!(J-m')!}}{(J+m-k)!k!(J-m'-k)!(m'-m+k)!} \\ &\quad \times a^{J+m-k} a^{*J-m'-k} b^k (-b^*)^{m'-m+k}. \end{aligned} \tag{105}$$

We now show that the representation is unitary. The normalization of the f_J^m was chosen so that

$$\begin{aligned} &\sum_{m=-J}^{J} f_J^m \left(f_J^m\right)^* \\ &\quad = \sum_{m=-J}^{J} \frac{1}{(J+m)!(J-m)!} |u|^{2(J+m)} |v|^{2(J-m)} \\ &\quad = \sum_{k=0}^{2J} \frac{|u|^{2k} |v|^{2(2J-k)}}{k!(2J-k)!} = \frac{(|u|^2 + |v|^2)^{2J}}{(2J)!} \end{aligned} \tag{106}$$

by binomial expansion. Similarly,

$$\begin{aligned} \sum_{m=-J}^{J} f_J^m \left(f_J'^m\right)^* &= \sum_{m=-J}^{J} \left|U(a,b) f_J^m\right|^2 \\ &= \sum_{m=-J}^{J} \frac{|au + bv|^{2(J+M)} \times |-b^* u + a^* v|^{2(J-M)}}{(J+m)!(J-m)!} \\ &= \frac{(|u|^2 + |v|^2)(|a|^2 + |b|^2)}{(2J)!} \\ &= \frac{(|u|^2 + |v|^2)}{(2J)!} = \sum_{m=-J}^{J} \left|f_J^m\right|^2. \end{aligned} \tag{107}$$

Therefore,

$$\sum_{m=-J}^{J}\sum_{m'} f_J^{m'} D^J_{m'n} \sum_{m''} \left(f_J^{m''}\right)^* \left(D^J_{m''m}\right) = \sum_{m=-J}^{J} f_J^m \left(f_J^m\right)^*. \tag{108}$$

Since the $(2J+1)^2$ functions $f_{m'}(f_{m''})^*$ are linearly independent, the representation matrices must satisfy

$$\sum_{m=-J}^{J} D^J_{m'm}\left(D^J_{m''m}\right)^* = \delta_{m'm''}. \tag{109}$$

That is, $DD^+ = I$ and the representation are unitary.

One can show that a matrix A which commutes with all matrices of the representation $D^J(a,\ b)$ must be a multiple of the unit matrix. Hence, by Schur's lemma, the representation matrices $D^J(a,\ b)$ are irreducible.

B. Glebsch–Gordan Coefficients for SU(2) and Isospin

The SU(2) group has three infinitesimal generators, J_1, J_2, and J_3, and the Casimir invariant $J^2 = J_1^2 + J_2^2 + J_3^2$. The $(2j+1)$-dimensional representations $D^{(j)}$ are generated by the basis functions $|jm\rangle$, where the labels are, respectively, the eigenvalues of $J^2(j(j+1))$ and $J_3(m)$. The electron spin in quantum mechanics is an outstanding example of SU(2) symmetry. Consider two different irreducible representations $D^{(j_1)}$ and $D^{(j_2)}$ (say, e.g., two electron spins). $D^{(j_1)} \times D^{(j_2)}$, a direct product of the two representations, is, in general, reducible. The Clebsch–Gordan (C–G) theorem essentially relates to the reduction of the representation

$$D^{(j_1)} \times D^{(j_2)} = \textstyle\sum_J D^{(J)} \quad \text{where } |j_1 + j_2| \leqq J \leqq (j_1 + j_2), \tag{110}$$

$$|JM\rangle = \textstyle\sum_{m_1 m_2} \langle j_1 m_1 j_2 m_2 | JM\rangle |j_1 m_1\rangle |j_2 m_2\rangle, \tag{111}$$

where $|j_1 m_1\rangle$ are the basis functions of the irreducible representations $D^{(j_1)}$ and similarly for $|j_2 m_2\rangle$; $|JM\rangle$ are the basis functions of the irreducible representations $D^{(J)}$; $\langle j_1 j_2 m_1 m_2 | JM\rangle$ are real numbers called C–G (or Wigner) coefficients.

This corresponds to the quantum mechanical vector addition of two commuting angular momentum vectors $\vec{\mathrm{J}}_1 + \vec{\mathrm{J}}_2 = \vec{\mathrm{J}}$. The derivation of these coefficients is given in several textbooks and Wigner's well-known classic (1959). Here we give instead a brief demonstration by induction for the case $j_2 = \frac{1}{2}$.

According to the C–G theorem we have

$$\begin{aligned}\left|J = j_1 + \frac{1}{2}, M = m\right\rangle &= \sqrt{\frac{j_1 + m + \frac{1}{2}}{2j_1 + 1}}\left|j_1, m_1 = m - \frac{1}{2}\right\rangle \\ &\times \left|j_2 = \frac{1}{2}, m_2 = \frac{1}{2}\right\rangle + \sqrt{\frac{j_1 - m + \frac{1}{2}}{2j_1 + 1}} \\ &\times \left|j_1, m_1 = m + \frac{1}{2}\right\rangle\left|j_2 = \frac{1}{2}, m_2 = -\frac{1}{2}\right\rangle,\end{aligned} \tag{112}$$

$$\begin{aligned}\left|j_1 - \frac{1}{2}, m\right\rangle &= -\sqrt{\frac{j_1 - m + \frac{1}{2}}{2j_1 + 1}} \\ &\times \left|j_1, m_1 = m - \frac{1}{2}\right\rangle\left|\frac{1}{2}\frac{1}{2}\right\rangle \\ &+ \sqrt{\frac{j_1 + m + \frac{1}{2}}{2j_1 + 1}}\left|j_1, m - \frac{1}{2}\right\rangle\left|\frac{1}{2} - \frac{1}{2}\right\rangle.\end{aligned} \tag{113}$$

It is true for $m_1 + \frac{1}{2}$; $|j_1 + \frac{1}{2},\ j_1 + \frac{1}{2}\rangle = |j_1 j_1\rangle|\frac{1}{2}\frac{1}{2}\rangle$. Assume it is true for m. We now show it is true for $m-1$. Once a basis function of a representation $D^{(j)}$ is known, its partners can be generated by means of step-up and step-down operators J_+ and J_- that satisfy

$$J_\pm |j_1 j_3\rangle = \sqrt{(j \pm j_3)(j \pm j_3 + 1)}\,|j_1 j_3 \pm 1\rangle. \tag{114}$$

Similar relations are well known in quantum mechanics where, for instance, $|l\mathrm{m}\rangle$ will be a spherical harmonic when J refers to the orbital angular momentum. Incidentally, it is important to note that SU(2) is the covering group for the three-dimensional rotation group $O(3)$ to which the spherical harmonics belong. Operating with $(\boldsymbol{J}_1 + \boldsymbol{J}_2)_- = J_-$ on the state $|j_1 + \frac{1}{2},\ m\rangle$ we have

$$\begin{aligned}J_-\left|j_1 + \frac{1}{2}, m\right\rangle &= \sqrt{\frac{j_1 + m + \frac{1}{2}}{2j_1 + 1}} \\ &\times \left\{\sqrt{\left(j_1 + m - \frac{1}{2}\right)\left(j_1 - m + \frac{3}{2}\right)}\right. \\ &\times \left|j_1, m - \frac{3}{2}\right\rangle\left|\frac{1}{2}, \frac{1}{2}\right\rangle \\ &+ \left.\left|j_1, m - \frac{1}{2}\right\rangle\left|\frac{1}{2}, -\frac{1}{2}\right\rangle\right\}\end{aligned}$$

$$
\begin{aligned}
&+\sqrt{\frac{j_1-m+\frac{1}{2}}{2j_1+1}}\\
&\times\sqrt{(j_1+m+\tfrac{1}{2})(j_1-m+\tfrac{1}{2})}\\
&\times\left|j_1,m-\frac{1}{2}\right\rangle\left|\frac{1}{2},-\frac{1}{2}\right\rangle\\
&=\sqrt{\frac{j_1-m+\frac{1}{2}}{2j_1+1}}(j_1+m-\tfrac{1}{2})\\
&\times\left(j_1-m+\frac{3}{2}\right)\left|j_1,m-\frac{3}{2}\right\rangle\\
&\times\left|\frac{1}{2},\frac{1}{2}\right\rangle+\left(j_1-m+\frac{3}{2}\right)\\
&\times\left|j_1,m-\frac{1}{2}\right\rangle\left|\frac{1}{2},-\frac{1}{2}\right\rangle. \qquad (115)
\end{aligned}
$$

The representations being unitary, the basis functions must be orthonormal. We normalize Eq. (115) by multiplying it with the factor

$$
\frac{1}{\sqrt{(j_1+m+\frac{1}{2})(j_1-m+\frac{3}{2})}}
$$

and obtain

$$
\begin{aligned}
\left|j_1+\frac{1}{2},m-1\right\rangle&=\sqrt{\frac{j_1+m-\frac{1}{2}}{2j_1+1}}\left|j_1,m-\frac{3}{2}\right\rangle\left|\frac{1}{2}\frac{1}{2}\right\rangle\\
&+\sqrt{\frac{j_1-m+\frac{3}{2}}{2j_1+1}}\left|j_1,m-\frac{1}{2}\right\rangle,\left|\frac{1}{2},-\frac{1}{2}\right\rangle, \qquad (116)
\end{aligned}
$$

and this means that the result is true for $m-1$, and therefore true for all m. In particular, when $m=j_1-\frac{1}{2}$ we have

$$
\begin{aligned}
\left|j_1+\frac{1}{2},j_1-\frac{1}{2}\right\rangle&=\sqrt{\frac{2j_1}{2j_1+1}}|j_1,j_1-1\rangle\left|\frac{1}{2},\frac{1}{2}\right\rangle\\
&+\sqrt{\frac{1}{2j_1+1}}|j_1,j_1\rangle\left|\frac{1}{2},-\frac{1}{2}\right\rangle. \qquad (117)
\end{aligned}
$$

Now for the case $j=j_1-\frac{1}{2}$ we note that the basis function $|j_1-\frac{1}{2},j_1-\frac{1}{2}\rangle$ must be normalized and orthogonal to the above state. Equation (114) gives, with $m=j_1-\frac{1}{2}$,

$$
\begin{aligned}
\left|j_1-\frac{1}{2},j_1-\frac{1}{2}\right\rangle&=-\sqrt{\frac{1}{2j_1+1}}|j_1,j_1-1\rangle\left|\frac{1}{2},\frac{1}{2}\right\rangle\\
&+\sqrt{\frac{2j_1}{2j_1+1}}|j_1,j_1\rangle\left|\frac{1}{2},-\frac{1}{2}\right\rangle, \qquad (118)
\end{aligned}
$$

which is indeed orthogonal to $|j_1+\frac{1}{2},j_1-\frac{1}{2}\rangle$ and is normalized. Hence the equation is correct for $m=j_1-\frac{1}{2}$. By going through a process of induction similar to the one for $j=j_1+\frac{1}{2}$, the proof is readily established.

C. Isospin Clebsch–Gordan Coefficients

Consider the reactions

$$
\begin{aligned}
&(1)\quad \pi^+ + p \to \pi^+ + p\\
&(2)\quad \pi^- + p \to \pi^- + p\\
&(3)\quad \pi^- + p \to \pi^0 + n
\end{aligned}
$$

The C–G expansion is

$$
\begin{aligned}
|JM\rangle&=\sum_{m_1+m_2=M}\langle lm_1\tfrac{1}{2}m_2|JM\rangle|lm_2\rangle|\tfrac{1}{2}m_2\rangle,\\
|JM\rangle&=\sum_{m_1+m_2=M}C^{JM}_{lm,\frac{1}{2}m_2}|l,m_1\rangle|\tfrac{1}{2},m_2\rangle,\\
|\tfrac{3}{2}\tfrac{3}{2}\rangle&=|1,\ 1\rangle|\tfrac{1}{2},\ \tfrac{1}{2}\rangle,\\
|\tfrac{3}{2}\tfrac{1}{2}\rangle&=\sqrt{\tfrac{1}{3}}|1,\ 2\rangle|\tfrac{1}{2}-\tfrac{1}{2}\rangle+\sqrt{\tfrac{2}{3}}|1,\ 0\rangle|\tfrac{1}{2}\tfrac{1}{2}\rangle,\\
|\tfrac{3}{2}-\tfrac{1}{2}\rangle&=\sqrt{\tfrac{1}{3}}|1,\ -1\rangle|\tfrac{1}{2}\tfrac{1}{2}\rangle+\sqrt{\tfrac{2}{3}}|1,\ 0\rangle|\tfrac{1}{2}-\tfrac{1}{2}\rangle,\\
|\tfrac{3}{2}-\tfrac{3}{2}\rangle&=|1\ -1\rangle|\tfrac{1}{2}-\tfrac{1}{2}\rangle,\\
-|\tfrac{1}{2},\tfrac{1}{2}\rangle&=\sqrt{\tfrac{1}{3}}|\tfrac{1}{2},\tfrac{1}{2}\rangle|1,\ 0\rangle-\sqrt{\tfrac{2}{3}}|\tfrac{1}{2}-\tfrac{1}{2}\rangle|1,\ 1\rangle,\\
|\tfrac{1}{2},-\tfrac{1}{2}\rangle&=\sqrt{\tfrac{1}{3}}|\tfrac{1}{2}-\tfrac{1}{2}\rangle|10\rangle-\sqrt{\tfrac{2}{3}}|\tfrac{1}{2}\tfrac{1}{2}\rangle|1,\ -1\rangle. \qquad (119)
\end{aligned}
$$

Identifying,

$$
\begin{aligned}
|11\rangle&=|\pi^+\rangle,\\
|10\rangle&=|\pi^0\rangle,\\
|1-1\rangle&=|\pi^-\rangle, \qquad (120)\\
|\tfrac{1}{2}\tfrac{1}{2}\rangle&=|\mathrm{p}\rangle,\\
|\tfrac{1}{2}-\tfrac{1}{2}\rangle&=|\mathrm{n}\rangle.
\end{aligned}
$$

One obtains

$$
\begin{aligned}
|\tfrac{3}{2},\tfrac{3}{2}\rangle&=|p,\ \pi^+\rangle,\\
|\tfrac{3}{2},\tfrac{1}{2}\rangle&=\sqrt{\tfrac{1}{3}}|n,\ \pi\rangle+\sqrt{\tfrac{2}{3}}|p,\ \pi^0\rangle,
\end{aligned}
$$

$$
\begin{aligned}
\left|\tfrac{3}{2}, -\tfrac{1}{2}\right\rangle &= \sqrt{\tfrac{1}{3}}|p,\ \pi\rangle + \sqrt{\tfrac{2}{3}}|n,\ \pi^0\rangle \\
\left|\tfrac{3}{2}, -\tfrac{3}{2}\right\rangle &= |n,\ \pi^-\rangle, \\
\left|\tfrac{1}{2}, \tfrac{1}{2}\right\rangle &= \sqrt{\tfrac{2}{3}}|n,\ \pi^+\rangle - \sqrt{\tfrac{1}{3}}|p,\ \pi^0\rangle, \\
\left|\tfrac{1}{2}, -\tfrac{1}{2}\right\rangle &= \sqrt{\tfrac{1}{3}}|n,\ \pi^0\rangle - \sqrt{\tfrac{2}{3}}|p,\ \pi^-\rangle,
\end{aligned} \tag{121}
$$

$$
\begin{aligned}
|p,\ \pi^+\rangle &= \left|\tfrac{3}{2}, \tfrac{3}{2}\right\rangle, \\
|p,\ \pi^-\rangle &= \sqrt{\tfrac{1}{3}}\left|\tfrac{3}{2}, -\tfrac{1}{2}\right\rangle - \sqrt{\tfrac{2}{3}}\left|\tfrac{1}{2}, -\tfrac{1}{2}\right\rangle, \\
|n,\ \pi^0\rangle &= \sqrt{\tfrac{2}{3}}\left|\tfrac{3}{2}, -\tfrac{1}{2}\right\rangle + \sqrt{\tfrac{1}{3}}\left|\tfrac{1}{2}, -\tfrac{1}{2}\right\rangle,
\end{aligned} \tag{122}
$$

By SU(2) invariance of (isospin conservation in) strong interactions,

$$
\frac{\langle \pi^+ p|T|\pi^+ p\rangle}{\langle \pi^- p|T|\pi^- p\rangle} = \frac{T^{(3/2)}}{\frac{1}{3}T^{(3/2)} + \frac{2}{3}T^{(1/2)}}
$$

and

$$
\frac{\langle \pi^+ p|T|\pi^+ p\rangle}{\langle \pi^0 n|T|\pi^- p\rangle} = \frac{T^{(3/2)}}{\sqrt{\frac{2}{3}}T^{(3/2)} - \sqrt{\frac{2}{3}}T^{(1/2)}}.
$$

At low energy we may take

$$
T^{(1/2)} = 0,
$$

and

$$
\begin{aligned}
&\sigma(\pi^+ p \to \pi^+ p) : \sigma(\pi^- p \to \pi^- p) : \sigma(\pi^- p \to \pi^0 n) \\
&= |\langle \pi^+ p|T|\pi^+ p\rangle|^2 : |\langle \pi^- p|T|\pi^- p\rangle|^2 : |\langle \pi^0 n|T|\pi p\rangle|^2 \\
&= 1 : (1/3)^2 : (\sqrt{2}/3)^2 \\
&= 9 : 1 : 2.
\end{aligned} \tag{123}
$$

In general, for $T^{(1/2)} = 0$,

$$
\sqrt{2}\langle \pi^0 n|T|\pi^- p\rangle + \langle \pi^- p|T|\pi^- p\rangle = \langle \pi^+ p|T|\pi^+ p\rangle.
$$

D. SU(3) and Particle Physics

SU(3) is the group of transformations $\psi'_a = U_{ab}\psi_b$, where U is any unitary, unimodular 3×3 matrix with determinant $\|U\| \neq 0$. This is the group of the three-dimensional isotropic harmonic oscillator in quantum mechanics (see Schiff 1968). In terms of Lie's infinitesimal generators U is given by

$$
U = \exp\left\{ i \sum_{i=1}^{8} \varepsilon_k F_k \right\}
$$

where $F_k \equiv \frac{1}{2}\lambda_k$ are the infinitesimal generators. An explicit matrix form for the λ_k's, due to Gell-Mann, in which λ_3 and λ_8 are diagonal is

$$
\begin{aligned}
\lambda_1 &= \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, & \lambda_2 &= \begin{bmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \\
\lambda_3 &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, & \lambda_4 &= \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \\
\lambda_5 &= \begin{bmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{bmatrix}, & \lambda_6 &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \\
\lambda_7 &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{bmatrix}, & \lambda_8 &= \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 0 & 1/\sqrt{3} & 0 \\ 0 & 0 & 2/\sqrt{3} \end{bmatrix}.
\end{aligned} \tag{124}
$$

The usual structure relations satisfied by these generators are

$$
[F_\alpha,\ F_\beta] = \sum_\gamma C^\gamma_{\alpha\beta} F_\gamma = i \sum_\gamma f_{\alpha\beta\gamma} F_\gamma.
$$

It is more convenient to express the structure constants in terms of $f_{\alpha\beta\gamma}$ rather than $C^\gamma_{\alpha\beta}$. The nonvanishing $f_{\alpha\beta\gamma}$ are

$$
\begin{aligned}
f_{123} &= 1, & f_{246} &= \tfrac{1}{2}, & f_{367} &= -\tfrac{1}{2}, \\
f_{147} &= \tfrac{1}{2}, & f_{257} &= \tfrac{1}{2}, & f_{458} &= \sqrt{3}/2 \\
f_{156} &= -\tfrac{1}{2}, & f_{345} &= \tfrac{1}{2}, & f_{678} &= \sqrt{3}/2,
\end{aligned} \tag{125}
$$

The $f_{\alpha\beta\gamma}$ are odd under permutation of any two indices.

We now define the combinations of generators, using the standard notation $T_\pm$, T_3 for isospin instead of $J_\pm$, J_3:

$$
\begin{aligned}
T_\pm &= F_1 \pm i F_2, & U_\pm &= F_6 \pm i F_7, \\
V_\pm &= F_4 \pm i F_5, & & \\
T_3 &= F_3, \qquad \text{and} & Y &= (2/\sqrt{3})F_8.
\end{aligned} \tag{126}
$$

The commutation relations satisfied by these operators can easily be derived. For example,

$$
\begin{aligned}
[T_3, T_\pm] &= [F_3, F_1 \pm i F_2] \\
&= [F_3, F_1] \pm i[F_3, F_2] \\
&= i F_2 \pm F_1 = \pm T_\pm.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&[Y, T_\pm] = 0 = [T_3, Y], [T_3, U_\pm] = \mp\tfrac{1}{2}U_\pm, \\
&[Y, U_\pm] = \pm U_\pm,
\end{aligned} \tag{127}
$$

$$[U_+, U_-] = \tfrac{3}{2}Y - T_3 = 2U_3,$$
$$[V_+, V_-] = \tfrac{3}{2}Y + T_3 = 2V_3, \quad \text{etc.}$$

The generators have been chosen so that T_3 and Y are diagonal with eigenvalues t_3 and y. The states ψ are labeled with t_3 and y. The action of the shift operators $T_\pm$, $U_\pm$, and $V_\pm$ on the state $\psi(t_3', y')$ is illustrated on the two-dimensional grid in Fig. 9. The action of $T_\pm$ is determined by the commutation relations

$$[T_3, T_\pm] = \pm T_\pm \qquad \text{and} \qquad [Y, T_\pm] = 0,$$

that of $U_\pm$ by

$$[T_3, U_\pm] = \mp\tfrac{1}{2}U_\pm \qquad \text{and} \qquad [Y, U_\pm] = \pm U_\pm,$$

and that of $V_\pm$ by

$$[T_3, V_\pm] = \pm\tfrac{1}{2}V_\pm \qquad \text{and} \qquad [Y, V_\pm] = \pm V_\pm. \quad (128)$$

One can generate all the states of an irreducible representation by repeated application of the shift operators to any one of them. The irreducible representations can be labeled by the pair of integers (p, q). If one begins at the unique state of maximum t_3, $\psi_{\max}$, the boundary of distribution of occupied sites (states), is p steps long in the -120° direction (i.e., $V_-^{p+1}\psi_{\max} = 0$) and $V_-^p\psi_{\max}$ is proportional to the state at the next corner, moving clockwise along the boundary from $\psi_{\max}$. If one now heads in the $-t_3$ direction, there are q steps until the next corner in the boundary is reached [i.e., $T_-^{q+1}(V_-^p\psi_{\max}) = 0$] and $T_-^q(V_-^p\psi_{\max})$ is proportional to the state at the second corner, clockwise along the boundary from $\psi_{\max}$ (the boundary is always convex).

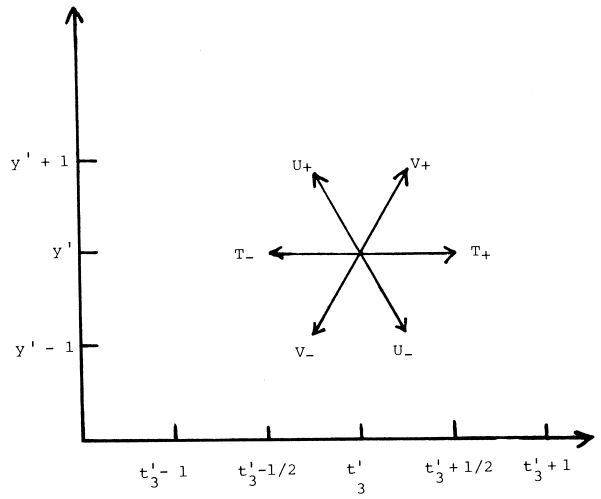


FIGURE 9 Results of shift operations on the state $\psi(t_3', y')$. [Reprinted with permission from Swamy, N. V. V. J., and Samuel, M. A. (1979). "Group Theory Made Easy for Scientists and Engineers," Wiley Interscience, New York. Copyright 1979 John Wiley and Sons.]

The multiplicity (dimensionality) of the representation (p, q) is $N = \frac{1}{2}(p+1)(q+1)(p+q+2)$. This can be seen as follows. Consider the case $p > q$. The boundary is, in general, a sixsided figure with sides of length p, q, p, q, p, q as one goes around clockwise from $\psi_{\max}$ and each with interior angles of 120°. There is one state at each site on the boundary. As one moves in from the boundary, each successive six-sided figure has the length of each side reduced by one, and the number of states at each site increased by one. This continues until the short sides are reduced to zero and one has an equilateral triangle with sides of length $(p-q)$ and multiplicity $(q+1)$. As one continues inward now, the multiplicity remains $(q+1)$.

The number of states on the boundary and inside the triangle is

$$(q+1)\sum_{k=1}^{p-q+1} k = \tfrac{1}{2}(q+1)(p-q+1)(p-q+2). \quad (129)$$

The number of states on the six-sided figures is given by

$$3\sum_{k=0}^{q-1}(q-k)(p-q+2k+2).$$

Therefore,

$$\begin{aligned} N &= \frac{1}{2}(q+1)(p-q+1)(p-q+2) \\ &\quad +3\sum_{k=0}^{q=1}(q-k)(p-q+2k-2), \\ &= \frac{1}{2}(p+1)(q+1)(p-q+2) \\ &\quad -\frac{1}{2}q(q+1)(p-q+2) \\ &\quad +3(p-q+2)\frac{1}{2}(q+1)+6\sum_{k=0}^{q-1}(q-k)k. \end{aligned} \quad (130)$$

and since

$$\sum_{k=0}^{q-1}(q-k)k = \tfrac{1}{6}(q+1)q(q-1),$$

we readily get

$$N = \tfrac{1}{2}(p+1)(q+1)(p+q+2).$$

This result, symmetric under interchange of p and q, is also valid for $p < q$.

If one now applies U_+ to the state of maximum t_3, $\psi_{\max}$, one moves along the boundary in the counterclockwise direction, reaching the next corner, ψ' after q steps. These $q+1$ states form a U-spin multiplet of

which $\psi_{\max} = |U = \frac{1}{2}q, u_3 = -\frac{1}{2}q\rangle$ is the lowest state and $\psi' = |U = \frac{1}{2}q, u_3 = +\frac{1}{2}q\rangle$ is the highest state. ψ' carries the maximum value of y that occurs in the representation $y_{\max}$, since continuing counterclockwise from ψ' the (convex) boundary runs parallel to the t_3 axis and then turns downward to smaller values of y, but u_3, y, and t_3 are not independent: $\frac{3}{2}y - t_3 = 2u_3$. Applying this equation to $\psi'(t_3 = \frac{1}{2}p, u_3 = \frac{1}{2}q)$ we find

$$\psi_{\max} = \tfrac{4}{3}\left(\tfrac{1}{3}q\right) + \tfrac{2}{3}\left(\tfrac{1}{2}p\right) = \tfrac{1}{3}(p + 2q). \qquad (131)$$

[The eigenvalue of Y for $\psi_{\max}$ is $\frac{1}{3}(p - q)$]. One can also find the value of t_3 carried by $\psi_{\max}$, $(t_3)_{\max}$ and hence the largest value of t, $t_{\max}$, which occurs in the representation. $t_{\max} = (t_3)_{\max} = \frac{1}{2}(p + q)$, since in moving back from ψ' to $\psi_{\max}$ the value of t_3 increases by $\frac{1}{2}$ for each of the q steps.

One makes here the usual association of the three (3) representation with the quarks and the three-star (3^*) representation with the antiquarks, y with hypercharge $y = B + S$ (B is the baryon number and S the strangeness), and t_3 with the third component of isospin. The formula $y_{\max} = \frac{1}{3}(p + 2q)$ gives $y_{\max} = \frac{1}{3}$ for the three representation, $y_{\max} = \frac{2}{3}$ for the three-star representation. Using $t_{\max} = \frac{1}{2}(p + q)$ one obtains $t_{\max} = \frac{1}{2}$ for both the three and three star. Using the Gell-Mann–Nishijima relation for the charge (in units of the proton charge) $Q = t_3 + \frac{1}{2}y$ we obtain the usual quantum numbers for the quarks and antiquarks (Table X). These representations are shown in Fig. 10. Figure 11 shows the usual association of the octet with the 0^- mesons and the $\frac{1}{2}^+$ baryons (there are other octets with $J^P = 1^-, 2^+$). The octet is easily constructed using the aforementioned rules and the values given by our formulas $y_{\max} = 1$ and $t_{\max} = 1$. The decuplets (Fig. 12) are easily constructed after determining that $t_{\max} = \frac{3}{2}$, $y_{\max} = 1$ for the 10 representation and $t_{\max} = \frac{3}{2}$, $y_{\max} = 2$ for the 10^* representation. The Gell-Mann–Nishijima relation is used to obtain the charges.

TABLE X Quantum Numbers for the Quarks and the Antiquarks

	B	t	t_3	y	S	Q
p	1/3	1/2	1/2	1/3	0	2/3
n	1/3	1/2	−1/2	1/3	0	−1/3
λ	1/3	0	0	−2/3	−1	−1/3
$\bar{p}$	−1/3	1/2	−1/2	−1/3	0	−2/3
$\bar{n}$	−1/3	1/2	1/2	−1/3	0	1/3
$\bar{\lambda}$	−1/3	0	0	2/3	1	1/3

E. Gell-Mann–Okubo Mass Formula

If one assumes that the mass operator is the sum of two terms, one which transforms like a U-spin scalar ($U = 0$) and the other which transforms like $U = 1$ (this is equivalent to the usual "octet enhancement" assumption), one

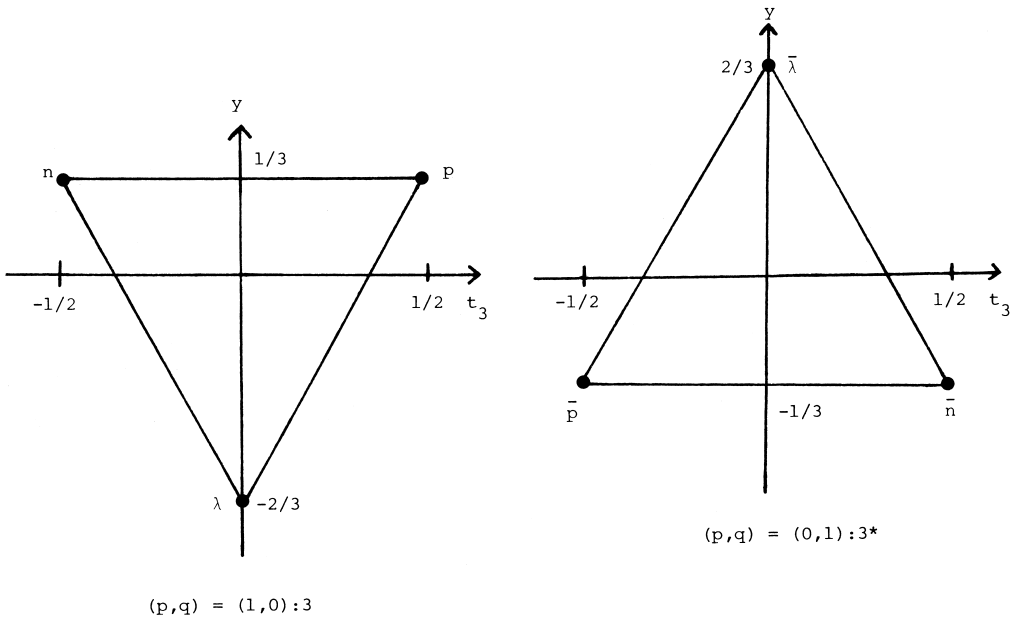


FIGURE 10 The fundamental triplet representation. [Reprinted with permission from Swamy, N. V. V. J., and Samuel, M. A. (1979). "Group Theory Made Easy for Scientists and Engineers," Wiley-Interscience, New York. Copyright 1979 John Wiley and Sons.]

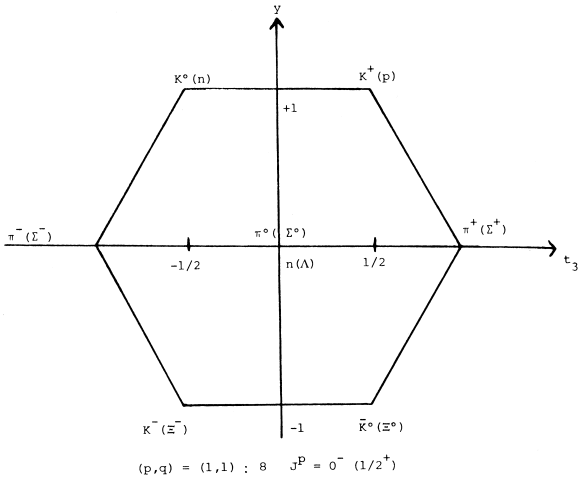


FIGURE 11 The octet. [Reprinted with permission from Swamy, N. V. V. J., and Samuel, M. A. (1979). "Group Theory Made Easy for Scientists and Engineers," Wiley-Interscience, New York. Copyright 1979 John Wiley and Sons.]

obtains $M = a + bU_3$ for a given U-spin multiplet. For the $\frac{1}{2}^+$ octet,

$$\begin{aligned} M(\Xi^0) &= a - b, \\ M(\Sigma') &= a = \tfrac{3}{4}M(\Lambda) + \tfrac{1}{4}M(\Sigma^0), \\ M(n) &= a + b, \end{aligned}$$

so that $M(\Xi^0) + M(n) = M(\Sigma')$, which leads to the famous mass formula

$$\tfrac{1}{2}(M(\Xi^0) + M(n)) = \tfrac{1}{4}(3M(\Lambda) + M(\Sigma^0)). \tag{132}$$

For the 0^- octet the corresponding formula (using M^2 for mesons instead of M) is

$$\begin{aligned} \tfrac{1}{2}(M^2(\bar{k}^0) + M^2(k^0)) &= M^2(k^0) \\ &= \tfrac{1}{4}(3M^2(\eta) + M^2(\pi^0)). \end{aligned} \tag{133}$$

For the $\frac{3}{2}^+$ decuplet,

$$\begin{aligned} M(\Omega^-) &= a - \tfrac{3}{2}b, \\ M(\Xi^*) &= a - \tfrac{3}{2}b, \\ M(\Sigma^-) &= a + \tfrac{1}{2}b, \\ M(\Delta^-) &= a + \tfrac{3}{2}b, \end{aligned} \tag{134}$$

which leads to the "equal spacing rule,"

$$\begin{aligned} M(\Omega^-) - M(\Xi^8) &= M(\Xi^*_-) - M(\Sigma^-) \\ &= M(\Sigma^-) - M(\Delta^-). \end{aligned}$$

F. SU(6) and the Quark Model

We have seen that the hadrons fit well into the representations of SU(3). This can be understood well in terms of the quarks (antiquarks) which fit into the fundamental 3 (and 3*) representations of SU(3). We now have to take spin into account. We assume that the quarks are fermions and have spin $\frac{1}{2}$. This gives us naturally the result that the mesons (qq states) have spin 0 or 1 and the baryons (qqq states) have spin $\frac{1}{8}$ or $\frac{3}{2}$

Our six-quark states (Kokkedee) consist of $|p\uparrow\rangle$, $|p\downarrow\rangle|n\uparrow\rangle$, $|n\downarrow\rangle$, $|\lambda\uparrow\rangle$, and $|\lambda\downarrow\rangle$, where the arrows refer to spin-up and spin-down. The SU(6) generalization for qq is

$$[6] \times [6] = [1] + [35]. \tag{135}$$

The SU(6) singlet has total spin zero. The 35-plet consists of an SU(3) octet with spin 0, another SU(3) octet with spin 1, and an SU(3) singlet with spin 1. This is represented as

$$[6] = \left[\{3\}, \tfrac{1}{2}\right], \qquad [\bar{6}] = \left[\{\bar{3}\}, \tfrac{1}{2}\right], \tag{136}$$

where the first entry is the SU(3) representation and the second is the spin. With this notation we have

$$[1] = [\{1\}, 0]$$

and

$$[35] = [\{1\}, 1] + [\{8\}, 0] + [\{8\}, 1]. \tag{137}$$

This accommodates very well the pseudo-scalar meson octet, the vector meson octet, and the vector meson singlet.

For the qqq states the SU(6) expansion is

$$[6] \times [6] \times [6] = [20] + [56] + [70] + [70]. \tag{138}$$

The expansions in terms of SU(3) representations and spin are

$$\begin{aligned} [20] &= \left[\{1\}, \tfrac{3}{2}\right] + \left[\{8\}, \tfrac{1}{2}\right], \\ [56] &= \left[\{8\}, \tfrac{1}{2}\right] + \left[\{10\}, \tfrac{3}{2}\right], \end{aligned}$$

and

$$[70] = \left[\{1\}, \tfrac{1}{2}\right] + \left[\{8\}, \tfrac{1}{2}\right] + \left[\{8\}, \tfrac{3}{2}\right] + \left[\{10\}, \tfrac{1}{2}\right]. \tag{139}$$

This describes very well the baryon singlets, octets, and decuplets.

G. Gauge Theories: Applications to Elementary Particle Physics

Group theory plays a very important role in our current understanding of the elementary particles, the basic building blocks of nature, and their fundamental interactions.

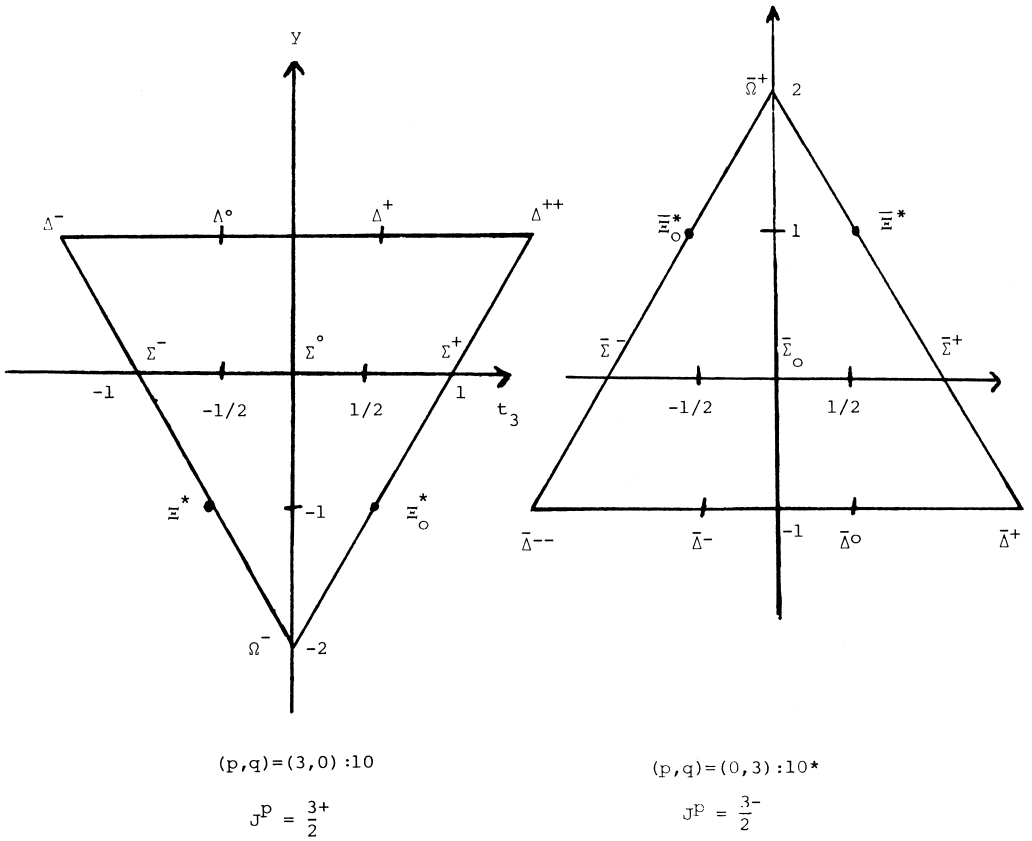


FIGURE 12 The decuplet. [Reprinted with permission from Swamy, N. V. V. J., and Samuel, M. A. (1979). "Group Theory Made Easy for Scientists and Engineers," Wiley-Interscience, New York. Copyright 1979 John Wiley and Sons.]

There are four fundamental interactions, electromagnetic, weak, strong, and gravitational. It is well established now that these are all gauge interactions based on various gauge groups which are continuous Lie groups. All the elementary particles falls neatly with the various representations of the gauge groups. The allowed interactions between the different particles are completely determined by the gauge symmetry, the mathematical requirement that the action be invariant under the transformation of the various fields under these group transformations. Historically, the pioneering step in this development was taken by H. Weyl who proposed (Weyl, Fock) that the electromagnetic interaction, currently known as quantum electrodynamics (QED), is invariant under a local $U(1)$ gauge symmetry. The experimental consequences of this symmetry had been tested to a very high degree of precision. Then, in 1954, Yang and Mills (Yang, Mills) extended this idea to include nonabelian gauge symmetry, such as SU(2), SU(3), etc. One new feature in this generalization is that the gauge bosons belonging to the adjoint representation of the gauge groups now can have interaction among themselves. [Such an interaction is not present in the $U(1)$ gauge symmetry. The photon cannot interact with itself at the tree level.] One problem in pure gauge theories is that all the gauge particles must be massless. While the photon, the gluons, and the graviton are massless, the gauge particles such as weak intermediate vector bosons, $W^{\pm}$, Z are not. Thus, part of the gauge symmetry must be broken so that some of these gauge particles can acquire masses. In 1964, P. Higgs incorporated (Higgs, Englert, Brout, Guralnik, Hagen, Kibble) the idea of spontaneous symmetry breaking with the introduction of additional scalar particles in the theory, called now the Higgs bosons. The self-interaction of these particles (the so-called Higgs potential) is such that the vacuum expectation values of some of these Higgs fields are nonzero and the gauge symmetry is broken spontaneously to a smaller gauge group. As a result, some of the gauge bosons, as well as matter fields, acquire masses. This idea of nonabelian gauge symmetry and the spontaneous symmetry braking was successfully integrated to build a unified theory of weak and electromagnetic interactions (Weinberg, Salam, Glashow). The gauge group is SU(2) $\times$ $U(1)$, which is spontaneously broken to $U_{\mathrm{EM}}(1)$. Thus the theory has one massless gauge

boson, the photon, and three massive gauge bosons, W^+, W^-, and Z. All three were discovered experimentally (Arnison *et al.*) in the CENN proton–antiproton collider at their theoretically predicted masses. All the detailed experimental predictions of this theory have been verified to better than 1% accuracy, and not a single deviation has been found. S. L. Glashow, A. Salam, and S. Weinberg were awarded the Nobel Prize in 1979 for proposing this theory, and C. Rubia and S. Van der Meer were awarded the Nobel Prize in 1984 for the experimental discovery of these theoretically predicted gauge bosons, $W^{\pm}$ and Z.

It is now well established that the protons and neutrons, which are the building blocks of the nuclei, are not elementary particles. They are made of elementary constituents called quarks, bound by interaction through a set of gauge particles called gluons. The gauge symmetry group for this interactions is SU(3), sometimes called SU(3) color. The quarks, antiquarks, and gluons have nonzero color charges under this symmetry, and the gluons can interact with themselves in addition to interacting with the quarks and antiquarks. This theory is called quantum chromodynamics (QCD). This interaction, which is responsible for the binding of three quarks, or a quarks and an antiquark, is very strong at low energy, and is responsible for the confinement of the quarks and gluons. At very high energy, the interaction becomes weaker and the theory becomes asymptotically free (Gross, Wilczek, Politzer). Many predictions of the theory at high energies have been tested experimentally to a good degree of accuracy.

Our current understanding of all three particle interactions, the electromagnetic, weak, and strong, is thus based on the gauge group $\mathrm{SU}(3) \times \mathrm{SU}(2) \times U(1)$ called the *standard model of particle physics*. All the existing elementary particles, the fermions, and gauge bosons fall neatly into various representations of this gauge group as shown in Table XI, where u_L stands for the left-handed up quark and u_R for the right-handed up quark. The index α is the SU(3) color index and takes the values 1, 2, 3. The same is true for the down quark, d. Here u_L and d_L are doublets under the weak SU(2), with u_L having $I_3 = +\frac{1}{2}$ and d_L having $I_3 = -\frac{1}{2}$. The electron-type neutrino, ν_{eL}, and the electron, e_L, have no color, and are doublets under the weak SU(2). All the right-handed fermions are singlets under SU(2). The last entries in the parentheses represent the values of all the $U(1)$ hypercharges, Y, and are related to the usual electric charge by the relation

$$Q = I_3 + \frac{Y}{2}. \tag{140}$$

Here I_3 is the third component of the weak isospin [the diagonal generator for SU(2)]. Note that all the right-handed particles are singlets under SU(2) weak, and hence they do not participate in the SU(2) weak interaction. This set of 15 chiral fermions $(u_{La}, d_{La}\nu_{eL}, e_L, u_{R\alpha}d_{R\alpha}, e_R)$ constitute what is called the first family of fermions. This is called the electron family. There are two other families, the muon family consisting of $(c_{La}, s_{La}, \nu_{\mu L}, \mu_L, c_{R\alpha}, s_{R\alpha}, \mu_R)$, and the tao family consisting of $(t_{L\alpha}, b_{L\alpha}, \nu_{\tau L}, \tau_L, t_{R\alpha}, b_{R\alpha}, \tau_R)$. The families are exact replica of each other except for the particles masses. The g_a $(a = 1\text{–}8)$ represent the eight gluons, and belong to the adjoint representation of the color SU(3). The weak gauge bosons $A_a (a = 1\text{–}3)$ belong to the adjoint representation of SU(2), and B is the gauge bosons for the $U(1)$ group. The observed gauge bosons $W^{\pm}$ are linear combination of A_1 and A_2, and are given by

$$W^{\pm} = \frac{1}{\sqrt{2}}(A_1 \mp i A_2) \tag{141}$$

while the observed photon γ and the Z boson are linear combinations of A_3 and B,

$$\begin{aligned} \gamma &= B\cos\theta_W + A_3 \sin\theta_W, \\ Z &= -B\sin\theta_W + A_3\cos\theta_W. \end{aligned} \tag{142}$$

The angle θ_W is known as the weak mixing angle. All the experimental observations, at both low and high energy scales (up to few hundred GeV) agree very well with the predictions of the standard model. However, the standard Model has several theoretically unsatisfactory features. Since it is based on the semisimple group containing the product of three group factors, it has three independent gauge couplings associated with gauge groups $U(1)$, SU(2), and SU(3), respectively. It will be theoretically much more beautiful as well as predictive if these three couplings can be unified into a simple group. Such a unification group (Georgi, Glashow, Pati, Salam), such as

TABLE XI Particle Contents of the Standard Model

Particles	SU(3)×SU(2)×U(1) representations
$u_{L\alpha}$	$(3, 2, \frac{1}{3})$
$d_{L\alpha}$	$(3, 2, \frac{1}{3})$
ν_{eL}	$(1, 2, -1)$
e_L	$(1, 2, -1)$
$u_{R\alpha}$	$(3, 1, \frac{4}{3})$
$d_{R\alpha}$	$(3, 1, -\frac{2}{3})$
e_R	$(1, 1, -2)$
g_a	$(8, 1, 0)$
A_a	$\begin{pmatrix} & & +1 \\ 1, & 3, & 0 \\ & & -1 \end{pmatrix}$
B	$(1, 1, 0)$
H	$(1, 2, 1)$

SU(5), has been proposed. This not only relates the three coupling strengths g, g', and g_3, it also predicts many interesting new phenomena such as proton decay, nonzero neutrino masses, matter–antimatter asymmetry etc.

So far, we have left out the gravitational interaction. This is also a gauge theory, the gauge group being that of a general coordinate transformation. Is it possible to unify also the gravity with the three gauge interactions of the standard model? Recent development of a new symmetry, called the supersymmetry (Wess, Zumino, Volkov, Akulov), based on graded Lie algebra, makes such unification not only possible, but a requirement. Supersymmetry is a generalization of a 10-parameter Poincaré group to a 14-parameter super-Poincaré group (Wess, Zumino). It has four additional fermionic generators. Making these fermionic parameters local requires the introduction of gravity. An exciting new development in the 1990s, known as string theory (Polchinski) (which includes local supersymmetry) allows the unification of the standard model with gravity. The most interesting prediction of the superstring theory is that space–time is 10-dimensional, with the extra six dimensions possibly being compact. Below we discuss the formalism of the standard model, grand unification, and supersymmetry in the context of the group theory point of view.

H. Standard Model

1. Quantum Chromodynamics

The standard model (SM) is based on the local gauge symmetry group $\mathrm{SU}(3) \times \mathrm{SU}(2) \times U(1)$. SU(3) corresponds to strong color interactions, while $\mathrm{SU}(2) \times U(1)$ corresponds to electroweak interactions. Let us discuss SU(3) color interactions first. The representations of the quarks and gluons under the SU(3) color are given in Table XI.

The Yang–Mills Lagrangian, for the pure gauge sector involving the gluons only, invariant under the local SU(3) gauge transformations, is given by

$$L_1 = -\tfrac{1}{4} F_{\mu\nu\alpha} F^{\mu\nu\alpha}, \tag{143}$$

where

$$F_{\mu\nu\alpha} = \partial_\mu A_{\nu\alpha} - \partial_\nu A_{\mu\alpha} + g_3 C_{\alpha\beta\gamma} A_{\mu\beta} A_{\nu\gamma}.$$

Here $A_{\mu\alpha}$ are the SU(3) gauge fields belong to the adjoint representation; μ, ν are space–time indices; α, β are the adjoint SU(3) indices (α, β, $\gamma = 1$–8); g_3 is the coupling constant; and the $C_{\alpha\beta\gamma}$ are the structure constants for SU(3) given by

$$[T_\alpha, T_\beta] = i C_{\alpha\beta\gamma} T_\gamma, \tag{144}$$

where the T_α are the SU(3) generators given by Eq. (124) with $T_\alpha = \frac{1}{2}\lambda_\alpha$. The transformation law for the gauge fields $A_{\mu\alpha}$ is given by

$$A^\alpha_\mu \longrightarrow A'^\alpha_\mu = A^\alpha_\mu - \frac{1}{g_3}\partial_\mu \varepsilon^\alpha + C_{\alpha\beta\gamma}\varepsilon^\beta A^\gamma_\mu, \tag{145}$$

where ε^α are the infinitesimal parameters of SU(3).

The Lagrangian describing the interaction of the quarks (q) with the gluons is given by

$$L_2 = \bar{q}(\partial_\mu - i g_3 T_\alpha A_{\mu\alpha}) q + M \bar{q} q, \tag{146}$$

where M is the mass of the quark. The SU(3) transformation law for the triplet quark fields q_i ($i = 1$–3) is given by

$$q_i \longrightarrow (e^{i\varepsilon_\alpha T_\alpha})_{ij} q_j, \tag{147}$$

where ε_α are the local (space–time-dependent) infinitesimal parameters. The sum of the Lagrangians $L_1 + L_2$ describes the complete SU(3) color interactions involving the quarks and the gluons and is known as quantum chromodynamics (QCD). This symmetry is exact, so that the gluons are massless. One very interesting feature of this nonabelian local gauge theory is that the coupling constant decreases logerithmically with the energy scale (this will be discussed in detail later, in the renormalization group equation section). Thus, the coupling constant vanishes at infinite energy, so the theory approaches the behavior of a free field theory as the energy scale approaches infinity. This is known as asymptotic freedom. This logarithmic decrease of the coupling parameter with energy has been tested up to few hundred GeV. On the other hand, as the energy scale decreases, the coupling increases logarithmically so that it becomes very large at low energy and the theory becomes nonperturbative. This is sometimes called infrared slavery, and it has been speculated that this may be responsible for the confinement of the quarks and the gluons.

2. Electroweak Theory

The $\mathrm{SU}(2) \times U(1)$ part of the SM is known as the electroweak theory (Weinberg, Salam, Glashow), since it describes the weak and EM interactions. The multiplet structure of the quarks, leptons, and the electroweak gauge bosons as given in Table XI. The gauge boson part of the Lagrangian under this symmetry is given by

$$L_1 = -\frac{1}{4} F_{\mu\nu\alpha} F^{\mu\nu\alpha} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu}, \tag{148}$$

where

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$$

$$F_{\mu\nu\alpha} = \partial_\mu A_{\nu a} - \partial_\nu A_{\mu a} + g \varepsilon_{abc} A_{\mu b} A_{\nu c}.$$

The transformation laws for the gauge fields B_μ and $A_{\mu a}$ are given by

$$B_\mu \longrightarrow B'_\mu = B_\mu - \frac{1}{g'}\partial_\mu \varepsilon,$$

$$A_{\mu a} \longrightarrow A'_{\mu a} = A_{\mu a} - \frac{1}{g}\partial_\mu \varepsilon_a - \varepsilon_{abc}\varepsilon_b A_{\mu c}.$$

Although the photon is massless, the gauge bosons $W^\pm$ and Z are massive ($M_W = 80$ GeV, $M_Z = 91$ GeV). Thus $SU(2) \times U(1)$ symmetry must be broken to $U_{EM}(1)$ at the 100-GeV scale. In analogy with the spontaneous breaking of the rotational symmetry in the case of ferromagnetism, this symmetry is also spontaneously broken by using suitable scalar fields, the so-called Higgs fields. The Higgs fields required is an SU(2) weak doublet, as shown in Table XI. The Higgs potential is such that although the Lagrangian is invariant under the symmetry group, the ground state (the vacuum state) is not. The required Higgs Lagrangian is

$$L_2 = D^\mu H^\dagger D_\mu H - V(H), \tag{149}$$

where

$$D_\mu = \partial_\mu - igT_i A_{\mu i} - i\frac{g'}{2} Y B_\mu$$

is known as the gauge-covariant derivative, and Y is the $U(1)$ hypercharge.

$$V(H) = -\mu^2 H^\dagger H + \lambda (H^\dagger H)^2 \tag{150}$$

is called the Higgs potential.

Note that for positive values of μ^2 and λ, $V(H)$ has minimum for nonzero values of $\langle H \rangle$, where $\langle H \rangle \equiv \langle 0|H|0\rangle \equiv V$ is the vacuum expectation value of H. Thus, the $SU(2) \times U(1)$ symmetry is broken spontaneously to $U_{EM}(1)$, which is a linear combination of the diagonal part of the SU(2) and the $U(1)$. The charged ($W^\pm$) and the neutral (Z) gauge bosons acquire masses

$$M_w = \frac{1}{2} gV, \qquad M_z = \frac{1}{2}\sqrt{g^2 + g'^2}\, V, \tag{151}$$

while the photon A_μ, corresponding to the unbroken $U_{EM}(1)$ gauge symmetry, remains massless. In terms of the original gauge fields (A_1, A_2, A_3) and B, the expressions for the $W^\pm$, Z, and γ fields are given by Eqs. (141) and (142). The expression for the weak mixing angle θ_w given by

$$\tan\theta_w = \frac{g'}{g}. \tag{152}$$

The Lagrangian for the fermionic part of the gauge interactions is given by

$$L_3 = \bar{\Psi}_L D_{\mu L} \Psi_L + \bar{\Psi}_R D_{\mu R} \Psi_R, \tag{153}$$

where

$$D_{\mu L} = \partial_\mu - igT_i A_{\mu i} - i\frac{g'}{2} Y_L B_\mu,$$

$$D_{\mu R} = \partial_\mu - i\frac{g'}{2} Y_R B_\mu.$$

Here Ψ_L represent left-handed quarks or leptons doublets, while Ψ_R represent right-handed quarks or leptons singlets under the SU(2) transformations given in Table XI.

$$\Psi_L = q_L = \begin{pmatrix} u \\ d \end{pmatrix}_L, \qquad \ell_L = \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L,$$

$$\Psi_R = u_R, d_R, e_R.$$

Finally, the Lagrangian for the Yukawa part of the interaction responsible for giving rise to the masses of the fermions is given by

$$L_4 = f_d \bar{q}_L d_R H + f_u \bar{q}_L U_R \tilde{H} + f_e \bar{\ell}_L e_R H,$$

where

$$\tilde{H} \equiv i\tau_2 H^* \qquad \text{and} \qquad \tau_2 = \begin{pmatrix} o & -i \\ i & o \end{pmatrix}. \tag{154}$$

When the symmetry is broken spontaneously, the fermions acquires masses.

$$m_u = f_u \frac{V}{\sqrt{2}}, \qquad m_d = f_d \frac{V}{\sqrt{2}}, \qquad m_e = f_e \frac{V}{\sqrt{2}}. \tag{155}$$

Thus, in electroweak gauge theory, the gauge boson masses are predicted by the gauge symmetry, while the fermion masses are parametrized in terms of the unknown Yukawa couplings f_u, f_d and f_e.

I. Grand Unification

We saw that the gauge theory based on the SM gauge group, $SU(3) \times SU(2) \times U(1)$, involves three independent coupling constants, g_3, g, and g' (or g_3, g_2, g_1, where $g_2 \equiv g$, $g_1 = \sqrt{5/3}g'$ in a different normalization). The question naturally arises whether the SM gauge group can be embedded into a simple group so that it will involve only one coupling constant. It turns out that there are such unification groups, the simplest being SU(5). (Since the SM gauge group has rank 4, the minimum rank of such a unifying symmetry group must be rank 4.) The unification of the three SM gauge interactions in a simple gauge group is known as the *grand unification theory* (GUT).

J. SU(5) Grand Unification

The SU(5) grand unification theory can accommodate the chiral nature of the fermion representation under $SU(2) \times U(1)$ as well as the absence of chiral anomalies

without introducing new fermions. One can also calculate the value of the weak mixing angle. The gauge bosons belong to the adjoint representation of SU(5). The 24 gauge bosons have the following decomposition under the SM, $SU(3) \times SU(2) \times U(1)$ representations:

$$24 = (8, 1, 0) + (1, 3, 0) + \left(3, 2, \tfrac{5}{3}\right) + \left(3^*, 2, -\tfrac{5}{3}\right) + (1, 1, 0),$$

where (8, 1, 0) represents the eight gluons, (1, 3, 0) represents the three SU(2) gauge bosons, and (1, 1, 0) represents the $U(1)$ gauge bosons. The particles in $(3, 2, \frac{5}{3})$ and $(3^*, 2, -\frac{5}{3})$ represent 12 new gauge bosons that are not present in the SM. These are denoted by X_α, Y_α, the (X, Y) being the doublet under the weak SU(2) and the triplet under the color SU(3) ($\alpha = 1, 2, 3$). The six gauge bosons in $(3^*, 2, -\frac{5}{3})$ are the complex conjugate fields of (X, Y). There are 24 generators, λ_A ($A =$ 1–24) of SU(5), where λ_α ($\alpha =$ 1–8) are the 5×5 matrix generalization of those given in Eq. (124). $\lambda_{20+\alpha} = 1, 2, 3$ corresponds to the generalization of three SU(2) generators, while λ_{24} corresponds to that of the hypercharge $U(1)$ generator. The generators $\lambda_A (A =$ 9–20) are the new ones that are not present in the SM gauge group. The corresponding gauge bosons X_α, Y_α can change quarks into leptons and thus violate baryon and lepton number conservation. This is a new feature of all grand unified theories.

The fermions belong to the SU(5) representations 5^* and 10, which under $SU(3) \times SU(2) \times U(1)$ representations decompose as follows:

$$5^* = \left(3^*, 1, \tfrac{2}{3}\right) + (1, 2, -1),$$
$$10 = \left(3^*, 1, -\tfrac{4}{3}\right) + \left(3, 2, \tfrac{1}{3}\right) + (1, 1, 2).$$

Expressing all fermions as left-handed Weyl fermions, $(3^*, 1, \frac{2}{3})$ contains the three color singlet SU(2) doublet-down-type antiquarks (d_1^c, d_2^c, d_3^c), $(1, 2, -1)$ contains the color singlet SU(2) doublet electron-type neutrino and the electron (ν_e, e), $(3, 2, \frac{1}{3})$ contains the color triplet SU(2) doublet-up and- down quarks, $(3^*, 1, -\frac{4}{3})$ contains the color triplet SU(2) singlet up-type antiquarks, and (1, 1, 2) contains both the SU(3) and SU(2) singlet positrons. Note that all the known fermions in one family fit neatly, and no extra fermions are needed.

K. Symmetry Breaking in SU(5)

SU(5) symmetry must be broken spontaneously at a very high energy scale ($\sim 10^{16}$GeV) because X, Y gauge bosons have baryon number-violating interactions, and will cause proton decay. The agreement with the current experimental lifetime of the proton demand $M_{x,y} \geq 10^{16}$ GeV. There are two stages of symmetry breaking:

$$\mathrm{SU}(5) \xrightarrow[\text{first stage}]{\mu \sim 10^{16}\mathrm{GeV}} \mathrm{SU}(3) \times \mathrm{SU}(2) \times U(1) \xrightarrow[\text{second stage}]{\mu \sim 10^{2}\mathrm{GeV}} \mathrm{SU}(3) \times U_{\mathrm{EM}}(1).$$

The first stage of symmetry breaking can be achieved using a 24-dimensional Higgs representation of SU(5), while the second stage of symmetry breaking can be achieved by using a 5-dimensional Higgs representation of SU(5). The gauge bosons X, Y acquire masses after the first symmetry breaking, while $W^\pm$, Z get their masses at the second symmetry breaking. Fermions also receive their masses from the second stage of symmetry breaking.

L. Renormalization Group Equations and Evolutions of Gauge Couplings

In grand unified theories (GUTs) with a simple gauge group G such as SU(5), we have only one coupling constant, g_G. Thus, at the energy scale M_{GUT} above which the symmetry is exact, the three coupling g_3, g_2, g_1 are equal. (Note that we are using a normalization for which $g_2 \equiv g$, $g_1 = \sqrt{5/3}g'$). The reason for the values of the three coupling being so different at low energy is due to the fact that they evolve in energy differently, making the effective couplings g_2, and g_1 much smaller than the strong coupling g_3. The renormalization group evolution of a coupling g_i is given by

$$\mu \frac{dg_i}{\partial \mu} = \beta_i(\{g\}), \tag{156}$$

Where $\beta_i(\{g\})$ is the beta-function for the coupling g_i. At the one-loop level, for the three SM gauge couplings, $\beta_i(\{g_i(\mu)\}) = b_i g_i^3(\mu)$ for $g_i \langle\langle 1$. The boundary condition is

$$g_1 = g_2 = g_3 = g_G \qquad \text{at} \quad \mu = M_{\mathrm{GUT}}. \tag{157}$$

For the SM,

$$b_3 = -\frac{1}{16\pi^2}\left(11 - \frac{2}{3}N_f\right),$$
$$b_2 = -\frac{1}{16\pi^2}\left(\frac{22}{3} - \frac{2}{3}N_f - \frac{1}{6}N_H\right), \tag{158}$$
$$b_1 = \frac{1}{16\pi^2}\left(\frac{2}{3}N_f + \frac{1}{10}N_H\right),$$

where N_f is the number of quark flavors ($N_f = 6$ for six quark flavors, u, d, c, s, t, b) and N_H is the number of Higgs doublets.

The experimental values of the three coupling at the low energy scale, such as at $\mu = M_Z$ is very well determined from the studies of Z-boson decay:

$$\alpha_3^{-1}(M_Z) = 8.5 \pm 0.5, \qquad \alpha_2^{-1}(M_Z) = 29.61 \pm 0.13,$$
$$\alpha_1^{-1}(M_Z) = 98.29 \pm 0.13, \tag{159}$$

where $\alpha_i(\mu) \equiv g_i^2(\mu)/4\pi$. Using these values at $\mu = M_Z$ and evolving the couplings to higher energy scale using Eqs. (156) and (157), we find that the couplings do not unify for this simple minimal SU(5) GUT. Also, the very approximate unification scale is too small, $M_X \sim 10^{13}$ GeV, so the dominant proton decay rate in this model, $P \rightarrow e^+\pi^o$ is too high compared to the experimental limit. Thus, this simple SU(5) GUT theory is ruled out experimentally. However, for supersymmetric SU(5) GUT, the unification works beautifully and the unification scale is $\sim 10^{16}$ GeV, in complete agreement with the current experimental proton lifetime limits. So, next we turn to supersymmetry, supersymmetric SM (also known as MSSM), and supersymmetric GUTS.

M. Supersymmetry

Supersymmetry is a beautiful mathematical generalization of 10-parameter Poincaré group to 14-parameter super-Poincaré or supersymmetry groups. The 10-parameter Poincaré group, with the generators $J_{\mu\nu}$ and P_μ ($\mu, \nu = 0, 1, 2, 3$), satisfy the algebra.

$$[J_{\mu\nu}, J_{p\sigma}] = g_{\mu\rho}J_{\nu\sigma} - g_{\nu\rho}J_{\mu\sigma} + g_{\mu\sigma}J_{\nu\rho} - g_{\nu\sigma}J_{\mu\rho}, \tag{160}$$
$$[P_\mu, P_\nu] = 0, \qquad [J_{\mu\nu}, P_\rho] = g_{\mu\rho}P_\nu - g_{\nu\rho}P_\mu.$$

Wess and Zumino (1974) discovered that this Poincaré algebra beautifully generalizes to a new symmetry algebra by introducing four new generators, S_α. S_α's are fermionic, the index α ($\alpha = 1, 2, 3, 4$) is a spinor index, and the corresponding parameters ε_α are anticommuting Majorana spinors. S_α's satisfy the following algebra:

$$[S_\alpha, P_\mu] = 0, \qquad [J_{\mu\nu}, S_\alpha] = \frac{1}{2}(\sigma_{\mu\nu})_{\alpha\beta}S_\beta, \tag{161}$$
$$\{S_\alpha, S_\beta\} = (\gamma^\mu c)_{\alpha\beta}P_\mu,$$

where c is the Dirac charge conjugation matrix.

In Eq. (161), the curly braces represent anticommutators. The introduction of anticommuting parameters and the associated fermonic generators, together with the bosonic ones, brought lot of excitement among the mathematicians. The usual Lie algebra has now been generalized to include both commutators and anticommutators and is known as graded Lie algebra. Graded Lie algebras have now been classified. The representation of the super Poincaré algebra, Eqs. (160) and (161), involves both bosons and fermions in the same representation and is called Fermi–Bose symmetry or supersymmetry. In a given representation, we have particles of both spin j and $j - \frac{1}{2}$. The simplest irreducible representation is called the chiral scalar superfield, in which we have a complex scalar field and a chiral fermionic field. Next is the vector representation, which has a massless vector field and a Majorana spinor field. Thus, in the SM, each representation will be associated with its superpartner. The particle content of the minimal supersymmetric extension of the SM is given in Table XII. The first column represent the usual SM particles, and the second column their supersymmetric partners. The intrinsic spins of the superpartners differ by half a unit. For example, $\tilde{e}_L$ has spin 0, a scalar particle. Note that an additional new feature is that that two Higgs doublets are needed to cancel chiral anomalies.

Since none of the superpartners has so far been observed, supersymmetry must be broken at a few hundred GeV scale or higher. Two popular supersymmetry breaking mechanisms are the gravity mediated and the gauge mediated (GMSB). The superpartners are expected to a have masses less than TEV for the supersymmetry to solve the gauge hierarchy problem, and are expected to be discovered at the Large Hadron Collider (LHC), if not at the upgraded Tevatron. Local supersymmetry, in which the fermionic parameters ε_α are arbitrary functions of space–time (x, t), necessarily needs the introduction of the gravity supermultiplet containing the spin-2 graviton and its supersymmetric partner, the gravitino.

TABLE XII Particle Contents of the Supersymmetric Standard Model

Particles	Superpartners	SU(3)× SU(2)×U(1) representations
$u_{L\alpha}$	$\tilde{u}_{L\alpha}$	$(3, 2, \frac{1}{3})$
$d_{L\alpha}$	$\tilde{d}_{L\alpha}$	$(3, 2, \frac{1}{3})$
ν_{eL}	$\tilde{\nu}_{eL}$	$(1, 2, -1)$
e_L	$\tilde{e}_L$	$(1, 2, -1)$
$u_{R\alpha}$	$\tilde{u}_{R\alpha}$	$(3, 1, \frac{4}{3})$
$d_{R\alpha}$	$\tilde{d}_{R\alpha}$	$(3, 1, -\frac{2}{3})$
e_R	$\tilde{e}_R$	$(1, 1, -2)$
g_a	$\tilde{g}_a$	$(8, 1, 0)$
A_a	$\tilde{A}_a$	$\begin{pmatrix} +1 \\ 1, 3, \ 0 \\ -1 \end{pmatrix}$
B	$\tilde{B}$	$(1, 1, 0)$
H_1	$\tilde{H}_1$	$(1, 2, 1)$
H_2	$\tilde{H}_2$	$(1, 2, -1)$

In supersymmetric SM, there will be additional contributions to the evolution of the gauge coupling above the supersymmetric thresholds. Thus, above threshold, the coefficients to the beta-factors b_1, b_2, and b_3 are modified to

$$\begin{aligned}\tilde{b}_3 &= -\frac{1}{16\pi^2}(9 - N_f),\\ \tilde{b}_2 &= -\frac{1}{16\pi^2}\left(6 - N_f - \frac{1}{2}N_H\right),\\ \tilde{b}_1 &= \frac{1}{16\pi^2}\left(N_f + \frac{3}{10}N_H\right).\end{aligned} \tag{162}$$

Starting with the experimental values of $\alpha_1(M_Z)$, $\alpha_2(M_Z)$, and $\alpha_3(M_Z)$ given in Eq. (162), and evolving these to higher energy scales, these coupling unify at an energy scales, $\mu = M_{\text{GUT}} \simeq 2 \times 10^{16}$ GeV. Thus, although in non-SUSY SU(5) the unification does not take place, unification does occur very accurately in SUSY SU(5). Also, because of the very high scale of the unifications, the proton decay limits are easily satisfied.

IX. APPLICATIONS IN GEOMETRICAL OPTICS

A nondegenerate, skew-symmetric bilinear form in $2n$ ($n = 2$, for example) dimensions is usually written as

$$\sum_{i=1}^{4}\sum_{k=1}^{4} a_{ik}, x_i y_k, \qquad a_{ik} = -a_{ki}. \tag{163}$$

In matrix form this can be expressed as $\mathbf{X}^T\mathbf{A}\mathbf{Y}$, where x_i and y_k are treated as components of the Vectors $\vec{\mathrm{X}}$ and $\vec{\mathrm{Y}}$, respectively, and the matrix $\mathbf{A}$ satisfies the relation

$$\mathbf{A}^T = -\mathbf{A}, \tag{164}$$

the superscript T meaning the transposition operation. We assume that the variables as well as the coefficients are real. By suitable transformations the above bilinear form can be reduced to "canonical form,"

$$\sum_{i=1}^{\nu=2}\left(\xi_i\eta_{i+\nu} - \xi_{i+\nu}\eta_i\right) = \vec{\xi}^{\,T}\mathbf{K}\vec{\eta} \tag{165}$$

Here $\vec{\xi}$ is the vector $(\xi_1, \xi_2, \xi_3, \xi_4)$, $\vec{\eta}$ the vector $(\eta_1, \eta_2, \eta_3, \eta_4)$, and $\mathbf{K}$ the matrix

$$\mathbf{K} = \begin{pmatrix} 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ -1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\end{pmatrix} \equiv \begin{pmatrix} 0 & 1\\ -1 & 0\end{pmatrix}. \tag{166}$$

The set of all matrices $\mathbf{S}$ that leave this skew-symmetric form invariant constitute a Lie group called the *symplectic group* Sp($2n$) (here $2n$ is 4), and this is a subgroup of the general linear group GL($2n$). The requirement of skew symmetry implies that the general element of this group of transformations $\mathbf{S}$ should satisfy

$$\mathbf{S}^T\mathbf{K}\mathbf{S} = \mathbf{K}. \tag{167}$$

In geometric optics a single ray incident on a refracting surface is usually defined by one of its points in object space, given by a vector $\vec{\mathrm{a}}$ and a vector $\vec{\mathrm{s}}$ which gives the direction of the ray. The length of $\vec{\mathrm{s}}$ is n, the refractive index of the medium in that space.

In image space the corresponding vectors describing the refracted ray are $\vec{\mathrm{a}}'$ and $\vec{\mathrm{s}}'$, n' being the appropriate refractive index there. A two-dimensional manifold of such rays is adequately described by the components of the vector expressed as continuous functions of two variable parameters u and v. It has been proven by Herzberger that a fundamental invariant governing image formation is $(\vec{a}_u \cdot \vec{s}_v) - (\vec{a}_v \cdot \vec{s}_u)$. In other words,

$$(\vec{a}_u' \cdot \vec{s}_v') - (\vec{a}_v' \cdot \vec{s}_u') = (\vec{a}_u \cdot \vec{s}_v) - (\vec{a}_v \cdot \vec{s}_u), \tag{168}$$

equivalent to the Lagrange bracket in analytical mechanics. Here $\vec{a}_u$ is a notation for the partial derivative $\partial\vec{a}/\partial u$. It turns out that the invariant is valid for any point on the ray, and as such it is convenient to refer to the intersection of the ray with a plane $Z = 0$, so that the components of $\vec{a}$ are (x, y). Let the components of $\vec{s}$ be $(\xi, \eta, \sqrt{n^2 - \xi^2 - \eta^2})$. Taking the pair of parameters u, v to be (x, y), (x, ξ), (x, η), (y, ξ), (y, η), and (ξ, η) [every pair of the four variables (x, y)], in turn, the above invariant relation yields the following six equations for the components of the vectors:

$$\begin{aligned}
x_x'\xi_y' + y_x'\eta_y' - x_y'\xi_x' - y_y'\eta_x' &= 0\\
x_x'\xi_\xi' + y_x'\eta_\xi' - x_\xi'\xi_x' - y_\xi'\eta_x' &= 1\\
x_x'\xi_\eta' + y_x'\eta_\eta' - x_\eta'\xi_x' - y_\eta'\eta_x' &= 0\\
x_y'\xi_\xi' + y_y'\eta_\xi' - x_\xi'\xi_y' - y_\xi'\eta_y' &= 0\\
x_y'\xi_\eta' + y_y'\eta_\eta' - x_\eta'\xi_y' - y_\eta'\eta_y' &= 1\\
x_\xi'\xi_\eta' + y_\xi'\eta_\eta' - x_\eta'\xi_\xi' - y_\eta'\eta_\xi' &= 0
\end{aligned} \tag{169}$$

This set of equations can be expressed in matrix form as

$$\mathbf{M}^T\mathbf{K}\mathbf{M} = \mathbf{K} \tag{170}$$

where $\mathbf{K}$ is given in Eq. (166). The matrix $\mathbf{M}$ now becomes

$$\mathbf{M} = \begin{pmatrix} x_x' & x_y' & x_\xi' & x_\eta'\\ y_x' & y_y' & y_\xi' & y_\eta'\\ \xi_x' & \xi_y' & \xi_\xi' & \xi_\eta'\\ \eta_x' & \eta_y' & \eta_\xi' & \eta_\eta'\end{pmatrix}$$

recalling that $y_\xi' \equiv \partial y'/\partial\xi$, etc. Thus we see that Eq. (170), sometimes called the "lens equation," is identical to

Eq. (167), and the latter describes the symplectic group Sp(4). Thus, the symplectic group is of fundamental importance in geometric optics.

An interesting application of the techniques based on the symplectic Lie group has been made to charged-particle beam optics, which is of importance in particle accelerator physics. The article in *Annual Review of Nuclear and Particle Science* cited in the Bibliography (Dragt *et al.*, 1988) gives an illuminating discussion of this application.

X. THE RENORMALIZATION GROUP

The *renormalization group* (RG) theory has applications in several areas of physics but we shall confine our treatment here to "critical phenomena". The importance of RG theory in elementary particle physics is discussed elsewhere in this article.

The concept of renormalization has its origins in quantum electrodynamics, which is the theory that explains the properties and behavior of electrons and photons or light quanta. According to quantum field theory, particles are the outcome of quantizing appropriate wave fields. For instance, photons are quanta resulting from quantizing the classical electromagnetic field of Maxwell. A disturbing feature of quantum electrodynamics has been the existence of infinities or divergent integrals. For instance, if we assume the charge of an electron e to be uniformly distributed in a sphere of radius a, the electrostatic self-energy of this sphere is $\frac{3}{5}(e^2/a)$ according to Maxwell's theory, and this becomes infinite in the limit of a going to zero. In other words, a point electron has infinite self-energy. Many such divergences exist, of which two important cases are the self-energy of the photon and the polarizability of the vacuum induced by an external electric field. This vacuum happens to be an infinite sea of occupied electron states of negative energy which are the inevitable consequences of Dirac's relativistic electron theory. According to Oppenheimer, the roots of charge renormalization lie in the efforts to overcome these infinities by asserting that the experimentally measured charge of the electron is the sum of the "true" and "induced" charges, and it is this prescription that is the genesis of renormalization of charge. The existence of a similar distinction between "bare electron mass" and "dressed electron mass" has been pointed out by Kramers in his study of the interaction between charged particles and the radiation field.

These concepts took concrete shape ten years later in the epoch-making contributions of Schwinger, Tomonoga, and Feynman, and the term *renormalization* has been in vogue ever since. According to Yukawa, the forces between particles are mediated by quanta and, in particular, the forces between electrons are mediated by light quanta or photons. A popular picture of this electron can thus be that of a bare particle modestly apparelled in photons. It is important to note that underlying all these ideas is the mass–energy equivalence of Einstein's special relativity theory.

As a preliminary to the introduction of the renormalization group, it is helpful to discuss certain ideas of "critical phenomena", to which it has its important application. We shall single out phase transitions in liquids for this purpose. It is common knowledge that under appropriate conditions of temperature and pressure, water exists in the three phases of solid (ice), liquid (water), and gas (steam). In a three-dimensional plot of pressure, density, and temperature, the domains of these phases and their boundaries are clearly demarcated. Across the boundary separating gas from a liquid, for instance, there is a discontinuity in density. In his experiments on the liquefaction of carbon dioxide gas by isothermal compression, Andrews noticed in 1869 the existence of a critical temperature T_c, above which there is a continuity of the gas and liquid phases. In other words, it is only below the critical temperature that condensation of a gas can take place. The pressure at this point is called the critical pressure P_c, and likewise the density ρ_c. As the temperature is increased to T_c, the density of the gas phase, ρ_g, tends to equal the density of the liquid phase, ρ_l, and the two become one ρ_c at the critical temperature T_c. For water, for instance, T_c is 648, P_c is 218, and ρ_c is 0.25 in appropriate units.

A classical theoretical foundation for the behavior of a gas at equilibrium temperature T is the equation of state relating the thermodynamic variables pressure p, volume (or density ρ) V, and temperature T. Taking into account possible intermolecular forces and the finite size of molecules, van der Waals derived in 1873 the equation of state

$$\left(p + \frac{a}{V^2}\right)(V - b) = NkT, \tag{171}$$

where a and b are constants pertaining to the gas, k is the universal Boltzmann constant, and N is the Avogadro number. This equation does predict the existence of critical constants according to the following relations:

$$\begin{aligned} &\frac{\partial p}{\partial V} = 0, \qquad \frac{\partial^2 p}{\partial V^2} = 0, \\ &p_c = \frac{a}{27b^2}, \qquad V_c = 3b, \qquad T_c = \frac{8a}{27bNk}. \end{aligned} \tag{172}$$

In terms of reduced variables $\rho = p/p_c$, $\nu = V/V_c$, $t = T/T_c$, the van der Waals equation can be written in the so-called universal form applicable to any gas:

$$\left(p + \frac{3}{\nu^2}\right)\left(\nu - \frac{1}{3}\right) = \frac{8}{3}t. \tag{173}$$

This is known as the *law of corresponding states*. The existence of such a universal relation, through not the exact quadratic equation, has been demonstrated by Guggenheim, who plotted the reduced density (ρ/ρ_c) as a function of the reduced temperature (T/T_c) for eight different fluids on which experimental measurements have been made. This has given rise to the suggestion that the *order parameter* $\rho_l - \rho_g$ satisfies a power law in the reduced temperature,

$$\rho_l - \rho_g \sim |\tau|^\beta, \qquad \tau = \frac{T}{T_c} - 1 = t - 1, \tag{174}$$

and β has since come to be known as a critical exponent. The other critical exponents are γ, related to the isothermal compressibility,

$$\kappa_T = -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_T = \frac{1}{\rho}\left(\frac{\partial \rho}{\partial p}\right)_T = |\tau|^{-\gamma}, \tag{175}$$

and α, related to specific heat at constant volume,

$$C_v = \left(\frac{\partial V}{\partial T}\right)_V = |\tau|^{-\alpha}. \tag{176}$$

From thermodynamic reasoning it has been shown that the exponents are not all independent but are coupled by a relation

$$\alpha + 2\beta + \gamma = 2. \tag{177}$$

One of the achievements of the RG theory has been the derivation of these critical exponents.

The concept of a renormalization group, like the concept of renormalization itself, originated in quantum electrodynamics. Stueckelberg and Petermann introduced this is quantum field theoretic language and worked out its details including infinitesimal transformations and their Lie algebra. However, this found little application and was almost ignored. The foundation for the currently used RG is the work of Gell-Mann and Low, also in the context of quantum electrodynamics. Recognizing this and making a systematic development of RG theory and its adaptability to critical phenomena analysis has been the indefatigable effort of Wilson.

The RG is a Lie group of transformations which are functions of certain variables and parameters that vary continuously over a given domain. The fundamental group postulate of multiplication here means that a transformation followed by a transformation is still another transformation with different values of the parameters. To illustrate this in one dimension let us assume x_i variable point on the X axis and s_j a parameter. The transformation with parameter s_i takes x_j to x_k on the line

$$f(-s_i, x_j)x_j = x_k. \tag{178}$$

A second transformation involving a parameter s_j takes x_k to x_l. A succession of these two transformations then moves x_j to x_l.

$$\{f(-s_j, x_k)x_k\}\{(-s_i, x_j)x_j\} = x_e. \tag{179}$$

However, there must also exist a transformation, say with parameter s_k, which takes x_j directly to x_l.

$$f(-s_k, x_j)x_j = x_e. \tag{180}$$

In other words, we have the group multiplication

$$\{f(-s_j, x_k)x_k\}\{f(-s_i, x_j)x_j\} = f(-s_k, x_j)x_j = x_e. \tag{181}$$

We assume this can happen provided the parameters satisfy the relation

$$s_j s_i = s_k. \tag{182}$$

For a certain value of the parameters, say $s = 1$, the transformation must be an identity and this means that

$$f(-1, x) = 1 \tag{183}$$

for any point x. It is generally agreed that the RG is really a semigroup, because the inverse transformation cannot always be uniquely defined. To be applicable to actual physical situations these transformation must be such that the transformed system describes the same physics as the original one. In statistical mechanics, for instance, the partition function will be the same in different systems. In quantum mechanics the original and transformed Hamiltonians relate to the same dynamical system.

The importance of the RG transformation is that it is iterative:

$$f(\mathbf{H}) = \mathbf{H}^1, \qquad f(\mathbf{H}^1) = \mathbf{H}^{11}, \text{ etc.} \tag{184}$$

In one dimension the functional equation is like a recursion relation,

$$f(x) = x^1, \qquad f(x^1) = x^{11}, \ldots, \tag{185}$$

and this implies that after very many iterations one can hope to approach a fixed point x^*,

$$f(x^*) = x^*. \tag{186}$$

As in the case of the roots of an algebraic equation, there can be several fixed points. Let us assume we have at a fixed point x^*,

$$\begin{aligned} x &= x^* + \delta x, \\ x^1 &= x^* + \delta x^1. \end{aligned} \tag{187}$$

In the vicinity of the fixed point a Taylor expansion gives

$$f(x^* + \delta x) = f(x^*) + \frac{\partial f}{\partial x^*}\delta x + \cdots = x^* + \delta x^1,$$

or

$$\delta x^1 = \frac{\partial f}{\partial x^*}\delta x + \text{higher powers of } \delta x, \tag{188}$$

We thus have, to a linear approximation in δx,

$$\delta x^1 = \frac{\partial f}{\partial x^*}\,\delta x \tag{189}$$

In higher dimensions the variables can be regarded as components of a vector and Eq. (154) now generalizes to

$$f(\vec{\mathrm{r}}) = \vec{\mathrm{r}}^{\mathrm{I}}, \qquad f\left(\vec{\mathrm{r}}^{\mathrm{I}}\right) = \vec{\mathrm{r}}^{\mathrm{II}}, \ldots, \qquad f(\vec{\mathrm{r}}^*) = \vec{\mathrm{r}}^*. \tag{190}$$

The linear approximation above now becomes

$$\delta\vec{\mathrm{r}}^{\,l} = \mathbf{M}(\vec{\mathrm{r}}^*)\delta\vec{\mathrm{r}}, \tag{191}$$

where the matrix elements of **M** are evaluated at the fixed point r^*. We presume **M** satisfies the eigenvalue equation

$$\mathbf{M}\,\vec{\xi}_i = \lambda_i\vec{\xi}_i. \tag{192}$$

The behavior of $\vec{\mathrm{r}}$ after several iterations is governed by the magnitude of the eigenvalue λ. If $|\lambda|\langle 1$, under iterations the higher powers of λ will then be progressively smaller and Eq. (155) shows that x will converge to (attracted to) the fixed point x^*, and for this reason this is said to be an "irrelevant" eigenvalue. On the other hand, when $|\lambda|\rangle 1$ the point x will drift farther and farther away from the fixed point. This is then a "relevant" eigenvalue. We will now illustrate this by a hypothetical example in two dimensions.

In component form let us assume the recursion equations

$$x^1 = \frac{2x}{1+x^3}, \qquad y^1 = \frac{y^3}{1+x^3}. \tag{193}$$

The fixed points (x^*, y^*), obtained by solving for $x^1 = x$, $y^1 = y$, are (0, 0), (0, 1), $(1,\sqrt{2})$, $(1, -\sqrt{2})$. We will now set

$$x + \delta x^1 = \frac{2(x+\delta x)}{1+(x+\delta x)^3}, \quad y + \delta y^1 = \frac{(y+\delta y)^3}{1+(x+\delta x)^3}. \tag{194}$$

In the linear approximation the matrix **M** of Eq. (192) is

$$\begin{pmatrix}\delta x^1\\ \delta y^1\end{pmatrix} = \begin{pmatrix}\dfrac{2-4x^3}{(1+x^3)^2} & 0\\ \dfrac{-3x^2y^3}{(1+x^3)^2} & \dfrac{3y^2}{(1+x^3)}\end{pmatrix}\begin{pmatrix}\delta x\\ \delta y\end{pmatrix}. \tag{195}$$

It is easy to see that the eigenvalues of this matrix are

$$\lambda_1 = \frac{2-4x^3}{(1+x^3)^2}, \qquad \lambda_2 = \frac{3y^2}{1+x^3}. \tag{196}$$

At the fixed point (0, 0), $\lambda_1 = 2, \lambda_2 = 0$, and we have one relevant eigenvalue and one marginal. At the fixed point (0, 1), $\lambda_1 = 2, \lambda_2 = 3$ and both are relevant. At $(1, \sqrt{2})$, $\lambda_1 = -\frac{1}{2}, \lambda_2 = 3$, and so one is irrelevant and one relevant. At $(1, -\sqrt{2})$, $\lambda_1 = -\frac{1}{2}, \lambda_2 = 3$ and this is the same case as the fixed point $(1, \sqrt{2})$. The marginal case can be decided only when one calculates beyond the linear approximation. As mentioned earlier, relevant parameters are those that tend to make the system unstable to small disturbances.

The RG transformations in practice are designed to reduce the degrees of freedom of the system from, say, N to N', and this involves a scale change of the vectors $\vec{\mathrm{r}}$ in coordinate space by the factor b,

$$b^d = (N/N'), \qquad \vec{\mathrm{r}}' = \left(\frac{\vec{\mathrm{r}}}{b}\right), \tag{197}$$

where d relates to its dimensionality. This makes the eigenvalues λ_i functions of the scale factor b and the iterative nature of the transformation makes

$$\lambda_i = b^{\eta_i}, \tag{198}$$

where η_i are independent of the scale factor. The use of this in thermodynamic functions like free energy eventually yields the critical exponents. A remarkable similarity in the values of these exponents in seemingly different phase transitions led to the hypothesis of "universality," which states that all phase-transition problems belong to a few classes determined by certain parameters such as dimensionality and the order. This is confirmed by RG theory, according to which these universality classes are determined by the different ranges of attraction of the fixed point.

ACKNOWLEDGMENT

We would like to dedicate this to the late Prof. Mark A. Samuel, who was a co-author on the first two editions of this article.

SEE ALSO THE FOLLOWING ARTICLES

ATOMIC PHYSICS • FIELD THEORY AND THE STANDARD MODEL • GROUP THEORY • NUCLEAR PHYSICS • QUANTUM CHROMODYNAMICS • QUANTUM MECHANICS • SET THEORY • PARTICLE PHYSICS, ELEMENTARY • SUPERSTRING THEORY

BIBLIOGRAPHY

Arnison, G., *et al.* (1983–1984). "Experimental observation of isolated large transverse energy electrons with associated missing energy $s = 540$ Gev," *Phys. Lett.* **122B,** 103 (1983). "Experimental observation of Lepton pairs of invariant mass around 95 Gev/c^2 at the CERN

collider," *Phys. Lett*. **126B**, 398 (1983). "Further evidence for charged intermediate vector bosons at the SPS collider," *Phys. Lett*. **129B,** 273 (1984). "Observation of the muonic decay of the charged intermediate vector boson," *Phys. Lett*. **134B,** 469 (1984). "Observation of muonic Z^o decay at the PP collider," *Phys. Lett*. **147B,** 241 (1984).

Campbell, J. E. (1966). "Introductory Treatise on Lie's theory of Finite Continuous Transformation Groups," Chelsea, New York.

Chang, T. P., and Li, L. F. (1984). "Gauge Theory of Elementary Particle Physics," Oxford University (Clarendon) Press, London/New York.

Cotton, F. A. (1970). "Chemical Applications of Group Theory," Wiley (Interscience), New York.

Domb, C. (1996). "The Critical Point," Taylor & Francis, London (all other references cited in the text can be found in this book).

Dragt, A. J., *et al*. (1988). "Annual Review of Nuclear & Particle Science," p. 455, Annual Reviews, Palo Alto, CA.

Englert, F., and Brout, R. (1964). "Broken symmetry and the mass of gauge vector mesons," *Phys. Rev. Lett*. **13,** 321.

Fock, V. (1927). Uber die invariante Form der Wellen- und der Beugungs gleichungen fur einen geladenen Massenpunkt. *Z. Physik* **39,** 226.

Gasiorowicz, S. (1966). "Elementary Particle Physics," Wiley, New York.

Gell-Mann, M., and Neeman, Y. (1964). "The Eight-Fold Way," Benjamin, New York.

Georgi, H. (1982). "Lie Algebras in Particle Physics," Frontiers in Physics, Benjamin/Cummings, New York.

Georgi, H., and Glashow, S. L. (1974). "Unity of all elementary particle forces," *Phys. Rev. Lett*. **32,** 438.

Glashow, S. L. (1961). "Particle symmetries of weak interactions," *Nuclear Phys*. **22,** 579.

Gross, D. J., and Wilczek, F. (1973). "Ultra violet behavior of non-Abelian gauge theories," *Phys. Rev. Lett*. **30,** 1343.

Guralnik, G. S., Hagen, C. R., and Kibble, T. W. (1964). "Global conservation laws and massless particles," *Phys. Rev. Lett*. **13,** 585.

Halzen, F., and Martin, A. D. (1984). "Quarks and Leptons," Wiley, New York.

Hamermesh, M. (1959). "Group Theory," Addison-Wesley, Reading, MA.

Herzberg, G. (1959). "Infrared and Raman Spectra," Van Nostrand, Princeton, NJ.

Herzberger, M. (1958). "Modern Geometrical Optics," Interscience, New York.

Higgs, P. W. (1964). "Broken Symmetries and the Masses of Gauge Bosons," *Phys. Rev. Lett*. **13,** 508.

Kibble, T. W. B. (1967). "Symmetry breaking in non-Abelian gauge theories," *Phys. Rev*. **155,** 1554.

Kokkedee, J. J. J. (1969). "The Quark Model," Benjamin, New York.

Lichtenberg, D. B. (1970). "Unitary Symmetry and Elementary Particles," Academic Press, New York.

Lie, S. (1893). "Theorie der Transformationgruppen," F. Engel, Leipzig.

Lipkin, H. J. (1965). "Lie Groups for Pedestrians," North-Holland, Amsterdam.

Luneburg, R. K. (1964). "Mathematical Theory of Optics," University of California Press, Los Angeles.

Miller, A. (1994). "Early Quantum Electrodynamics," Cambridge University Press, Cambridge, U.K.

Pati, J. C., and Salam, A. (1973). "Is baryon number conserved?" *Phys. Rev. Lett*. **31,** 661.

Polchinski, J. (1998). "String Theory," Vols. I and II, Cambridge University Press, Cambridge, U.K.

Politzer, H. D. (1973). "Reliable perturbative results for strong interactions," *Phys. Rev. Lett*. **30,** 1346.

Salam, A. (1968). "Elementary Particle Physics," p. 367, Almquist and Wiksells, Stockholm, Sweden.

Schiff, L. I. (1968). "Quantum Mechanics," McGraw-Hill, New York.

Schwinger, J. (ed). (1958). "Quantum Electrodynamics," Dover, New York. The original papers can be found in the books of Miller and Schwinger.

Segré, E. (1977). "Nuclei and Particles," Benjamin, New York.

Slater, J. C. (1972). "Symmetry and Energy Bands in Crystals," Dover, New York.

Stavroudis, O. N. (1972). "The Optics of Rays, Wavefronts and Caustics," Academic Press, New York.

Swamy, N. V. V. J., and Samuel, M. A. (1979). "Group Theory Made Easy for Scientists & Engineers," Wiley-Interscience, New York.

't Hooft, G. (1992). "Conference on Lagrangian Field Theory," Marseille.

Tinkham, M. (1975). "Group Theory and Quantum Mechanics," McGraw-Hill, New York.

Volkov, D. V., and Akulov, V. P. (1972). "Universal neutrino interaction," *JETP Lett*. **16,** 438.

Weinberg, S. (1967). "A model of leptons," *Phys. Rev. Lett*. **19,** 1264.

Wess, J., and Zumino, B. (1974). "Supergauge transformation in four dimensions," *Nuclear Phys*. **B70,** 39.

Weyl, H. (1929). "Elektron und Gravitation," *Z. Physik* **56,** 330.

Wigner, E. P. (1959). "Group Theory and Its Application to the Quantum Mechanics of Atomic Spectra," Academic Press, New York.

Wilson, R. G. (1971). "Renormalization group and strong interactions," *Phys. Rev*. **D3,** 1818.

Yang, C. N., and Mills, R. L. (1954). "Conservation of isotopic spin and isotopic gauge invariance," *Phys. Rev*. **96,** 191.

Yeomans, J. M. (1992). "Statistical Mechanics of Phase Transitions," Oxford University Press, New York.



Integral Equations

Ram P. Kanwal
Pennsylvania State University

I. Definitions, Classification, and Notation
II. The Method of Successive Approximations
III. The Fredholm Alternative
IV. The Fredholm Operator
V. Hermitian Kernels and the Hilbert–Schmidt Theory
VI. Singular Integral Equations on the Real Line
VII. The Cauchy Kernel and the Riemann–Hilbert Problem
VIII. Wiener–Hopf Integral Equation
IX. Nonlinear Integral Equations
X. A Taylor Expansion Technique

GLOSSARY

Cauchy representation Representation of a function $f(z)$:

$$F(z) = \frac{1}{2\pi i}\int_C \frac{f(\zeta)}{\zeta - z}\,d\zeta$$

where z and ζ are points in the complex plane $\mathbb{C}$ while C is a contour in $\mathbb{C}$.

Compact operator Operator that transforms any bounded set in a Hilbert space onto a precompact set.

Eigenvalue Complex numbers λ_n of the integral equation $g(x) = \lambda \int_a^b K(x, y)g(y)\,dy$.

Eigenfunctions Nonzero solutions $g_n(x)$ of the integral equation $g(x) = \lambda \int_a^b K(x, y)g(y)\,dy$.

Fredholm alternative Alternative that specifies the conditions under which a linear integral equation can have a unique solution.

Green's function Kernel obtained when a differential operator is inverted into an integral operator.

Hermitian kernel Kernel $K(x, y)$ if $K(x, y) = \overline{K(y, x)}$, where bar indicates complex conjugate.

Hilbert–Schmidt theory Theory pertaining to the series expansion of a function $f(x)$ which can be represented in the form $\int_a^b K(x, y)h(y)\,dy$.

Kernel Function $K(x, y)$ occurring under the integral sign of the integral equation $\phi(x)g(x) = f(x) + \lambda \int_a^b K(x, y)F\{y, g(y)\}\,dy$.

Neumann series Series obtained by solving the integral equation by successive approximations.

Riemann–Hilbert problem Conversion of a singular integral equation with Cauchy kernel into an algebraic

equation in terms of the boundary values of the function.

Separable kernel Kernel $K(x, y) = \sum_{i=1}^{n} a_i(x)b_i(y)$; also called degenerate.

Singular integral equation Equation in which either the kernel in the integral equation is singular or one or both of the limits of integration are infinite.

INTEGRAL EQUATIONS are equations in which the unknown function appears under the integral sign. They arise in the quest for the integral representation formulas for the solution of a differential operator so as to include the boundary and initial conditions. They also arise naturally in describing phenomena by models which require summation over space and time. Among the integral equations which have received the most attention are the Fredholm- and Volterra-type equations. In the study of singular integral equations, the prominent ones are the Abel, Cauchy, and Carleman type.

I. DEFINITIONS, CLASSIFICATION, AND NOTATION

An integral equation is a functional equation in which the unknown variable $g(x)$ appears under the integral sign. A general example of an integral equation is

$$\phi(x)g(x) = f(x) + \lambda \int_a^b F\{x, y; g(y)\}\, dy \qquad a \le x \le b, \quad (1)$$

where $\phi(x)$, $f(x)$, and $F\{x, y, g(y)\}$ are known functions and $g(x)$ is to be evaluated. The quantity λ is a complex parameter. When $F\{x, y, g(y)\} = K(x, y)g(y)$, Eq. (1) becomes a *linear* integral equation:

$$\phi(x)g(x) = f(x) + \lambda \int_a^b K(x, y)g(y)\, dy \qquad a \le x \le b, \quad (2)$$

where $K(x, y)$ is called a kernel. Four special cases of Eq. (2) are extensively studied. In the *Fredholm* integral equation of *the first kind* $\phi(x) = 0$, and in his equation of the *second kind* $\phi(x) = 1$; in both cases a and b are constants. The *Volterra integral equations* of the first and second kinds are like the corresponding Fredholm integral equations except that now $b = x$. If $f(x) = 0$ in either case, the equation is called homogeneous.

A nonlinear integral equation may occur in the form (1) or the function $f\{x, y, g(y)\}$ may have the form $K(x, y)F(y, g(y))$ where $F(y, g(y))$ is nonlinear in $g(y)$.

When one or both limits of integration become infinite or when the kernel becomes infinite at one or more points within the range of integration, the integral equation is called *singular*.

We shall mainly deal with functions which are either continuous or integrable or square integrable. A function $g(x)$ is square integrable if $\int_a^b |g(x)|^2\, dx < \infty$, and is called an $\mathcal{L}_2$ function. The kernel $K(x, y)$ is an $\mathcal{L}_2$ function if

$$\int_a^b |K(x, y)|^2\, dx < \infty$$

$$\int_a^b |K(x, y)|^2\, dy < \infty \quad (3)$$

$$\int_a^b \int_a^b |K(x, y)|^2\, dx\, dy < \infty.$$

We shall use the inner product (or scalar product) notation

$$\langle \phi, \psi \rangle = \int_a^b \phi(x)\bar{\psi}(x)\, dx, \quad (4)$$

where the bar indicates complex conjugate. The functions ϕ and ψ are orthogonal if $\langle \phi, \psi \rangle = 0$. The norm of the function ϕ is $||\phi|| = (\langle \phi, \phi \rangle)^{1/2}$. If $\|\phi\| = 1$, then ϕ is called *normalized*. In terms of this norm the famous Cauchy–Schwarz inequality can be written as

$$|\langle \phi, \psi \rangle| \le \|\phi\| \|\psi\| \quad (5)$$

while the Minkowski inequality is

$$\|\phi + \psi\| \le \|\phi\| + \|\psi\|. \quad (6)$$

NOTATION. We shall sometimes write the right-hand side of Eq. (2) as $f + \lambda K g$ and call K the *Fredholm operator*. Furthermore, for Fredholm integral equations it will be assumed that the range of integration is a to b unless the contrary is stated. The limits a and b will be omitted.

II. THE METHOD OF SUCCESSIVE APPROXIMATIONS

Our aim is to solve the inhomogeneous Fredholm integral equation

$$g(x) = f(x) + \lambda \int K(x, y)g(y)\, dy, \quad (7)$$

where we assume that $f(x)$ and $K(x, y)$ are in the space $\mathcal{L}_2[a, b]$, by Picard's method of successive approximations. The method is based on choosing the first approximation as $g_0(x) = f(x)$. This is substituted into Eq. (7) under the integral sign to obtain the second approximation and the process is then repeated. This results in the sequence

$$\begin{aligned} g_0(x) &= f(x) \\ g_1(x) &= f(x) + \lambda \int K(x, y) g_0(y)\, dy \\ &\vdots \\ g_m(x) &= f(x) + \lambda \int K(x, y) g_{m-1}(y)\, dy. \end{aligned} \tag{8}$$

The analysis is facilitated when we utilize the iterated kernels defined as

$$\begin{aligned} K_1(x, y) &= K(x, y) \\ K_2(x, y) &= \int K(x, s) K_1(s, y)\, ds \\ &\vdots \\ K_m(x, y) &= \int K(x, s) K_{m-1}(s, y)\, ds. \end{aligned} \tag{9}$$

It can be proved that $K_{m+n}(x, y) = \int K_m(x, s) K_n(s, y)\, ds$. Thereby, we can express the mth approximation in (8) as

$$g_m(x) = f(x) + \lambda \int \left[\sum_{n=1}^{m} \lambda^{n-1} K_n(x, y) \right] f(y)\, dy \tag{10}$$

When we let $m \to \infty$, we obtain formally the so-called *Neumann series*

$$\begin{aligned} g(m) &= \lim_{m \to \infty} g_m(x) \\ &= f(x) + \sum_{m=1}^{\infty} \lambda^m \int K_m(x, y) f(y)\, dy. \end{aligned} \tag{11}$$

In order to examine the convergence of this series, we apply the Cauchy–Schwarz inequality (5) to the general term $\int K_m(x, y) f(y)\, dy$ and get

$$\begin{aligned} &\left| \int K_m(x, y) f(y)\, dy \right|^2 \\ &\quad \le \left(\int |K_m(x, y)|^2\, dy \right) \int |f(y)|^2\, dy. \end{aligned} \tag{12}$$

Let us denote the norm $\|f\|$ as D and the upper bound of the integral $\int |K_m(x, y)|^2\, dy$ as C_m^2, so that relation (12) becomes

$$\left| \int K_m(x, y) f(y)\, dy \right|^2 \le C_m^2 D^2 \tag{13}$$

We can connect the estimate C_m^2 with C_{m-1}^2 by applying the Cauchy–Schwarz inequality to relation (9) and then integrating with respect to y so that $\int |K_m(x, y)|^2\, dy \le B^2 C_{m-1}^2$ where $B^2 = \int\int |K(x, y)|^2\, dx\, dy$. Continuing this process we get the relation $C_m^2 \le B^{2m-2} C_1^2$. Substituting it in (13) we arrive at the inequality

$$\left| \int K_m(x, y) f(y)\, dy \right|^2 < C_1^2 D^2 B^{2m-2}. \tag{14}$$

This means that the infinite series (11) converges faster than the geometric series with common ratio $|\lambda| B$. Thus, if $|\lambda| B < 1$, the Neumann series converges uniformly and absolutely. Fortunately, the condition $|\lambda| B < 1$ also assures us that solution (11) is unique as can be easily proved. In view of the uniform convergence of series (11) we can change the order of integration and summation in it and write it as

$$g(x) = f(x) + \lambda \int \Gamma(x, y; \lambda) f(y)\, dy, \tag{15}$$

where $\Gamma(x, y; \lambda) = \sum_{m=1}^{\infty} \lambda^{m-1} K_m(x, y)$ is called the *resolvent kernel*. This series is also convergent at least for $|\lambda| B < 1$. Indeed, the resolvent kernel is an analytic function of λ, regular at least inside the circle $|\lambda| B < 1$. From the uniqueness of the solution it can be proved that the resolvent kernel is unique.

A few remarks are in order:

1. We can start with any other suitable function for the first approximation $g_0(s)$.
2. The Neumann series, in general, cannot be summed in closed form.
3. The solution of Eq. (7) may exist even if $|\lambda| B > 1$.

The same iterative scheme is applicable to the Volterra integral equation of the second kind:

$$g(x) = f(x) + \lambda \int_a^x K(x, y) g(y)\, dy. \tag{16}$$

In this case the formulas corresponding to (11) and (15) are

$$g(x) = f(x) + \sum_{m=1}^{\infty} \lambda^m \int_a^x K_m(x, y) f(y)\, dy \tag{17}$$

and

$$g(x) = f(x) + \lambda \int_a^x \Gamma(x, y; \lambda) f(y)\, dy, \tag{18}$$

where the iterated kernel $K_m(x, y)$ satisfies the recurrence formula $K_m(x, y) = \int_y^x K(x, s) K_{m-1}(s, y)\, ds$, with $K_1(x, y) = K(x, y)$ as before. The resolvent kernel is given by the same formula as given previously and is an entire function of λ for any given (x, y).

III. THE FREDHOLM ALTERNATIVE

Let us consider the inhomogeneous Fredholm integral equation of the second kind

$$g(x) = f(x) + \lambda \int K(x, y)g(y)\,dy, \tag{19}$$

when the kernel is *degenerate* (*separable*) (i.e., $K(x, y) = \sum_{k=1}^{n} a_k(x)b_k(y)$, where $a_k(x)$ and $b_k(y), k = 1, \ldots, n$, are linearly independent functions). Thus, Eq. (19) becomes

$$g(x) = f(x) + \lambda \sum_{k=1}^{n} a_k(x) \int b_k(y)g(y)\,dy, \tag{20}$$

where we have exchanged summation with integration. It emerges that the technique of solving Eq. (20) depends on the choice of the complex parameter λ and on the constants c_k defined as

$$c_k = \int b_k(y)g(y)\,dy, \tag{21}$$

which are unknown because $g(y)$ is so. Thereby, Eq. (20) takes the algebraic form

$$g(x) = f(x) + \lambda \sum_{k=1}^{n} c_k a_k(x) \tag{22}$$

Next, we multiply both sides of (22) by $b_i(x)$ and integrate from a to b so that we have a set of linear algebraic equations

$$c_i = f_i + \lambda \sum_{k=1}^{n} a_{ik}c_k, \qquad i = 1, 2, \ldots, n, \tag{23}$$

where

$$\begin{aligned} f_i &= \int b_i(x)f(x)\,dx \\ a_{ik} &= \int b_i(x)a_k(x)\,dx. \end{aligned} \tag{24}$$

Let us write the algebraic system (23) in the matrix form

$$(\mathbf{I} - \lambda\mathbf{A})\mathbf{c} = \mathbf{f}, \tag{25}$$

where $\mathbf{I}$ is the identity matrix of order n, $\mathbf{A}$ is the matrix a_{ik}, while $\mathbf{c}$ and $\mathbf{f}$ are column matrices.

The determinant $D(\lambda)$ of the algebraic system (23) is

$$D(\lambda) = \begin{vmatrix} 1 - \lambda a_{11} & -\lambda a_{12} & \cdots & -\lambda a_{1n} \\ -\lambda a_{21} & 1 - \lambda a_{22} & \cdots & -\lambda a_{2n} \\ \vdots & & & \\ -\lambda a_{n1} & -\lambda a_{n2} & \cdots & 1 - \lambda a_{nn} \end{vmatrix}, \tag{26}$$

which is a polynomial of degree at most n in λ. Note that $D(\lambda)$ is not identically zero because when $\lambda = 0$, it reduces to unity. Accordingly, for all values of λ for which $D(\lambda) \neq 0$, the algebraic system (25) and thereby integral equation (19) has a unique solution. On the other hand, for all values of λ for which $D(\lambda) = 0$, algebraic system (25), and with it integral equation (19), is either insoluble or has an infinite number of solutions. We discuss both these cases.

The Case $D(\lambda) \neq 0$. In this case the algebraic system (25) has only one solution given by Cramer's rule

$$\begin{aligned} c_k &= \left(\frac{D_{1k}f_1 + \cdots + D_{hk}f_h + \cdots + D_{nk}f_n}{D(\lambda)} \right) \\ & k = 1, 2, \ldots, n, \end{aligned} \tag{27}$$

where D_{hk} denotes the cofactor of the (h, k)th element of the determinant (26). When we substitute (27) in (22) we obtain the unique solution

$$\begin{aligned} g(x) &= f(x) + \lambda \frac{\sum_{j=1}^{n} \sum_{k=1}^{n} D_{jk}f_i a_k(x)}{D(\lambda)} \\ &= f(x) + \frac{\lambda}{D(\lambda)} \\ &\quad \times \int \left\{ \sum_{j=1}^{n} \sum_{k=1}^{n} D_{jk}b_j(y)a_k(x) \right\} f(y)\,dy, \end{aligned} \tag{28}$$

where we have used relation (24). This expression can be put in an elegant form if we introduce the determinant

$$\begin{aligned} &D(x, y; \lambda) \\ &= \begin{vmatrix} 0 & -a_1(x) & -a_2(x) & \cdots & -a_n(x) \\ b_1(y) & 1 - \lambda a_{11} & -\lambda a_{12} & \cdots & -\lambda a_{1n} \\ b_2(y) & -\lambda a_{21} & 1 - \lambda a_{22} & \cdots & -\lambda a_{2n} \\ \vdots & & & & \\ b_n(y) & -\lambda a_{n1} & -\lambda a_{n2} & \cdots & 1 - \lambda a_{nn} \end{vmatrix} \end{aligned} \tag{29}$$

which is called the *Fredholm minor*. Then Eq. (28) takes the form

$$g(x) = f(x) + \lambda \int \Gamma(x, y; \lambda)f(y)\,dy \tag{30a}$$

where the resolvent kernel Γ is the ratio of two determinants, that is,

$$\begin{aligned} \Gamma(x, y; \lambda) &= D(x, y; \lambda)/D(\lambda) \\ &= \frac{1}{D(\lambda)} \sum_{j=1}^{n} \sum_{k=1}^{n} D_{jk}b_j(y)a_k(x). \end{aligned} \tag{30b}$$

It is clear from the above analysis that if we start with the homogeneous integral equation,

$$g(x) = \lambda \int K(x, y)g(y)\,dy, \tag{31}$$

we shall obtain the homogeneous algebraic system

$$(\mathbf{I} - \lambda\mathbf{A})\mathbf{c} = 0. \tag{32}$$

When $D(\lambda) \neq 0$, this algebraic system and the homogeneous integral equation (31) have only the trivial solutions.

The Case $D(\lambda) = 0$. In this case the algebraic system (32) and hence the homogeneous integral equation (31) may have either no solution or infinitely many solutions. To examine these possibilities it is necessary to discuss the subject of eigenvalues and eigenfunctions of the homogeneous problem (31). Strictly speaking we should write it as $\int K(x, y)g(y)\,dy = \omega g(x)$ for ω to be an eigenvalue, but in the theory of integral equations it has become customary to call the parameter $\lambda \neq 0$, for which the homogeneous equation (31) has a nontrivial solution, its eigenvalue. The corresponding solution $g(x)$ is called the eigenfunction of the operator K. From the above analysis it follows that the eigenvalues of (31) are the solutions of the polynomial $|\mathbf{I} - \lambda\mathbf{A}| = 0$. There may exist more than one eigenfunction corresponding to a specific eigenvalue. Let us denote the number r of such eigenfunctions as $g_{i1}, g_{i2}, \ldots, g_{ir}$ corresponding to the eigenvalue λ_i. The number r is called the *index* of the eigenvalue λ_i (it is also called the *geometric* multiplicity of λ_i while the *algebraic* multiplicity m means that $D(\lambda) = 0$ has m equal roots). We know from linear algebra that if p is the rank of the determinant $D(\lambda_i) = |\mathbf{I} - \lambda_i\mathbf{A}|$, then $r = n - p$. If $r = 1$, λ_i is called a simple eigenvalue. Let us assume that the eigenfunctions $g_{i1}, \ldots, g_{ir}$ have been normalized (i.e., $\|g_{ij}\| = 1$ for $j = 1, \ldots, r$). Then to each eigenvalue λ_i of index $r = n - p$, there corresponds a solution $g_i(x)$ of the homogeneous integral equation (31) of the form $g_i(x) = \sum_{k=1}^{r} \alpha_k g_{ik}(x)$, where α_k are arbitrary constants.

For studying the case when the inhomogeneous integral equation (19) has a solution even when $D(\lambda) = 0$, we need the integral equation

$$\psi(x) = f(x) + \lambda \int \overline{K(y, x)}\psi(y)\,dy, \tag{33}$$

which is called the *transpose* (or *adjoint*) of Eq. (19) which is then the transpose of Eq. (33). For the separable kernel $K(x, y)$ as considered in this section, the transpose kernel is $\overline{K(y, x)} = \sum_{k=1}^{n} \overline{a_k(y)}\overline{b_k(x)}$. When we follow the same steps that we followed for integral equation (19) we find that the transposed integral equation (33) leads to the algebraic system $(\mathbf{I} - \lambda\bar{\mathbf{A}}^{\mathrm{T}})\mathbf{c} = \mathbf{f}$, where $\mathbf{A}^{\mathrm{T}}$ stands for the transpose of $\mathbf{A}$ while c_k and f_k are now defined as

$$c_k = \int \overline{a_k(y)}\psi(y)\,dy,$$

and (34)

$$f_k = \int \overline{a_k(y)}f(y)\,dy,$$

respectively. Clearly, the determinant of this algebraic system is $\overline{D(\bar{\lambda})}$. Accordingly, the transposed integral equation (33) also possesses a unique solution whenever (19) does. Also, the eigenvalues of the homogeneous part of Eq. (33), that is,

$$\psi(x) = \lambda \int \overline{K(y, x)}\psi(y)\,dy \tag{35}$$

are the complex conjugates of those for (31). The eigenvectors of the homogeneous system $(\mathbf{I} - \lambda\bar{\mathbf{A}}^{\mathrm{T}})\mathbf{c} = 0$, are, in general, different from the corresponding eigenvectors of system (32). The same applies to the eigenfunctions of the transposed integral equation (35). Because the geometric multiplicity r of λ_i for (31) is the same as that of $\bar{\lambda}_i$ for (35), the number of linearly independent eigenfunctions of the transposed equation (35) corresponding to $\bar{\lambda}_i$ are also r in number, say, $\psi_{i1}, \psi_{i2}, \ldots, \psi_{ir}$ which we assume to be normalized. Accordingly, any solution $\psi_i(x)$ of (35) corresponding to the eigenvalue λ_i is of the form $\psi_i(x) = \sum_{k=1}^{n} \beta_k\psi_{ik}(x)$, where β_i are arbitrary constants. Incidentally, it can be easily proved that the eigenfunctions $g(x)$ and $\psi(x)$ corresponding to eigenvalues λ_1 and $\bar{\lambda}_2$ $(\lambda_1 \neq \lambda_2)$ of the homogeneous integral equation (31) and its transpose (35) respectively, are orthogonal.

This analysis is sufficient for us to prove that the necessary and sufficient condition for Eq. (19) to have a solution for $\lambda = \lambda_i$, a root of $D(\lambda) = 0$, is that $f(x)$ be orthogonal to the r eigenfunctions ψ_{ij}, $j = 1, \ldots, r$, of the transposed equation (35). The necessary part follows from the fact that if Eq. (19) for $\lambda = \lambda_i$ admits a certain solution $g(x)$, then

$$\begin{aligned}
&\int f(x)\overline{\psi_{ij}(x)}\,dx \\
&\quad = \int g(x)\overline{\psi_{ij}(x)}\,dx \\
&\qquad - \lambda_i \int \overline{\psi_{ij}(x)}\,dx \int K(x, y)g(y)\,dy \\
&\quad = \int g(x)\overline{\psi_{ij}}(x)\,dx \\
&\qquad - \left(\lambda_i \int \overline{g(y)}\,dy \int \overline{K(x, y)}\psi_{ij}(x)\,dx\right) \\
&\quad = 0
\end{aligned}$$

because $\bar{\lambda}_i$ and $\psi_{ij}(x)$ are an eigenvalue and a corresponding eigenfunction of (35). For the proof of the sufficiency, we appeal to the corresponding condition of orthogonality for the linear algebraic system which assures us that the inhomogeneous system (25) reduces to only $n - r$ independent equations (i.e., the rank of the matrix $(I - \lambda\mathbf{A})$ is exactly $p = n - r$ and therefore the system ($(\mathbf{I} - \lambda\mathbf{A})\mathbf{c} = \mathbf{f}$ is soluble). Substituting this value of $\mathbf{c}$ in (22) we have the required solution of (19).

This analysis is true for a general integrable kernel. Fredholm gave three theorems in this connection and they bear his name. These theorems are of great importance in general discussion but are of little use in constructing closed form solutions or obtaining solutions numerically. Fredholm's first theorem gives the same formula as (30a) where the resolvent kernel is

$$\Gamma(x, y; \lambda) = D(x, y; \lambda)/D(\lambda), \qquad D(\lambda) \neq 0, \quad (36)$$

which is a meromorphic function of the complex variable λ, being the ratio of two entire functions $D(x, y; \lambda)$ and $D(\lambda)$ which are given by suitable Fredholm series of form similar to (30b). The other two theorems discuss the case $D(\lambda) = 0$ for a general kernel. The discussion given above and these three theorems add up to the following important result.

A. The Fredholm Alternative Theorem

For a fixed λ, either the integral equation (19) possesses one and only one solution $g(x)$ for integrable functions $f(x)$ and $K(x, y)$ (in particular the solution $g(x) = 0$ for the homogeneous equation (31)), or the homogeneous equation (31) possesses a finite number r of linearly independent solutions g_{ij}, $j = 1, \ldots, r$, with respect to the eigenvalue $\lambda = \lambda_i$. In the first case, the transposed inhomogeneous equation (33) also possesses a unique solution. In the second case, the transposed homogeneous equation (35) also has r linearly independent solutions $\psi_{ij}(x)$, $j = 1, \ldots, r$, corresponding to the eigenvalues λ_i; and the inhomogeneous integral equation (19) has a solution if and only if the given function $f(x)$ is orthogonal to all the eigenfunctions $\psi_{ij}(x)$. In this case the general solution of the integral equation (19) is determined only up to an additive linear combination $\sum_{j=1}^{r} \alpha_j g_{ij}(x)$.

IV. THE FREDHOLM OPERATOR

We have observed in Section I that the Fredholm operator $Kg(x) = \int K(x, y)g(y)\,dy$ is linear (i.e., $K(\alpha g_1 + \beta g_2) = \alpha K g_1 + \beta K g_2$, where α and β are arbitrary complex numbers). In this section we study some general results for this operator. For this purpose we consider a linear space of an infinite dimension with inner product defined by (4) (i.e., $\langle f, g\rangle = \int f(x)\bar{g}(x)\,dx$). This inner product is a complex number and satisfies the following axioms:

(a) $\langle f, f\rangle = 0$ iff $f = 0$.
(b) $\langle \alpha f_1 + \beta f_2, g\rangle = \alpha\langle f_1, g\rangle + \beta\langle f_2, g\rangle$.
(c) $\langle f, g\rangle = \overline{\langle g, f\rangle}$.

The norm $\|f\|$ defined by (4) generates the natural metric $d(f, g) = \|f - g\|$. Furthermore, we have the Cauchy–Schwarz and Minkowski inequalities as given by (5) and (6), respectively.

An important concept in the study of metric spaces is that of completeness. A metric space is called complete if every Cauchy sequence of functions in this space is a convergent sequence (i.e., the limit is in this space). A *Hilbert space* H is an inner product linear space that is complete in its natural metric. An important example is the space of square integrable functions on the interval $[a, b]$. It is denoted as $\mathscr{L}_2[a, b]$, called $\mathscr{L}_2$ space in the sequel.

An operator K is called *bounded* if there exists a constant $M > 0$, such that $\|Kg\| \leq M\|g\|$ for all $g \in \mathscr{L}_2$. We can prove that the Fredholm operator with an $\mathscr{L}_2$ kernel is bounded by starting with the relation $f = Kg$. Then by using the Cauchy–Schwarz inequality we have

$$|f(x)|^2 = \left| \int K(x, y)g(y) \right|^2 dy \leq \int |K(x, y)|^2\,dy \int |g(y)|^2\,dy.$$

Integrating both sides of this relation we find that $\|f\| = \|Kg\| \leq \|g\|[\iint |K(x, y)|^2\,dx\,dy]^{1/2}$ and we have established the boundedness of K. The norm $\|K\|$ of an operator K is defined as

$$\|K\| = \sup(\|Kg\|/\|g\|) \quad (37a)$$

or

$$\|K\| = (\sup \|Kg\|; \|g\| = 1). \quad (37b)$$

The operator K is called continuous in a Hilbert space if whenever $\{g_n\}$ is a sequence in the domain of K with limit g, then $Kg_n \to Kg$. A linear operator is continuous if it is bounded and conversely.

A set S is called precompact if a convergent subsequence can be extracted from any sequence of elements in S. A bounded linear operator K is called compact if it transforms any bounded set in H onto a precompact set. Any bounded operator K, whose range is finite-dimensional, is compact because it transforms a bounded set in H into a bounded finite-dimensional set which is necessarily precompact. Many interesting integral operators are compact. For instance, if the Hilbert space is $\mathscr{L}_2$ space and $K(x, y)$ is a degenerate kernel $\sum_{i=1}^{n} a_i(x)b_i(y)$, then K is a compact operator. This follows by observing that

$$Kg = \int \left[\sum_{i=1}^{n} a_i(x)b_i(y)g(y) \right] dy = \sum_{i=1}^{n} c_i a_i(x)$$

(i.e., the range of K is a finite-dimensional subspace of $\mathscr{L}_2$). Furthermore,

$$\|Kg\| = \left\| \sum_{i=1}^{n} c_i a_i(x) \right\| \le \sum_{i=1}^{n} |c_i| \|a_i\|$$

$$\le \sum_{i=1}^{n} \|a_i\| \int |b_i(y)||g_i(y)|\, dy.$$

Finally, by applying the Cauchy–Schwarz inequality we have $\|Kg\| \le M\|g\|$, where $M = \sum_{i=1}^{n} \|a_i\|\|b_i\|$. Accordingly, K is a bounded linear operator with finite-dimensional range and hence it is compact.

The following property of compact operators will prove useful in the next section. Let K be a compact operator on the Hilbert space H, and L be a bounded operator. Then both KL and LK are compact operators. To prove this property for the case of LK we let $\{f_n\}$ be a uniformly bounded sequence in H. Since K is compact it contains a subsequence $\{f_{n'}\}$ such that $\|LKf_{n'} - LKf_{m'}\| \le \|L\|\|Kf_{n'} - Kf_{m'}\|$. This means that $\{LKf_n\}$ is also a Cauchy sequence. In the case of KL, we observe that if $\{f_n\}$ is uniformly bounded, and L is bounded, then $\{Lf_n\}$ is also uniformly bounded. The compactness of K now assures us that there exists a subsequence $\{KLf_{n'}\}$ which is a Cauchy sequence. Thus, KL is also compact. As a particular case we find that if K is compact then so is K^2.

There are many interesting results regarding the compact operators. Some of them are as follows.

1. If $\{K_n\}$ is a sequence of compact operators on a Hilbert space H, such that for some K we have $\lim_{n\to\infty} \|K - K_n\| = 0$, then K is compact.
2. If $K(x, y)$ is continuous for all $a \le x, y \le b$ then K is a compact operator on $\mathscr{L}_2[a, b]$.
3. An $\mathscr{L}_2$ kernel K (i.e., $\iint |K(x, y)|^2\, dx\, dy < \infty$) is a compact operator.

V. HERMITIAN KERNELS AND THE HILBERT–SCHMIDT THEORY

Let us now consider a technique quite different from the Neumann and Fredholm series. This technique is based on considering the eigenvalues and eigenfunctions of the homogeneous integral equation with *Hermitian* kernel (i.e., $K(x, y) = \overline{K(y, x)}$ (then K is called the Hermitian operator). For real K it becomes $K(x, y) = K(y, x)$ (i.e., a symmetric kernel). We shall restrict ourselves to the Hilbert space $\mathscr{L}_2$ of square integrable functions and will benefit from the concepts of the previous section. Because

$$\begin{aligned}\langle K\phi, \psi\rangle &= \int \bar{\psi}(x)\left[\int K(x, y)\phi(y)\, dy\right] dx \\ &= \int \phi(y)\left[\int K(x, y)\bar{\psi}(x)\, dx\right] dy \\ &= \int \phi(x)\left[\int K(y, x)\bar{\psi}(y)\, dy\right] dx \\ &= \langle \phi, \bar{K}\psi\rangle, \end{aligned} \tag{38}$$

the operator $\bar{K}(y, x)$ is called the *adjoint* operator. When K is Hermitian, this result becomes $\langle K\phi, \psi\rangle = \langle \phi, K\psi\rangle$ (i.e., K is *selfadjoint*). On the other hand, $\langle K\phi, \phi\rangle = \overline{\langle \phi, K\phi\rangle}$. Combining these two results we observe that the inner product $\langle K\phi, \phi\rangle$ is always real and the converse is also true.

Systems of orthogonal functions play an important role in this section so we give a brief account here. A finite or an infinite set $\{\phi_k\}$ is said to be an orthogonal set if $\langle \phi_i, \phi_j\rangle = 0, i \ne j$. If none of the elements of this set is zero vector, it is said to be a proper orthogonal set. A set is orthonormal if $\langle \phi_i, \phi_j\rangle = \delta_{ij}$, where δ_{ij} is the Kronecker delta which is zero for $i \ne j$ and 1 if $i = j$. As defined in section 1. a function ϕ for which $\|\phi\| = 1$ is said to be normalized. Given a finite or a countably infinite independent set of functions $\{\psi_n\}$, we can replace them by an equivalent set $\{\phi_n\}$ which is orthonormal. This is achieved by the so-called *Gram–Schmidt* procedure:

$$\begin{aligned}\phi_1 &= \frac{\psi_1}{\|\psi_1\|} \\ \phi_1 &= \frac{\psi_2 - \langle \psi_2, \phi_1\rangle\phi_1}{\|\psi_2 - \langle \psi_2, \phi_1\rangle\phi_1\|} \\ &\vdots \\ \phi_n &= \frac{\psi_n - \sum_{i=1}^{n-1} \langle \psi_n, \phi_i\rangle\phi_i}{\|\psi_n - \sum_{i=1}^{n-1} \langle \psi_n, \phi_i\rangle\phi_i\|}\end{aligned}$$

as is easily verified. In case we are given a set of orthogonal functions then we can convert it into an orthonormal set simply by dividing each function by its norm.

Starting from an arbitrary orthonormal system we can construct the theory of Fourier series on the same lines as the trigonometric series. This is achieved if we attempt to find the best approximation of an arbitrary function $\psi(x)$ in terms of a linear combination of an orthonormal set $\{\phi_n\}$. By this we mean that we choose the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ in order to minimize

$$\left\| \psi - \sum_{i=1}^{n} \alpha_i \phi_i \right\|^2 = \|\psi\|^2 + \sum_{i=1}^{n} |\langle \psi, \phi_i\rangle - \alpha_i|^2 - \sum_{i=1}^{n} |\langle \psi, \phi_i\rangle|^2. \tag{39}$$

Clearly, the minimum is achieved by setting $\alpha_i = \langle \psi, \phi_i \rangle = a_i$ (say), which are called the Fourier coefficients of ψ relative to the orthonormal system $\{\phi_n\}$. Then (39) becomes

$$\left\| \psi - \sum_{i=1}^{n} \alpha_i \phi_i \right\|^2 = \|\psi\|^2 - \sum_{i=1}^{n} |a_i|^2,$$

from which it follows that $\sum_{i=1}^{n} |a_i|^2 \leq \|\psi\|^2$. For an infinite set $\{\phi_n\}$ this inequality is

$$\sum_{i=1}^{\infty} |a_i|^2 \leq \|\psi\|^2, \tag{40}$$

which is called the *Bessel inequality*. If an orthonormal system of functions can be found in $\mathcal{L}_2$ space such that its every element can be represented linearly in terms of this system, it is called an *orthonormal basis*. Thus, we have $\psi = \sum_i \langle \psi, \phi_i \rangle \phi_i = \sum_i a_i \phi_i$. From this relation we readily derive the Parseval identity:

$$\|\psi\|^2 = \sum_{i=1}^{\infty} |\langle \psi, \phi_i \rangle|^2$$

Incidentally, if $\|\psi - \sum_{i=1}^{n} \alpha_i \phi_i\| \to 0$ as $n \to \infty$, then ψ is said to *converge in the mean* to the series $\sum_{i=1}^{\infty} \alpha_i \phi_i$. The reason for this terminology is that the norm entails the integration of the square of the function.

We are now ready to discuss the solutions of the integral equation $\lambda K\phi = \phi$. We take $K(x, y)$ to be a Hermitian kernel so that, in view of relation (38) K is self-adjoint. As such, this operator has some very interesting properties as we establish below.

A. Property 1

1. Existence of an Eigenvalue

If the Hermitian kernel is also an $\mathcal{L}_2$ function then at least one of the quantities $\pm(\|K\|)^{-1}$, where this norm is defined by (37), must be an eigenvalue of $\lambda K f = f$.

2. Proof

Because K is a Hermitian kernel, it is self-adjoint because of (38), and being an $\mathcal{L}_2$ function it is compact as mentioned in the previous section. Accordingly, we consider a sequence $\{f_n\}$ such that $\|f_n\| = 1$ and $\|Kf_n\|$ converges to $\|K\|$. Note that

$$\begin{aligned} 0 &\leq \left\| K^2 f_n - \|Kf_n\|^2 f_n \right\|^2 \\ &= \left\| K^2 f_n \right\|^2 - 2\|Kf_n\|^2 \langle K^2 f_n, f_n \rangle + \|Kf_n\|^4 \\ &= \left\| K^2 f_n \right\|^2 - \|Kf_n\|^4 \end{aligned}$$

But $\|Kf_n\|$ converges to $\|K\|$, so that $\lim_{n\to\infty} \|K^2 f_n - \|Kf_n\|^2 f_n\| = 0$. In this relation we can use the fact that K^2 being the product of compact operators is compact and, therefore, we can extract a subsequence $\{f_{n'}\}$ from $\{f_n\}$ so that $\{K^2 f_{n'}\}$ converges to a function, say $\|K\|^2 g$. Thus $\lim_{n\to\infty} \|\|K^2\| g - \|K\|^2 f_{n'}\| = 0$, so that $\{f_{n'}\}$ converges to g and $K^2 g = \|K\|^2 g$. But this means that $((\|K\|)^{-1} K - 1)((\|K\|)^{-1} K + 1) g = 0$ and the result follows.

As mentioned in section III, an eigenvalue is simple if there is only one corresponding eigenfunction, otherwise the eigenvalue is called degenerate. The spectrum of K is the set of all eigenvalues, in view of the above-mentioned property we find that the *spectrum of a Hermitian kernel is never empty*.

B. Property 2

The eigenvalues of a Hermitian operator K are all real.

1. Proof

Suppose $\lambda K\phi = \phi$. Then by taking the inner product of this relation with ϕ, we have $\lambda \langle K\phi, \phi \rangle = \|\phi\|^2$ or $\lambda = \|\phi\|^2 / \langle K\phi, \phi \rangle$. As already observed $\langle K\phi, \phi \rangle$ is real for a Hermitian kernel. The quantity $\|\phi\|^2$ is also real. Thus λ is real.

C. Property 3

All eigenfunctions of a Hermitian operator K corresponding to distinct eigenvalues are orthogonal.

1. Proof

Let ϕ_1 and ϕ_2 be the eigenfunctions corresponding respectively, to the distinct eigenvalues λ_1 and λ_2. Then we have $\lambda_1 K\phi_1 = \phi_1$ and $\lambda_2 K\phi_2 = \phi_2$ so that $\langle K\phi_1, \phi_2 \rangle = \lambda_1^{-1} \langle \phi_1, \phi_2 \rangle$ and also $\langle K\phi_1, \phi_2 \rangle = \langle \phi_1, K\phi_2 \rangle = \lambda_2^{-1} \langle \phi_1, \phi_2 \rangle$. Thus $\lambda_2 \langle \phi_1, \phi_2 \rangle = \lambda_1 \langle \phi_1, \phi_2 \rangle$. Since $\lambda \neq \lambda_2$, we find that $\langle \phi_1, \phi_2 \rangle = 0$, as desired.

D. Property 4

The multiplicity of any nonzero eigenvalue is finite for every Hermitian operator K with $\mathcal{L}_2$ kernel.

1. Proof

Let the functions $\phi_{1\lambda}(x), \phi_{2\lambda}(x), \ldots, \phi_{n\lambda}(x), \ldots$ be the linearly independent eigenfunctions which correspond to a nonzero eigenvalue λ. We appeal to the Gram–Schmidt process and then find linear combinations of

these functions which form an orthonormal system $\{u_{k\lambda}(x)\}$. Their complex conjugate functions $\{\overline{u_{k\lambda}(x)}\}$ also form an orthonormal system. Let the function $K(x, y)$ with fixed x be written as $K(x, y) \sim \sum_i a_i \overline{u_{i\lambda}}(y)$, where $a_i = \int K(x, y) u_{i\lambda}(y)\, dy = u_{i\lambda}(x)/\lambda$. By applying the Bessel inequality (40) to this series, we have $\int |K(x, y)|^2\, dy \geq \sum_i (\lambda^{-1})^2 |u_{i\lambda}(x)|^2$, which, when integrated, yields the inequality

$$\int\int |K(x, y)|^2\, dx\, dy \geq \sum_i (\lambda^{-1})^2 = \frac{m}{\lambda^2}$$

where m is the multiplicity of λ. But K is an $\mathscr{L}_2$ kernel so the left-hand side of this relation is finite. It follows that m is finite.

E. Property 5

The sequence of eigenfunctions of a Hermitian kernel K can be made orthonormal.

1. Proof

Suppose that, corresponding to a certain eigenvalue, there are m linearly independent eigenfunctions. Because K is a linear operator, every linear combination of these functions is also an eigenfunction. Thus, by the Gram–Schmidt procedure, we can get an equivalent number of eigenfunctions which are orthonormal. On the other hand, for distinct eigenvalues, the corresponding eigenfunctions are orthogonal and can be readily normalized. Combining these two parts we have established the required property.

F. Property 6

The eigenvalues of a Hermitian operator K with $\mathscr{L}_2$ kernel form a finite or an infinite sequence $\{\lambda_n\}$ with no finite limit point. Furthermore, if we include each eigenvalue in the sequence a number of times equal to its algebraic multiplicity, then

$$\sum_{n=1}^{\infty} \left(\lambda_n^{-1}\right)^2 \leq \int\int |K(x, y)|^2\, dx\, dy$$

The proof of this property can be presented by a slight extension of the arguments given in the analysis of Property 4.

G. Property 7

The set of eigenvalues of the nth iterated kernel coincide with the set of nth powers of the eigenvalues of the kernel K(x, y).

1. Proof

Let λ be an eigenvalue of K with corresponding eigenfunction $\phi(x)$ so that $(I - \lambda K)\phi = 0$, where I is the identity operator. The result of operating on both sides of this equation with the operator $(I + \lambda K)$ yields $(I - \lambda^2 K^2)\phi = 0$, or

$$\phi(x) - \lambda^2 \int K_2(x, y)\phi(y)\, dy = 0,$$

which proves that λ^2 is an eigenvalue of the kernel $K_2(x, y)$. Conversely, if λ^2 is an eigenvalue $K_2(x, y)$, with $\phi(x)$ as its corresponding eigenfunction $\phi(x)$, we have

$$(I - \lambda^2 K^2)\phi = 0$$

or

$$(I - \lambda K)(I + \lambda K)\phi = 0$$

If λ is an eigenvalue of K this equation is satisfied and we have established the property for $n = 2$. Otherwise, we set $[(I + \lambda K)]\phi(x) = \mathscr{X}(x)$ so that the above equation becomes $(I - \lambda K)\mathscr{X}(x) = 0$. Since we have assumed that λ is not an eigenvalue of K, it follows that $\mathscr{X}(x) \equiv 0$. This means that $(I + \lambda K)\phi = 0$, or $-\lambda$ is an eigenvalue and we have again proved our result for $n = 2$. The general result follows by continuing this process.

In the next stage we wish to expand a nonzero Hermitian kernel in terms of its eigenfunctions. It may have a finite or infinite number of real eigenvalues. We order them in the sequence $\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots$ in such a way that

1. Each eigenvalue is repeated as many times as its multiplicity, and
2. we enumerate these eigenvalues in an order which corresponds to their absolute values (i.e., $0 < |\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_n| \leq |\lambda_{n+1}| \leq \cdots$).

Let $\phi_1(x), \phi_2(x), \ldots, \phi_n(x), \ldots$ be the sequence of corresponding orthonormalized eigenfunctions which are arranged in such a way that they are no longer repeated and are linearly independent in each group corresponding to the same eigenvalue. Thus to each eigenvalue λ_k there corresponds just one eigenfunction $\phi_k(x)$. We shall have this ordering in mind in the sequel.

Property 1 has assured us that a nonzero Hermitian kernel always has a finite, nonzero lowest eigenvalue λ_1. Let ϕ_1 be the corresponding eigenfunction. Now we remove this eigenvalue from the spectrum of K by defining the truncated kernel $K^{(2)}(x, y)$:

$$K^{(2)}(x, y) = K(x, y) - (\phi_1(x)\bar{\phi}_1(y)/\lambda_1)$$

which is also Hermitian. If $K^{(2)}(x, y)$ is nonzero then let λ_2 be its lowest eigenvalue with corresponding eigenfunction $\phi_2(x)$. Because

$$K^{(2)}\phi_1 = K\phi_1 - (\phi_1(x)/\lambda_1)\int \phi_1(y)\phi_1(y)\,dy = 0$$

we observe that $\phi_1 \neq \phi_2$ even if $\lambda_1 = \lambda_2$. Similarly, the third truncated Hermitian kernel is

$$K^{(3)}(x, y) = K^{(2)}(x, y) - \phi_2(x)\frac{\overline{\phi_2}(y)}{\lambda_2}$$
$$= K(x, y) - \sum_{k=1}^{2}\frac{\phi_k(x)\overline{\phi_k}(y)}{\lambda_k}$$

Continuing this process we have

$$K^{(n+1)}(x, y) = K(x, y) - \sum_{k=1}^{n}\frac{\phi_k(x)\overline{\phi_k}(y)}{\lambda_k}, \qquad (41)$$

which yields the $(n+1)$th lowest eigenvalue and the corresponding eigenfunction $\phi_{n+1}(x)$. Thereby we find that either this process terminates after n steps (i.e., $K^{(n+1)}(x, y) = 0$), and the kernel $K(x, y)$ is a degenerate kernel $\sum_{k=1}^{n}(\phi_k(x)\overline{\phi_k}(y)/\lambda_k)$, or the process can be continued indefinitely and there are an infinite number of eigenvalues and eigenfunctions so that

$$K(x, y) = \sum_{k=1}^{\infty}\frac{\phi_k(x)\overline{\phi_k}(y)}{\lambda_k}. \qquad (42)$$

This is called the bilinear form of the kernel. Recall that we meet a similar situation for a Hermitian matrix **A**. Indeed, by transforming to an orthonormal basis of the vector space consisting of the eigenvectors of **A** we can transform it to a diagonal matrix.

From the bilinear form (42) we derive a useful inequality. Let the sequence $\{\phi_k(x)\}$ be all the eigenfunctions of a Hermitian $\mathcal{L}_2$ kernel $K(x, y)$ with $\{\lambda_k\}$ as the corresponding eigenvalues as described and arranged in the above analysis. Then the series

$$\sum_{n=1}^{\infty}\frac{|\phi_n(x)|^2}{\lambda_n^2} < C_1^2, \qquad (43)$$

where C_1^2 is an upper bound of the integral $\int |K(x, y)|^2\,dy$. The proof follows by observing that the Fourier coefficients a_n of the function $K(x, y)$ with fixed x, with respect to the orthonormal system $\overline{\phi_n}(y)$ are $a_n = \langle K(x, y), \bar{\phi}_n(y)\rangle = \phi_n(x)\lambda_n$. Substituting these values of a_n in the Bessel inequality (40) we derive (43).

The eigenfunctions do not have to form a complete set in order to represent the functions in $\mathcal{L}_2$. Indeed, any function which can be written as "sourcewise" in terms of the kernel K (i.e., any function $f = Kh$) can be expanded in a series of the eigenfunctions of K. This is not surprising because integration smooths out irregularities or if the functions f, K, and h are represented by the series, then the convergence of the series representing f will be better than that of the series representing K and h. This important concept for a Hermitian kernel is embodied in the theorem.

H. Hilbert–Schmidt Theorem

If $f(x)$ can be written in the form

$$f(x) = \int K(x, y)h(y)\,dy, \qquad (44)$$

where K and h are in $\mathcal{L}_2$ space, then $f(x)$ can be expanded in an absolutely and uniformly convergent Fourier series with respect to the orthonormal system of eigenfunctions of K, that is,

$$f(x) = \sum_{n=1}^{\infty} f_n\phi_n(x),\ f_n = \langle f, \phi_n\rangle. \qquad (45a)$$

The Fourier coefficients f_n are related to the corresponding coefficients h_n of $h(x)$ as

$$f_n = h_n/\lambda_n = \langle h, \phi_n\rangle/\lambda_n \qquad (45b)$$

and λ_n are the eigenvalues of K.

1. Proof

The Fourier coefficients of the function $f(x)$ with respect to the orthonormal system $\{\phi_n(x)\}$ are

$$f_n = \langle f, \phi_n\rangle = \langle Kh, \phi_n\rangle = \langle h, K\phi_n\rangle$$
$$= \langle h, \phi_n\rangle/\lambda_n = h_n/\lambda_n$$

because K is self-adjoint and $\lambda_n K\phi_n = \phi_n$. Accordingly, we can write the correspondence

$$f(x) \sim \sum_{n=1}^{\infty} f_n\phi_n(x) = \sum_{n=1}^{\infty}\frac{h_n\phi_n(x)}{\lambda_n}. \qquad (46)$$

The estimate of the remainder term for this series is

$$\left|\sum_{k=n+1}^{n+p} h_k\frac{\phi_k(x)}{\lambda_k}\right|^2 \leq \sum_{k=n+1}^{n+p} h_k^2\sum_{k=n+1}^{n+p}\frac{|\phi_k(x)|^2}{\lambda_k^2}$$
$$\leq \sum_{k=n+1}^{n+p} h_k^2\sum_{k=1}^{\infty}\frac{\left|\phi_k^2(x)\right|}{\lambda_k^2} \qquad (47)$$

Now the series $\sum_{k=1}^{\infty}|\phi_k^2(x)|/\lambda_k^2$ is bounded in view of relation (43) while the partial sum $\sum_{k=n+1}^{n+p} h_k^2$ can be made arbitrarily small because $h(x) \in \mathcal{L}_2$ and, as such, the series $\sum_{k=1}^{\infty} h_k^2$ is convergent. Thus, the estimate (47) can be made arbitrarily small so that series (46) converges absolutely and uniformly. Next, we show that this series converges to $f(x)$ in the mean and for this purpose we denote its partial sum as $\psi_n(x) = \sum_{m=1}^{n}(h_m/\lambda_m)\phi_m(x)$ and estimate the value $\|f(x) - \psi_n(x)\|$. Because

$$f(x) - \psi_n(x) = Kh - \sum_{m=1}^{n}\frac{h_m}{\lambda_m}\phi_m(x) = K^{(n+1)}h,$$

where K^{n+1} is the truncated kernel (41), we find that

$$\begin{aligned}\|f(x)-\psi_n(x)\|^2 &= \left\|K^{(n+1)}h\right\|^2\\ &= \left\langle K^{(n+1)}h,\, K^{(n+1)}h\right\rangle\\ &= \left\langle h,\, K^{(n+1)}K^{(n+1)}h\right\rangle\\ &= \left\langle h,\, K_2^{(n+1)}h\right\rangle \qquad (48)\end{aligned}$$

in view of the self-adjointness of the kernel $K^{(n+1)}$ and the relation $K^{(n+1)}K^{(n+1)} = K_2^{(n+1)}$. Now we use Property 7 for the Hermitian kernels and find that the least eigenvalue of the kernel $K_2^{(n+1)}$ is λ_{n+1}^2. On the other hand, Property 1 implies that

$$\begin{aligned}1/\lambda_{n+1}^2 &= \sup\left\|K^{(n+1)}h\right\|^2/\|h\|^2\\ &= \sup\left(\left\langle h,\, K_2^{(n+1)}h\right\rangle\right)/\|h\|^2.\end{aligned}$$

Combining it with (48) we get $\|f(x)-\psi_n(x)\|^2 \le \|h\|^2/\lambda_{n+1}^2$. Because $\lambda_{n+1}\to\infty$, we have proved that $\|f(x)-\psi_n(x)\|\to 0$ as $n\to\infty$.

In order to prove that $f=\psi$, where ψ is the series with partial sum ψ_n, we use the Minkowski inequality (6) and get $\|f-\psi\| \le \|f-\psi_n\| + \|\psi_n-\psi\|$. The first term on the right-handside of this inequality tends to zero as proved above. Because series (47) converges uniformly, the second term can be made as small as we want (i.e., given an arbitrarily small and positive ε we can find n large enough that $|\psi_n-\psi|<\varepsilon$). One integration then yields $\|\psi_n-\psi\| < \varepsilon(b-a)^{1/2}$ and we have proved the result.

Let us now use the foregoing theorem for solving the inhomogeneous Fredholm integral equation of the second kind:

$$g(x) = f(x) + \lambda\int K(x,y)g(y)\,dy, \qquad (49)$$

with a Hermitian $\mathscr{L}_2$ kernel. First, we assume that λ is not an eigenvalue of K. Because $g(x)-f(x)$ in this equation has the integral representation of the form (44) we expand both $g(x)$ and $f(x)$ in terms of the eigenfunctions $\phi_n(x)$ given by the homogeneous equation $\phi_n(x)=\lambda_n\int K(x,y)\phi_n(y)\,dy$. Accordingly, we set

$$g(x) = \sum_{n=1}^{\infty} g_n\phi_n(x), \qquad f(x) = \sum_{n=1}^{\infty} f_n\phi_n(x), \qquad (50)$$

where $g_n = \langle g, \phi_n\rangle$ is unknown and $f_n = \langle f, \phi_n\rangle$ is known. Substituting these expansions in (49) we obtain

$$\sum_{n=1}^{\infty} g_n\phi_n(x) = \sum_{n=1}^{\infty} f_n\phi_n(x) + \lambda\int K(x,y)\sum_{n=1}^{\infty} g_n\phi_n(y)\,dy. \qquad (51)$$

In view of the uniform convergence of the expansion, we can interchange the order of integration and summation, and get

$$\sum_{n=1}^{\infty} g_n\phi_n(x) = \sum_{n=1}^{\infty} f_n\phi_n(x) + \lambda\sum_{n=1}^{\infty}\frac{g_n\phi_n(x)}{\lambda_n} \qquad (52)$$

Now we multiply both sides of (52) by $\overline{\phi_k(x)}$ and integrate from a to b and appeal to the orthogonality of the eigenfunctions and obtain

$$g_k = f_k + (\lambda/\lambda_k)g_k \qquad (53a)$$

or

$$g_k = f_k + (\lambda/(\lambda_k-\lambda))f_k. \qquad (53b)$$

Substitution of (53) into (50) leads us to the required solution

$$\begin{aligned}g(x) &= \sum_{n=1}^{\infty}\left(f_n + \frac{\lambda}{(\lambda_n-\lambda)}f_n\right)\phi_n(x)\\ &= f(x) + \lambda\sum_{n=1}^{\infty}\int\frac{\phi_n(x)\overline{\phi_n}(y)}{(\lambda_n-\lambda)}f(y)\,dy\\ &= f(x) + \lambda\int\Gamma(x,y;\lambda)f(y)\,dy, \qquad (54)\end{aligned}$$

where the resolvent kernel $\Gamma(x,y;\lambda)$ is expressed by the series

$$\Gamma(x,y;\lambda) = \sum_{n=1}^{\infty}\frac{\phi_n(x)\overline{\phi_n}(y)}{\lambda_n-\lambda}, \qquad (55)$$

and we have again interchanged the integration and summation. It follows from expression (55) that the singular points of the resolvent kernel Γ corresponding to a Hermitian $\mathscr{L}_2$ kernel are simple poles and every pole is an eigenvalue of the kernel.

In the event that λ in Eq. (49) is equal to one of the eigenvalues, say, λ_p of K, solution (55) becomes infinite. To remedy it we return to relation (53) which for $k=p$ becomes $g_p = f_p + g_p$. Thus, g_p is arbitrary and $f_p = 0$. This implies that $\int f(x)\phi_p(x)\,dx = 0$ (i.e., $f(x)$ is orthogonal to the eigenfunction $\phi_p(x)$). If this is not the case, we have no solution. If λ_p has the algebraic multiplicity m, then there are m coefficients g_p which are arbitrary and $f(x)$ is orthogonal to all these m functions.

Integral equations arise in the process of inverting ordinary and partial differential operators. In the quest for the representation formula for the solutions of these operators so as to include the initial or boundary values in it, we arrive at integral equations. In the process there arises the theory of Green's functions which are symmetric and become the kernels of the integral equations. If they are not symmetric, then they can be symmetrized. We illustrate these concepts with the help of the Sturm–Liouville differential operator

$$L = -\frac{d}{dx}\left\{p(x)\frac{d}{dx}\right\} + q(x)$$

where $p(x)$ and $q(x)$ are continuous in the interval $[a, b]$ and in addition $p(x)$ has a continuous derivative and does not vanish in this interval. We discuss two kinds of equations, that is,

$$Ly = f(x), \qquad a \le x \le b \tag{56}$$

and

$$Ly - \lambda r(x)y = 0, \tag{57}$$

where $f(x)$ and $r(x)$ are given functions. The function $r(x)$ is continuous and nonnegative in $[a, b]$. Each of these equations is subject to the boundary conditions

$$\begin{aligned} \alpha_1 y(a) + \alpha_2 y'(a) &= 0 \\ \beta_1 y(b) + \beta_2 y'(b) &= 0. \end{aligned} \tag{58}$$

Let us assume that it is not possible to obtain a nonzero solution of system (57) and (58) for the case $\lambda = 0$ (this means that there is no eigenfunction corresponding to the eigenvalue $\lambda = 0$). Accordingly, we assume that a function ϕ_1 satisfies the boundary condition $(58)_1: \alpha_1\phi_1(a) + \alpha_2\phi_1'(a) = 0$ and another function ϕ_2 (independent of ϕ_1) satisfies the boundary condition $(58)_2: \beta_1\phi_2(b) + \beta_2\phi_2'(b) = 0$. This amounts to solving two initial value problems, namely, $L\phi_1 = 0$, $\phi_1(a) = -\alpha_2$, $\phi_1'(a) = \alpha_1$ and $L\phi_2 = 0$, $\phi_2(b) = -\beta_2$, $\phi_2'(b) = \beta_1$. With the help of these two linearly independent solutions ϕ_1 and ϕ_2 we use the method of variation of parameters and assume the solution of the inhomogeneous equation (56) as

$$y(x) = C_1(x)\phi_1(x) + C_2(x)\phi_2(x), \tag{59}$$

where C_1 and C_2 must be found from the relations

$$C_1'(x)\phi_1(x) + C_2'(x)\phi_2(x) = 0 \tag{60a}$$

$$C_1'(x)\phi_1'(x) + C_2'(x)\phi_2'(x) = -f(x)/p(x). \tag{60b}$$

We need one more relation for ϕ_1 and ϕ_2 to facilitate the solution of (56). This is found from the fact that ϕ_1 and ϕ_2 are two linearly independent solutions of the homogeneous equation $Ly = 0$. As such

$$\begin{aligned} 0 &= \phi_2 L\phi_1 - \phi_1 L\phi_2 \\ &= -\phi_2\frac{d}{dx}\left(p\frac{d\phi_1}{dx}\right) + \phi_1\frac{d}{dx}\left(p\frac{d\phi_2}{dx}\right) \\ &= -\frac{d}{dx}\left\{p\left(\phi_2\frac{d\phi_1}{dx} - \phi_1\frac{d\phi_2}{dx}\right)\right\}, \end{aligned}$$

so that the quantity within the braces is a constant. Inasmuch as ϕ_1 and ϕ_2 can be determined up to a constant factor, we can choose this expression to be

$$p\left(\phi_2\frac{d\phi_1}{dx} - \phi_1\frac{d\phi_2}{dx}\right) = -1. \tag{61}$$

From relations (60) and (61) we find that $C_1'(x) = -\phi_2(x)f(x)$ and $C_2'(x) = \phi_1(x)f(x)$. Thus

$$C_1(x) = \int_x^b \phi_2(\xi)f(\xi)\,d\xi$$

$$C_2(x) = \int_a^x \phi_1(\xi)f(\xi)\,d\xi,$$

with convenient constants of integration. Substituting these values in (59) we arrive at the solution

$$\begin{aligned} y(x) &= \phi_1(x)\int_x^b \phi_2(\xi)f(\xi)\,d\xi \\ &\quad + \phi_2(x)\int_a^x \phi_1(\xi)f(\xi)\,d\xi \\ &= \int_a^b G(x, \xi)f(\xi)\,d\xi, \end{aligned} \tag{62}$$

where the function

$$G(x, \xi) = \begin{cases} \phi_1(\xi)\phi_2(x), & \xi \le x \\ \phi_1(x)\phi_2(\xi), & \xi \ge x \end{cases} \tag{63}$$

is called the Green's function of this boundary value problem. It can be written in an elegant form by defining the regions

$$x_< = \min(x, \xi) = \begin{cases} x, & a \le x \le \xi \\ \xi, & \xi \le x \le b \end{cases}$$

$$x_> = \max(x, \xi) = \begin{cases} \xi, & a \le x \le \xi \\ x, & \xi \le x \le b. \end{cases}$$

Then $G(x, \xi) = \phi_1(x_<)\phi_2(x_>)$.

Finally, we attend to Eq. (57) whose solution follows from (62) to be

$$y(x) = \lambda\int r(\xi)G(x, \xi)y(\xi)\,d\xi. \tag{64}$$

This is an integral equation with kernel $r(\xi)G(x, \xi)$. Although this kernel is not symmetric, it can be symmetrized by setting $[r(x)]^{1/2}y(x) = g(x)$ and defining the symmetric kernel $K(x, y) = G(x, \xi)[r(x)]^{1/2}[r(\xi)]^{1/2}$. Then (64) becomes the integral equation

$$g(x) = \lambda\int K(x, y)g(y)\,dy. \tag{65}$$

If there was a term on the right-hand side of (57) we would have arrived at an inhomogeneous integral equation of the second kind. Incidentally, it can be readily verified that Eq. (64) satisfies the given boundary conditions. The case

when $\lambda = 0$ has a nonzero eigenfunction can be handled by a slight extension of the foregoing arguments.

The theory of Green's functions as derived above can be displayed very elegantly with the help of the Dirac delta function and other generalized functions.

VI. SINGULAR INTEGRAL EQUATIONS ON THE REAL LINE

An integral equation is said to be singular either if the kernel is singular within the range of integration or if one or both limits of integration are infinite. In this section we study some famous singular integral equations. They arise very frequently in various branches of physics and engineering. No general theory is available for these equations but methods are available for solving some special cases. We start with the Abel integral equation.

A. The Abel Integral Equation

This equation is

$$f(x) = \int_a^x \frac{g(y)}{(x-y)^\alpha}\,dy, \qquad 0 < \alpha < 1. \tag{66}$$

To solve it we multiply both sides of this equation by $dx/(u-x)^{1-\alpha}$ and integrate with respect to x from a to u so that we have

$$\int_a^u \frac{f(x)\,dx}{(u-x)^{1-\alpha}} = \int_a^u \frac{dx}{(u-x)^{1-\alpha}} \int_a^x \frac{g(y)\,dy}{(x-y)^\alpha}.$$

When we change the order of integration on the right-hand side of this equation we have

$$\int_a^u \frac{f(x)\,dx}{(u-x)^{1-\alpha}} = \int_a^u g(y)\,dy \int_y^u \frac{dx}{(u-x)^{1-\alpha}(x-y)^a}. \tag{67}$$

Next, we set $t = (u-x)/(u-y)$ in the second integral so that

$$\int_y^u (u-x)^{\alpha-1}(x-y)^{-\alpha}\,dx = \int_0^1 t^{\alpha-1}(1-t)^{-\alpha}\,dt = \frac{\pi}{\sin\alpha\pi},$$

where we have used the value of the Eulerian beta function $B(\alpha, 1-\alpha) = \pi \csc \alpha\pi$. Thus relation (67) becomes

$$\frac{\sin\alpha\pi}{\pi} \int_a^u \frac{f(x)\,dx}{(u-x)^{1-\alpha}} = \int_a^u g(y)\,dy. \tag{68}$$

Differentiation of this relation finally yields the solution

$$g(y) = \frac{\sin\alpha\pi}{\pi} \frac{d}{dy}\left[\int_a^y \frac{f(x)}{(y-x)^{1-\alpha}}\,dx\right]. \tag{69}$$

Similarly, the solution of the integral equation

$$\int_s^b \frac{g(y)}{(y-x)^\alpha}\,dy = f(s), \qquad 0 < \alpha < 1 \tag{70}$$

is

$$g(y) = -\frac{\sin\alpha\pi}{\pi} \frac{d}{dy}\left[\int_y^b \frac{f(x)}{(x-y)^{1-\alpha}}\,dx\right]. \tag{71}$$

There are many related integral equations which can be solved by similar steps. For instance, the solution of the integral equation

$$f(x) = \int_a^x \frac{g(y)\,dy}{[h(x)-h(y)]^\alpha}, \qquad 0 < \alpha < 1, \tag{72}$$

where the function $h(x)$ is strictly increasing differentiable function with nonzero $h'(x)$ over some interval $a \le x \le b$, is

$$g(y) = \frac{\sin\alpha\pi}{\pi} \frac{d}{dy}\left[\int_a^y \frac{h'(u)f(u)\,du}{[h(y)-h(u)]^{1-\alpha}}\right]. \tag{73}$$

Similarly, the solution of the integral equation

$$f(x) = \int_x^b \frac{g(y)\,dy}{[h(y)-h(x)]^\alpha}, \qquad 0 < \alpha < 1 \tag{74}$$

is

$$g(y) = -\frac{\sin\alpha\pi}{\pi} \frac{d}{dy}\left[\int_y^b \frac{h'(u)f(u)\,du}{[h(u)-h(y)]^{1-\alpha}}\right]. \tag{75}$$

These relations remain valid when $a \to -\infty$ and $b \to \infty$.

B. The Cauchy Integral Equation

The equation

$$g(x) = f(x) + \lambda \int_0^1 \frac{g(y)\,dy}{y-x} \tag{76}$$

is the inhomogeneous Cauchy integral equation. Here the integral is the Cauchy principal value. To solve this equation we appeal to the identity

$$\int_0^u \frac{dy}{(u-y)^{\alpha-1}y^\alpha(y-x)} = \begin{cases} \dfrac{\pi\cot\alpha\pi}{(u-x)^{1-\alpha}x^\alpha}, & 0 < x < u \\ \dfrac{-\pi\csc\alpha\pi}{(x-u)^{1-\alpha}x^\alpha}, & u < x \end{cases} \tag{77}$$

and then define the function $\phi(x, u)$ as

$$\phi(x,u) = \frac{1}{(u-x)^{1-\alpha}x^\alpha}, \qquad 0 < x < u, \tag{78}$$

where α is such that $-\pi\cot\alpha\pi = (1/\lambda)$. Then $\phi(x, u)$ is the solution of the integral equation

$$-\lambda \int_0^u \frac{\phi(y,u)}{y-x}\,dy = \phi(x,u), \qquad 0 < x < u \tag{79}$$

while

$$\int_0^u \frac{\phi(y,u)}{y-x}\,dy = -\frac{\pi \csc \alpha\pi}{(x-u)^{1-\alpha}x^{\alpha}}, \qquad u < x. \quad (80)$$

When we multiply (76) by x we get

$$\lambda \int_0^1 \frac{yg(y)\,dy}{y-x} = xg(x) - xf(x) + c, \quad (81)$$

where $c = \lambda \int_0^1 g(y)\,dy$. Next, we multiply both sides of (81) by $\phi(x,u)$ as defined by (78), integrate from 0 to u and change the order of integration. The result is

$$\begin{aligned} &-\lambda \int_0^u yg(y)\,dy \int_0^u \frac{\phi(x,u)\,dx}{x-y} \\ &-\lambda \int_u^1 yg(y)\,dy \int_0^u \frac{\phi(x,u)\,dx}{x-y} \\ &= \int_0^u xg(x)\phi(x,u)\,dx \\ &\quad - \int_0^u xf(x)\phi(x,u)\,dx + c\int_0^u \phi(x,u)\,dx. \end{aligned}$$

With the help of relations (79) and (80) and the fact that $\int_0^u \phi(x,u)\,dx = \pi \csc \alpha\pi$, the above relation becomes

$$\begin{aligned} &\lambda\pi \csc\alpha\pi \int_u^1 \frac{y^{1-\alpha}g(y)}{(y-u)^{1-\alpha}}\,dy \\ &\quad = -\int_0^u xf(x)\phi(x,u)\,dx + c\pi \csc\alpha\pi \end{aligned} \quad (82)$$

This is an Abel-type integral equation whose solution is found from the previous analysis to be

$$\begin{aligned} \lambda y^{1-\alpha}g(y) = \frac{\sin^2\alpha\pi}{\pi^2}\frac{d}{dy}\Bigg[\int_y^1 \int_0^u (u-y)^{-\alpha} \\ \times (u-x)^{\alpha-1}x^{1-\alpha}f(x)\,dx\,dy\Bigg] + \frac{c\sin\alpha\pi}{\pi(1-y)^{\alpha}} \end{aligned} \quad (83)$$

Now we use the relation $-\pi \cot\alpha\pi = 1/\lambda$ and do a little algebraic manipulation and obtain the required solution as

$$\begin{aligned} g(x) = -\frac{f(x)}{1+\pi^2\lambda^2} + \frac{\lambda}{(1+\pi^2\lambda^2)x^{1-\alpha}(1-x)^{\alpha}} \\ \times \int_0^1 \frac{(1-y)^{\alpha}y^{1-\alpha}f(y)\,dy}{y-x} \\ + \frac{c}{x^{1-\alpha}(1-x)^{\alpha}\sqrt{1+\pi^2\lambda^2}}. \end{aligned} \quad (84)$$

Finally, we set $y = (y'-a)/(b-a)$, and find from the above analysis that the solution of the integral equation

$$g(x) = f(x) + \lambda \int_a^b \frac{g(y)\,dy}{y-x} \quad (85)$$

is

$$\begin{aligned} g(x) = -\frac{f(x)}{1+\pi^2\lambda^2} + \frac{\lambda}{(1+\pi^2\lambda^2)(x-a)^{1-\alpha}(b-x)^{\alpha}} \\ \times \int_a^b \frac{(b-y)^{\alpha}(y-a)^{1-\alpha}f(y)\,dy}{y-x} \\ + \frac{c}{(x-a)^{1-\alpha}(b-x)^{\alpha}}, \end{aligned} \quad (86)$$

where c is an arbitary constant.

The solution of the Cauchy-type integral equation of the first kind:

$$\int_a^b \frac{g(y)\,dy}{y-x} = f(x), \qquad a < x < b \quad (87)$$

can be obtained with a very similar analysis and is

$$\begin{aligned} g(x) = \frac{1}{\pi^2\sqrt{(x-a)(b-x)}} \\ \times\Bigg[\int_a^b \frac{\sqrt{(y-a)(b-y)}}{x-y}f(y)\,dy + \pi c\Bigg]. \end{aligned} \quad (88)$$

In particular, when $a=-1$, $b=1$, it follows that the solution of the *airfoil equation*

$$\frac{1}{\pi}\int_{-1}^1 \frac{g(y)\,dy}{y-x} = f(x), \qquad -1 < x < 1 \quad (89)$$

is

$$g(x) = \frac{1}{\pi\sqrt{1-x^2}}\int_{-1}^1 \frac{\sqrt{(1-y^2)}f(y)}{x-y}\,dy + \frac{c}{\sqrt{1-x^2}}. \quad (90)$$

C. Singular Integral Equations with a Logarithmic Kernel

We start with the integral equation

$$\int_{-1}^1 \ln|x-y|g_0(y)\,dy = 1, \qquad -1 < x < 1. \quad (91)$$

By setting $x = \cos\alpha$, $y = \cos\beta$, Eq. (91) becomes

$$\int_0^{\pi} \ln|\cos\alpha - \cos\beta|G(\beta)\,d\beta = 1, \quad 0 < \alpha < \pi, \quad (92)$$

where $G(\beta) = g_0(\cos\beta)\sin\beta$. Let us now expand $G(\beta) = \sum_{n=0}^{\infty} b_n \cos n\beta$ and use the summation formula

$$\ln|\cos\alpha - \cos\beta| = -\ln 2 - 2\sum_{n=1}^{\infty}\frac{\cos n\alpha \cos n\beta}{n}. \quad (93)$$

Then relation (92) becomes

$$\int_0^\pi \left[-\ln 2 - 2\sum_{n=1}^{\infty} \frac{\cos n\alpha \cos n\beta}{n} \right] \times \left[\sum_{m=0}^{\infty} b_m \cos m\beta \right] d\beta = 1,$$

from which it follows, due to orthogonality of cosine functions, that

$$-\pi b_0 \ln 2 - \sum_{n=1}^{\infty} \pi b_n \frac{\cos n\alpha}{n} = 1.$$

Thus, $b_0 = -(1/(\pi \ln 2))$, $b_n = 0$, $n \geq 1$, and we find that the solution of Eq. (91) is

$$g_0(y) = -\frac{1}{\pi \ln 2} \frac{1}{\sqrt{1-y^2}}. \tag{94}$$

In passing we observe that by substituting solution (94) in (91) we have the useful identity

$$\int_{-1}^{1} \frac{\ln|x-y|}{(1-y^2)^{1/2}}\, dy = -\pi \ln 2, \quad -1 < x < 1. \tag{95}$$

Next, we consider the integral equation

$$\int_{-1}^{1} \ln|x-y| g(y)\, dy = f(x), \quad -1 < x < 1. \tag{96}$$

Differentiation with respect to x gives

$$\int_{-1}^{1} \frac{g(y)}{x-y}\, dy = f'(x), \quad -1 < x < 1,$$

whose solution follows from (90) to be

$$g(x) = \frac{1}{\pi^2} \int_{-1}^{1} \left[\frac{1-y^2}{1-x^2} \right]^{1/2} \frac{f'(y)}{y-x}\, dy + \frac{C}{\pi\sqrt{1-x^2}}, \tag{97}$$

where $C = \int_{-1}^{1} g(y)\, dy$. To find the constant C, we multiply (96) by $1/\sqrt{(1-x^2)}$ and integrate it with respect to x from -1 to 1 and change the order of integration. The result is

$$\int_{-1}^{1} g(y)\, dy \int_{-1}^{1} \frac{\ln|x-y|}{(1-x^2)^{1/2}}\, dx = \int_{-1}^{1} \frac{f(x)}{(1-x^2)^{1/2}}\, dx,$$

which, in view of identity (95), becomes

$$(-\pi \ln 2)C = \int_{-1}^{1} \frac{f(x)}{(1-x^2)^{1/2}}\, dx.$$

Thus

$$C = -\frac{1}{\pi \ln 2} \int_{-1}^{1} \frac{f(x)}{(1-x^2)^{1/2}}\, dx,$$

which when substituted in (97) yields the solution

$$g(x) = \frac{1}{\pi^2} \int_{-1}^{1} \left(\frac{1-y^2}{1-x^2} \right)^{1/2} \frac{f'(y)}{y-x}\, dx - \frac{1}{\pi^2 \ln 2 (1-x^2)^{1/2}} \int_{-1}^{1} \frac{f(y)}{(1-y^2)^{1/2}}\, dy. \tag{98}$$

Various other forms of integral equations with logarithmic kernels can be solved in a similar fashion.

VII. THE CAUCHY KERNEL AND THE RIEMANN–HILBERT PROBLEM

For the study of the singular equations in the complex plane $\mathbb{C}$, we require a few important results from the analysis of a complex variable. We present some of these concepts needed for the Cauchy kernel. Let C be a simple, smooth, and closed curve in the complex z plane endowed with the counterclockwise orientation. The complement $\mathbb{C}\backslash C$ consists of two parts, one interior (bounded) part S_+ and the other exterior part S_-. A function $F(z)$ defined and analytic in the complement $\mathbb{C}\backslash C$ is called a *sectionally analytic* function with discontinuity contour C. Let $f(\zeta)$ be a continuous function defined for $\zeta \in C$. The Cauchy (or analytic) representation of f is the sectionally analytic function

$$F(z) = F\{f(\zeta); z\} = \frac{1}{2\pi i} \int_C \frac{f(\zeta)\, d\zeta}{\zeta - z}, \qquad z \in \mathbb{C}\backslash C. \tag{99}$$

The boundary values $F_\pm(\omega)$ of this function on both sides of C satisfy the Plemelj relations

$$F_+(\omega) = \frac{1}{2} f(\omega) - \frac{i}{2} H(f)$$
$$F_-(\omega) = -\frac{1}{2} f(\omega) - \frac{i}{2} H(f), \tag{100}$$

where $H(f) = (1/\pi) \int_C (f(\zeta)/(\zeta - \omega))\, d\zeta, \omega \in C$, is called the Hilbert transform of f. Solving (100) for f and $H(f)$ we get

$$f = F_+ - F_- = [F]$$
$$H(f) = \frac{1}{\pi} \int_C \frac{f(\zeta)}{\zeta - \omega}\, d\zeta = i(F_+ + F_-), \tag{101}$$

where $[F]$ is called the jump of F across C.

Now let $\Phi_1(\zeta)$ and $\Phi_2(\zeta)$ be two continuous functions defined on C. The Riemann–Hilbert problem is to find the sectionally analytic function $Y(z)$ defined on $\mathbb{C}\backslash C$ whose boundary values satisfy

$$\Phi_1(\zeta) Y_+(\zeta) - \Phi_2(\zeta) Y_-(\zeta) = \Psi(\zeta), \tag{102}$$

where $\Psi(\zeta)$ is a function given on C. We assume that Φ_1 and Φ_2 never vanish on C. It is called the normality condition. When we divide both sides of the above equation by $\Phi_1(\zeta)$ we get

$$Y_+(\zeta) = \Phi(\zeta)Y_-(\zeta) + \psi(\zeta), \tag{103}$$

where $\Phi = \Phi_2/\Phi_1$ and $\psi = \Psi/\Phi_1$. When $\psi = 0$. Eq. (103) becomes the homogeneous Riemann–Hilbert problem

$$X_+(\zeta) = \Phi(\zeta)X_-(\zeta). \tag{104}$$

To solve (103) we first reduce it to the simple form

$$W_+(\zeta) = W_-(\zeta) + \psi(\zeta) \tag{105}$$

which is obtained from (103) by taking $\Phi(\zeta) = 1$ because the solution of (105) is known to have the analytic representation

$$W(z) = F\{\psi(\zeta); z\} = \frac{1}{2\pi i}\int_C \frac{\psi(\zeta)\,d\zeta}{\zeta - z}, \tag{106}$$

where we have appealed to definition (99). We first solve the homogeneous Riemann–Hilbert problem (104). For this purpose we take the logarithm of both sides of (104), and get

$$\ln X_+(\zeta) = \ln X_-(\zeta) + \ln \Phi(\zeta), \tag{107}$$

where we assume, for the time being, that $\ln \Phi(\zeta)$ is single valued on C. A particular solution of (107) is given by the analytic representation

$$\begin{aligned} \ln X(z) &= F\{\ln \Phi(\zeta); z\} \\ &= \frac{1}{2\pi i}\int_C \frac{\ln \Phi(\zeta)\,d\zeta}{\zeta - z}. \end{aligned} \tag{108}$$

or $X(z) = e^{F\{\ln \Phi(\zeta); z\}}$, which is a sectionally analytic function that never vanishes on $\mathbb{C}\backslash C$ and whose boundary values satisfy (104) because

$$\begin{aligned} \frac{X_+(\zeta)}{X_-(\zeta)} &= \exp[F_+\{\ln \Phi(\omega); \zeta\} - F_-\{\ln \Phi(\omega); \zeta\}] \\ &= \exp[\ln \Phi(\zeta)] = \Phi(\zeta). \end{aligned}$$

Note that this basic solution is normal because $X(\infty) = 1$. Now, if $Y(z)$ is any other solution of the homogeneous problem (104) then the function $Y(z)/X(z)$, which is known to be analytic on $\mathbb{C}\backslash C$, is also analytic on C because its jump across C vanishes:

$$\left[\frac{Y}{X}\right]_+ - \left[\frac{Y}{X}\right]_- = \frac{\Phi Y_-}{\Phi X_-} - \frac{Y_-}{X_-} = 0.$$

Thus, Y/X is an entire function and the most general solution of the homogeneous Riemann–Hilbert problem is $Y(z) = P(z)X(z)$ where $P(z)$ is an entire function. It is called the fundamental solution of the Riemann–Hilbert problem (fundamental in the sense that all other solutions can be obtained from it in a suitable way).

Let us now consider the case when $\ln \Phi(\zeta)$ is multiple valued on C and introduce the number k

$$k = \frac{1}{2\pi i}\Delta_c(\ln \Phi(\zeta)) = \frac{1}{2\pi}\Delta_c(\arg \ln \Phi(\zeta)), \tag{109}$$

where $\Delta_c(f(\zeta))$ denotes the increment of the function $f(\zeta)$ when the curve C is transversed in the positive direction. Thus, k is the index of the point $z = 0$ with respect to the curve C', the image of the curve C under the function $\Phi(\zeta)$. The number k, which is always an integer, is called the *index* of the Riemann–Hilbert problem.

Let $Y(z)$ be a solution of the homogeneous Riemann–Hilbert problem (104); we define the sectionally analytic function $\bar{Y}(z)$ as

$$\begin{aligned} \bar{Y}(z) &= Y(z), && z \in S_+ \\ \bar{Y}(z) &= (z - z_0)^k Y(z), && z \in S_- \end{aligned} \tag{110}$$

Thus, $\bar{Y}(z)$ satisfies the following boundary value problem where z_0 is an arbitrary point of S_+:

$$\begin{aligned} \bar{Y}_+(\zeta) &= \Phi_0(\zeta)\bar{Y}_-(\zeta), \\ \Phi_0(\zeta) &= (z - z_0)^{-k}\Phi(\zeta) \end{aligned} \tag{111}$$

Thereby $\ln \Phi_0(\zeta)$ has become single valued and we can apply the previous analysis to conclude that the solution of (111) is $\bar{Y}(z) = P(z)\bar{X}(z)$, where $P(z)$ is a polynomial and where $\bar{X}(z) = \exp[F(\ln \phi_0(\zeta); z)]$ is the fundamental solution. Then it follows from (110) that the solutions of (104) are of the form $Y(z) = P(z)X(z)$ where $P(z)$ is an arbitrary polynomial and where the fundamental solution $X(z)$ is given by

$$X(z) = \exp\left[\frac{1}{2\pi i}\int_C \frac{\ln \Phi_0(\zeta)\,d\zeta}{\zeta - z}\right], \quad z \in S_+ \tag{112a}$$

$$X(z) = (z - z_0)^{-k}\exp\left[\frac{1}{2\pi i}\int_C \frac{\ln \Phi_0(\zeta)\,d\zeta}{\zeta - z}\right] \quad z \in S_- \tag{112b}$$

Once a fundamental solution of the Riemann–Hilbert problem has been obtained, we can solve the inhomogeneous problem (103) as follows. Let $X(z)$ be a fundamental solution and let $Y(z)$ be a solution of (103) which we can write as

$$\frac{Y_+}{X_+} = \frac{Y_-}{X_-} + \frac{\psi}{X_+} \tag{113}$$

because $\Phi = X_+/X_-$. The solution of this equation with polynomial behavior at $z = \infty$ is

$$\frac{Y(z)}{X(z)} = P(z) + F\left\{\frac{\psi(\zeta)}{X_+(\zeta)}; z\right\}.$$

Thus

$$Y(z) = P(z)X(z) + X(z)F\left\{\frac{\psi(\zeta)}{X_+(\zeta)}; z\right\}. \tag{114}$$

This formula gives the solution of the Riemann–Hilbert problem with polynomial behavior at $z=\infty$. For obtaining the solutions that vanish at $z=\infty$, it is necessary to consider the sign of the index k. When $k \geq 0$, (114) will be a solution provided the degree of P does not exceed $k-1$. When $k<0$, however, for the solution to vanish at $z=\infty$, the polynomial $P(z)$ should vanish and so should the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_{-k}$ of the Taylor expansion

$$F\left\{\frac{f(\zeta)}{X_+(\zeta)}; z\right\} = \frac{\alpha_1}{z} + \frac{\alpha_2}{z^2} + \cdots \quad \text{at} \quad z=\infty.$$

We are now ready to solve one of the most important singular integral equations, namely, the Carleman integral equation:

$$a(\zeta)g(\zeta) + \frac{b(\zeta)}{\pi i}\oint_C \frac{g(\omega)}{\omega-\zeta}\,d\omega = f(\zeta) \tag{115}$$

over a closed contour C, where $a(\zeta)$, $b(\zeta)$, and $f(\zeta)$ are given functions on C subject to the normality condition $a^2(\zeta) - b^2(\zeta) \neq 0$. To solve this equation we appeal to the analytic representation

$$G(z) = \frac{1}{2\pi i}\oint_C \frac{g(\omega)\,d\omega}{\omega - z} \tag{116}$$

of the unknown function $g(\zeta)$. Next, we substitute the Plemelj formulas (100) in (115) and get

$$G_+(\zeta) = \Phi(\zeta)G_-(\zeta) + \psi(\zeta), \tag{117}$$

where

$$\Phi(\zeta) = \frac{a(\zeta)-b(\zeta)}{a(\zeta)+b(\zeta)}, \quad \psi(\zeta) = \frac{f(\zeta)}{a(\zeta)+b(\zeta)}. \tag{118}$$

Thus, we have the Riemann–Hilbert problem (117) to solve. For this purpose we consider the index (109) and then define the fundamental solution (112). Now

$$\begin{aligned} X_+(\zeta) &= \exp\left\{F_+\left[\ln[\Phi(\omega)(\omega-z_0)^{-k}]; \zeta\right\}\right. \\ &= \exp\left\{\tfrac{1}{2}\ln\left[\Phi(\zeta)(\zeta-z_0)^{-k}\right] + \gamma(\zeta)\right\} \\ &= \sqrt{\Phi(\zeta)}e^{\gamma(\zeta)}/(\zeta-z_0)^{k/2}, \end{aligned} \tag{119}$$

where

$$\gamma(\zeta) = \frac{1}{2\pi i}\oint_C \frac{\ln\left[\Phi(\omega)(\omega-z_0)^{-k}\right]}{\omega-\zeta}\,d\omega. \tag{120}$$

Similarly,

$$X_-(\zeta) = e^{\gamma(\zeta)}\Big/\sqrt{\Phi(\zeta)}(\zeta-z_0)^{k/2}. \tag{121}$$

If $k \geq 0$, the solution of the Riemann–Hilbert problem (117) is given as

$$G(z) = X(z)P(z) + X(z)F\left\{\frac{f(\zeta)}{(a(\zeta)+b(\zeta))X_+(\zeta)}; z\right\}, \tag{122}$$

where $P(z)$ is a polynomial of degree $(k-1)$ at most. If $k<0$, the solution is

$$G(z) = X(z)F\left\{\frac{f(\zeta)}{(a(\zeta)+b(\zeta))X_+(\zeta)}; z\right\}, \tag{123}$$

provided

$$\oint_C \left\{(f(\zeta)\zeta^j/(a(\zeta)+b(\zeta))X_+(\zeta))\right\}d\zeta = 0$$

for

$$0 \leq j \leq -k-1.$$

Because

$$\begin{aligned} &(a(\zeta)+b(\zeta))X_+(\zeta) \\ &\quad = (a(\zeta)+b(\zeta))\sqrt{\Phi(\zeta)}(\zeta-z_0)^{-k/2}e^{\gamma(\zeta)} \\ &\quad = \sqrt{a^2(\zeta)+b^2(\zeta)}(\zeta-z_0)^{-k/2}e^{\gamma(\zeta)}, \end{aligned}$$

it follows by using the Plemelj formula $g(\zeta) = G_+(\zeta) - G_-(\zeta)$, and relations (119), (121), and (123) that the solution of the integral equation (115) is

$$\begin{aligned} g(\zeta) = \frac{a(\zeta)f(\zeta)}{a^2(\zeta)-b^2(\zeta)} &- \frac{e^{\gamma(\zeta)}(\zeta-z_0)^{-k/2}b(\zeta)}{\sqrt{a^2(\zeta)-b^2(\zeta)}\pi_i} \\ &\times \oint_C \frac{f(\omega)e^{-\gamma(\omega)}(\omega-z_0)^{k/2}\,d\omega}{\sqrt{a^2(\omega)-b^2(\omega)}(\omega-\zeta)}\omega \end{aligned} \tag{124}$$

so long as $k<0$ and

$$\begin{aligned} \oint_C \frac{f(\omega)e^{\gamma(\omega)}(\omega-z_0)^{k/2}\omega^j}{\sqrt{a^2(\omega)-b^2(\omega)}}\,d\omega = 0 \\ 0 \leq j \leq -k-1 \end{aligned} \tag{125}$$

Similarly, when $k \geq 0$ the solution becomes

$$\begin{aligned} g(\zeta) = \frac{a(\zeta)f(\zeta)}{a^2(\zeta)-b^2(\zeta)} &- \frac{e^{\gamma(\zeta)}(\zeta-z_0)^{-k/2}b(\zeta)}{\sqrt{a^2(\zeta)-b^2(\zeta)}\,\pi_i} \\ &\times\left[\oint_C \frac{f(\omega)e^{-\gamma(\omega)}(\omega-z_0)^{k/2}\,d\omega}{\sqrt{a^2(\omega)-b^2(\omega)}(\omega-\zeta)} + P(\zeta)\right], \end{aligned} \tag{126}$$

where $P(\zeta)$ is a polynomial whose degree does not exceed $k-1$.

For the special case when a and b are constants. Eq. (115) reduces to the Cauchy integral equation

$$ag(\zeta) + \frac{b}{\pi i}\oint_C \frac{g(\omega)}{\omega - \zeta}\, d\omega = f(\zeta) \tag{127}$$

while its solution follows by observing that $k = 0$, and Φ and γ are constants. Thus, we appeal to relation (126) and find the solution to be

$$g(\zeta) = \frac{af(\zeta)}{a^2 - b^2} - \frac{b}{(a^2 - b^2)\pi i}\oint_C \frac{f(\omega)\, d\omega}{\omega - \zeta}. \tag{128}$$

VIII. WIENER–HOPF INTEGRAL EQUATION

The integral equation of the type

$$g(x) + \lambda \int_0^\infty K(x - y)g(y)\, dy = f(x), \quad 0 < x < \infty \tag{129}$$

is called the Wiener–Hopf integral equation. Its distinguishing features are the difference kernel and the semi-infinite interval. This is an eigenvalue problem in which the eigenvalue happens to be known from certain physical considerations.

To use a two-sided transform, we have to change Eq. (129) in such a way that it is valid for $x < 0$. To achieve this objective we set

$$\lambda \int_{-\infty}^\infty K(x - y)g(y) = \begin{cases} -g(x) + f(x), & 0 < x < \infty \\ h(x), & -\infty < x < 0, \end{cases} \tag{130}$$

where $h(x)$ is unknown. Next we extend the definition of the function $f(x)$, $g(x)$, and $h(x)$ as follows:

$$f(x) = 0, \quad x < 0, \quad g(x) = 0, \quad x < 0, \qquad h(x) = 0, \quad x > 0. \tag{131}$$

This enables us to write Eq. (130) as

$$g_+(x) + \lambda \int_{-\infty}^\infty K(x - y)g_+(y)\, dy = f_+(x) + h_-(x), \quad -\infty < x < \infty, \tag{132}$$

where the subscripts $\pm$ indicate the $\pm$ half line on which the function is nonvanishing. Thereby we have succeeded in transforming the Wiener–Hopf integral Eq. (129) into a convolution-type integral equation.

Let us observe some interesting features of Eq. (132): (1) We assume that the kernel $K(x - y)$ is known for all values of its arguments, namely, $-\infty < x - y < \infty$. (2) The function $h_-(x)$ is unknown and can only be evaluated after we have found $g_+(x)$. (3) Because the integral in this equation extends from $-\infty$ to $+\infty$, we must have growth limits on the functions appearing in Eq. (132). They are: $K(x) = 0(e^{-c|x|})$, $f(x) = 0(e^{dx})$, as $x \to \infty$, the function $h_-(x) = 0(e^{-e|x|})$ as $x \to -\infty$, where $c > 0$ and $d < c$. (4). We look for a solution $g_+(x) = 0(e^{dx})$ as $x \to \infty$.

Let us now apply the Fourier transform $\hat{\psi}(u)$

$$\hat{\psi}(u) = \int_{-\infty}^\infty \psi(x)e^{iux}\, dx \tag{133}$$

to both sides of equation (132) and get

$$(1 + \lambda\hat{K}(u))\hat{g}_+(u) = \hat{f}_+(u) + \hat{h}_-(u). \tag{134}$$

These Fourier transforms have the following strips of definition $\hat{K}(u)$ in $|\mathrm{Im}\, u| < c$, $\hat{g}_+(u)$ in $\mathrm{Im}\ u > d$, $f_+(u)$ in $\mathrm{Im}\ u > d$ and $\hat{h}(u)$ in $\mathrm{Im}\ u < c$. Accordingly, Eq. (134) is a well-defined equation in the strip $d < \mathrm{Im}\, u < c$.

At this stage, two splitting techniques help us in solving Eq. (134). The first one is called the quotient splitting. This is achieved by setting

$$(1 + \lambda)\hat{K}(u) = \frac{K_+(u)}{K_-(u)}, \tag{135}$$

where K_+ and K_- have their respective regions of analyticity. When we substitute Eq. (135) in Eq. (134) we obtain

$$K_+(u)\hat{g}_+(u) = K_-(u)\hat{f}_+(u) + K_-(u)\widehat{h_-}(u). \tag{136}$$

Next we split the mixed term $K_-\hat{f}_+$ in the previous equation as

$$K_-(u)\hat{f}_+(u) = p_+(u) + p_-(u) \tag{137}$$

so that Eq. (136) becomes

$$K_+(u)\hat{g}_+(u) - p_+(u) = K_-(u)\widehat{h_-}(u) + p_-(u). \tag{138}$$

Thereby we have reduced our problem to solving the Riemann–Hilbert boundary value problem as studied in the previous section.

We illustrate all the above-mentioned concepts with the help of the following example.

Let us solve the integral equation:

$$\frac{1}{4}g(x) + \int_0^\infty e^{-|x - y|}g(y)\, dy = 1. \tag{139}$$

The Fourier transform of the kernel $K(x) = e^{-|x|}$ is $\hat{K}(u) = 2/(1 + u^2)$. Because 1_+ is the Heaviside function, which is 0 for $x < 0$ and 1 for $x > 0$, we have $(1_+)^\wedge(u) = i/u$. Note that $2/(1 + u^2)$ is analytic in the strip $-1 < \mathrm{Im} < 1$, while i/u is analytic in the half plane $\mathrm{Im}\ u > 0$.

Now we process Eq (139) by the Wiener–Hopf technique, as explained previously, and find that this integral equation transforms into

$$\frac{q+u^2}{4(1+u^2)}\hat{g}_+(u) = \frac{i}{u} + \hat{h}_-(u), \quad 0 < \operatorname{Im} u < 1. \quad (140)$$

Next we split the coefficient of $g_+(u)$ in (140) as the quotient

$$\frac{9+u^2}{4(1+u^2)} = \frac{(4+3i)(4-3i)}{4(u+i)(u-i)} = \left(\frac{4+3i}{4(u+i)}\right)\Big/\left(\frac{u-i}{u-3i}\right) = \frac{K_+}{K_-}.$$

Then Eq. (140) takes the form

$$\frac{u+3i}{4(u+i)}g_+(u) = \frac{u-i}{u-3i}\frac{i}{4} + \frac{u-i}{u-3i}h_-(u). \quad (141)$$

At this stage we split the first term on the right side of (141) as the sum

$$\frac{u-i}{u-3i}\frac{i}{4} = \frac{i}{3u} + \frac{2i}{3(u-3i)},$$

so that Eq. (141) becomes

$$\frac{4+3i}{4(u+i)}\hat{g}_+(u) - \frac{i}{3u} = \frac{2i}{3(u-3i)} + \frac{u-i}{u-3i}\hat{h}_-(u). \quad (142)$$

The left side of the equation (142) is analytic in $\operatorname{Im} u > 0$, while the right side is valid in the strip $0 < \operatorname{Im} u < 1$. Because these functions are equal on $\operatorname{Im} u = 0$, they are equal everywhere by analytic continuation. Accordingly, they are equal to the same entire function $E(u)$. As the $\operatorname{Im} u \to \infty$, the left side of Eq. (142) tends to zero. Similarly, as $\operatorname{Im} u \to -\infty$, the right side of (142) tends to zero. Therefore $E(u) \equiv 0$. Since our interest is to evaluate $g_+(u)$, we evaluate the part

$$\frac{u+3i}{4(u+i)}\hat{g}_+(u) - \frac{i}{3u} = 0. \quad (143)$$

When inverted, Eq. (143) yields

$$g_+(x) = \frac{4i}{3}\frac{1}{2\pi}\int_{ia-\infty}^{ia+\infty} e^{-iux}\frac{(u+i)\,du}{u(u+3i)}. \quad (144)$$

We evaluate this integral by choosing the contour in the lower half plane. This contour consists of the line $-ia - R$ to $ia + R$, and the semicircle of radius R with center at $(0, a)$ and then we let $R \to \infty$. The value of the integral on the semicircle vanishes. Finally, with the help of the theory of residues we find that the solution is

$$g(x) = \frac{4}{9}(1 + 2e^{-3x})H(x). \quad (145)$$

IX. NONLINEAR INTEGRAL EQUATIONS

Let us finally present an elementary discussion of nonlinear integral equations. Because there are no analytic techniques by which these equations are solved, their treatment is available on ad hoc basis. A nonlinear Fredholm integral equation is of the form

$$g(x) = f(x) = \lambda \int F\{x, y, g(y)\}\,dy, \quad (146)$$

where both x and y lie in the domain (a, b) and we have omitted the limits a, b of integration as in the previous discussion. Equation (146) is called an *Uryson* equation. Perhaps the most important particular case is the *Hammerstein* equation:

$$g(x) = f(x) + \lambda \int K(x, y)F(y, g(y))\,dy. \quad (147)$$

These nonlinear integral equations are being presently studied by modern numerical methods. We shall limit ourselves to giving the iterative scheme for solving the general case (146). This scheme is similar to the one given for the linear case in Section II. For this purpose we impose the following conditions on the functions occurring in this equation. We assume that $f(x)$ is continuous in (a, b). The function $F(x, y, g(y))$ is continuous with respect to x and y in the rectangle $a \le x, y \le b$ and satisfies the Lipschitz condition with respect to $g(y)$ (i.e., $|F\{x, y, g_1(y)\} - F\{x, y, g_2(y)\}| \le L|g_1(y) - g_2(y)|$), where L is a positive constant. The continuity of F with respect to x and y assures that $|F\{x, y, g(y)\}| < M$ for bounded $g(y)$, where M is a positive constant.

Now we follow the iterative technique as given in Section II and set the first approximation for $g(x)$ to be $g^{(0)}(x) = f(x)$ and subsequent ones as

$$g^{(n)}(x) = f(x) + \lambda \int F\left\{x, y, g^{(n-1)}(y)\right\} dy \qquad n \ge 1 \quad (148)$$

Because

$$g^{(n)}(x) = \sum_{k=1}^{n}\left\{g^{(k)}(x) - g^{(k-1)}(x)\right\} + g^{(0)}(x)$$

the convergence of the sequence $g^{(n)}(x)$ is equivalent to the convergence of the series whose kth term is $(g^{(k)}(x) - g^{(k-1)}(x))$. To examine the convergence of this series we set

$$g^{(1)}(x) - g^{(0)}(x) = \lambda \int F(x, y, f(x)) = \phi(x)$$

as $g^{(0)}(x) = f(x)$, and assume that $|\phi(x)| < A$, a constant. Then we find that

$$
\begin{aligned}
&\left|g^{(k)}(x)-g^{(k-1)}(x)\right| \\
&\quad=|\lambda|\left|\int F\left\{x, y, g^{(k-1)}(y)\right\}-F\left\{x, y, g^{(k-2)}(y)\right\} dy\right| \\
&\quad\leq|\lambda|\int\left|F\left\{x, y, g^{(k-1)}(y)\right\}-F\left\{x, y, g^{(k-2)}(y)\right\}\right| dy \\
&\quad\leq|\lambda| L\int\left|g^{(k-1)}(y)-g^{(k-2)}(y)\right| dy \\
&\quad\leq\cdots\leq[|\lambda| L(b-a)]^{k-1} A.
\end{aligned}
$$

Thus, if $|\lambda| L(b-a)<1$, the series will be absolutely and uniformly convergent and $g^{(n)}(x)$ will tend to a function $g(x)$ which will be the solution of Eq. (146).

To prove the uniqueness of the solution, we let $g(x)$ and $h(x)$ be two solutions. Then the value of the difference $\psi(x)=g(x)-h(x)$ is

$$
\psi(x)=\lambda\int[F\{x, y, g(y)\}-F\{x, y, h(y)\}] dy.
$$

When we denote by $\psi_{\max}$, the maximum value of $\psi(x)$ in (a, b) and appeal to the Lipschitz continuity of F, we find that

$$
|\psi(x)|<|\lambda|\int|\psi(y) dy|\leq|\lambda| L(b-a)\psi_{\max}(x),
$$

which means that $\psi_{\max}\leq|\lambda| L(b-a)\psi_{\max}$. But $|\lambda| L(b-a)<1$, so we have $\psi_{\max}=0$ and the uniqueness of the solution is proved.

Let us now consider the nonlinear Volterra integral equation

$$
g(x)=f(x)+\lambda\int_0^x F\{x, y, g(y)\} dy, \tag{149}
$$

with the same conditions on the function $f(x)$ and $F\{x, y, g(y)\}$ as given above. The iterative scheme again yields the sequence $g^{(0)}(x)=f(x)$ and

$$
g^{(n)}(x)=f(x)+\lambda\int_0^x F\left\{x, y, g^{(n-1)}(y)\right\} dy \qquad n\geq 1 \tag{150}
$$

and its convergence is equivalent to that of the series whose kth term is $g^{(k)}(x)-g^{(k-1)}(x)$. To study this convergence we observe that

$$
\begin{aligned}
g^{(1)}(x)-g^{(0)}(x)&=\lambda\int_0^x F\{x, y, f(y)\} dy \\
&\leq|\lambda|\int_0^x|F\{x, y, f(y)\}| dy \\
&\leq|\lambda|\int_0^x M dy=M|\lambda| x
\end{aligned} \tag{151}
$$

Then from (150) it follows that

$$
\begin{aligned}
&\left|g^{(k)}(x)-g^{(k-1)}(x)\right| \\
&\quad=|\lambda|\left|\int_0^x F\left\{x, y, g^{(k-1)}(y)\right\}-F\left\{x, y, g^{(k-2)}(y)\right\} dy\right| \\
&\quad\leq|\lambda|\int_0^x L\left|g^{(k-1)}(y)-g^{(k-2)}(y)\right| dy \\
&\quad\leq\cdots\leq\frac{(L|\lambda||x|)^k}{k!} M
\end{aligned} \tag{152}
$$

But the last term in (152) is the kth term of the power series for $M\exp\{L|\lambda||x|\}$, so that the series $g^{(0)}(x)+\sum_{k=1}^{\infty}\{g^{(k)}(x)-g^{(k-1)}(x)\}$ is absolutely and uniformly convergent for all values of λ and its sum $\lim_{n\to\infty} g^{(n)}(x)$ is the solution of the integral equation (149). The uniqueness of this solution can be proved by a slight extension of the arguments presented above for the Fredholm case.

X. A TAYLOR EXPANSION TECHNIQUE

In the previous analysis we have presented the Neumann and Hilbert-Schmidt expansion techniques for solving the integral equations. Recently, it has been discovered that both the linear and nonlinear integral equations can also be solved with the help of the Taylor series. To present the basic ideas of this method, we consider the Fredholm integral equation of the second kind

$$
g(x)=f(x)+\int_a^b K(x, y)g(y) dy \tag{153}
$$

and differentiate it n times with respect to x so that we have

$$
g^{(n)}(x)=f^{(n)}(x)+\int_a^b\frac{\partial^n K(x, y)}{\partial x^n} g(y) dy.
$$

For $x=0$, the above relation becomes

$$
g^{(n)}(0)=f^{(n)}(0)+\int_a^b\left.\frac{\partial^n K(x, y)}{\partial x^n}\right|_{x=0} g(y) dy. \tag{154}
$$

When we substitute the Taylor series

$$
g(y)=\sum_{m=0}^{\infty}\frac{1}{m!} g^{(m)}(0) y^m \tag{155}
$$

for $g(y)$ in (154) we get

$$
\begin{aligned}
g^{(n)}(0)&=f^{(n)}(0) \\
&\quad+\int_a^b\left.\frac{\partial^n K(x, y)}{\partial x^n}\right|_{x=0}\left(\sum_{m=0}^{\infty}\frac{1}{m!} g^{(m)}(0) y^m\right) dy.
\end{aligned} \tag{156}
$$

Next, we set

$$T_{mn} = \frac{1}{m!}\int_a^b \left.\frac{\partial^n K(x, y)}{\partial x^n}\right|_{x=0} y^m\, dy \tag{157}$$

in (139) and obtain

$$g^{(n)}(0) = f^{(n)}(0) + \sum_{m=0}^{\infty} T_{mn} g^{(m)}(0), \tag{158}$$

$n = 0, 1, 2, 3, \ldots$. Accordingly, the evaluation of the solution $g(x)$ of integral equation (136) reduces to solving the above infinite system of algebraic equations for the Taylor coefficients $g^m(0)$. This is achieved by truncating this system in a suitable manner so that we get a determinate system at every step. Thus, we get an approximate solution to the desired order of accuracy. For instance, let us take (i) $n = 0$, $m = 1$, (ii) $n = 1$, $m = 1$; (iii) $n = 2$, $m = 2$, in (158) so that we get the determinate system of three equations

$$(T_{00} - 1)g(0) + T_{01} g^{(1)}(0) = -f(0) \tag{159a}$$

$$T_{10} g(0) + (T_{11} - 1)g^{(1)}(0) = -f^{(1)}(0) \tag{159b}$$

$$T_{20} g(0) + T_{21} g^{(1)}(0) + (T_{22} - 1)g^{(2)}(0) = -f^{(2)}(0) \tag{159c}$$

Equations (159a) and (159b) yield the values of $g(0)$ and $g^{(1)}(0)$, which when substituted in (159c), give the value of $g^{(2)}(0)$. Thus, we obtain the solution $g(x)$ to $O(x^2)$.

The elegance of the method lies in the fact that frequently we get an exact solution of an integral equation. We illustrate this point with the help of the integral equation

$$g(x) = (x + 1)^2 + \int_{-1}^{1} (xy + x^2y^2) g(y)\, dy. \tag{160}$$

Accordingly, $f(x) = (x + 1)^2$, $K(x, y) = xy + x^2y^2$, and we have

$$f(0) = 1, \quad f^{(1)}(0) = 2, \quad f^{(2)}(0) = 2$$

$$T_{00} = 0, \qquad T_{01} = 0, \qquad T_{10} = 0, \quad T_{11} = \frac{2}{3}$$

$$T_{20} = \frac{4}{3}, \qquad T_{21} = 0, \qquad T_{22} = \frac{2}{5}.$$

When we substitute these values in Eq. (159) and solve for the first three Taylor coefficients of $g(x)$, we get

$$g(0) = 1, \qquad g^{(1)}(0) = 6, \qquad g^{(2)}(0) = \frac{50}{9}.$$

Thus, we have

$$g(x) = 1 + 6x + \frac{25}{9}x^2,$$

which happens to be the exact solution of Eq. (160).

SEE ALSO THE FOLLOWING ARTICLES

CALCULUS • DIFFERENTIAL EQUATIONS, ORDINARY • GREEN'S FUNCTIONS

BIBLIOGRAPHY

Cochran, J. A., (1972). "The Analysis of Linear Integral Equations," McGraw-Hill, New York.

Estrada, R., and Kanwal, R. P. (1999). "Singular Integral Equations," Birkhäuser, Boston.

Hochstadt, H. (1973). "Integral Equations," Wiley, New York.

Jerry, A. J. (1999). "Introduction to Integral Equations with Applications," Wiley, New York.

Kanwal, R. P. (1997). "Linear Integral Equations," Second Edition, Birkhäuser, Boston.

Muskhelishvili, N. I. (1953). "Singular Integral Equations," Noordhoff, Groningen, Holland.

Peters, A. S. (1969). Some integral equations related to Abel's equation and the Hilbert transform. Comm. Pure Appl. Math., **22,** 539–560.

Pipkin, A. C. (1991). "A Course on Integral Equations," Springer Verlag, New York.

Stakgold, I. (1968). "Boundary Value Problems of Mathematical Physics, Vol. II," Macmillan, New York.



Knots

Louis H. Kauffman
University of Illinois

I. Knot Tying and the Reidemeister Moves
II. Invariants of Knots and Links, A First Pass
III. The Jones Polynomial
IV. The Bracket State Sum
V. Vassiliev Invariants
VI. Vassiliev Invariants and Lie Algebras
VII. A Quick Review of Quantum Mechanics
VIII. Knot Amplitudes
IX. Topological Quantum Field Theory, First Steps

GLOSSARY

Coloring A labeling of a combinatorial structure (such as a graph or a knot diagram) with elements of a chosen set (called colors) according to rules that are specified in a given context.

Diagram A graphical structure intended to illustrate or embody a mathematical structure.

Fundamental group The fundamental group of a topological space is a group that is naturally defined by mappings of circles to the space. This group is used throughout topology.

Group An algebraic structure with one binary operation, satisfying the axioms that there is an identity element, every element has an inverse, and associativity. The concept of a group formalizes a general notion of symmetry.

Knot An embedding of a circle in three-dimensional space. A knot is said to be knotted if this embedding cannot be transformed topologically to a flat circle. The intent of this mathematical definition is to capture the topological aspect of a knotted rope in the space of physical experience.

Knot polynomial A method of assigning to each knot or link a polynomial that is a topological invariant of the knot or link.

Link An embedding of several disjoint circles in three-dimensional space. Two links are said to be topologically equivalent if there is an (orientation-preserving) homeomorphism of three-dimensional space that carries one link to the other one.

Quandle An algebraic invariant of knots and links closely related to the fundamental group.

Topological invariant A method of assigning a mathematical object (such as a number, a polynomial, or a group) to each of a class of topological spaces (or in the case of knots and links, to a space and collection of subspaces) so that topologically equivalent spaces (or configurations of spaces) receive the same assigned mathematical structure.

Topology The study of topological spaces—sets endowed with a notion of neighborhoods (called open sets) closed under finite intersections and arbitrary unions, such that the whole space and the empty set are both open. Topological spaces encapsulate the concept of continuity in the structure of the neighborhoods.

THIS ARTICLE constitutes an introduction to the theory of knots as it has been influenced by developments concurrent with the discovery of the Jones polynomial in 1984 and the subsequent explosion of research that followed this signal event in the mathematics of the 20th century.

I hope to give the flavor of these extraordinary events in this exposition. Even the act of tying a shoelace can become an adventure. The familiar world of string, rope, and the third dimension becomes an inexhaustible source of ideas and phenomena.

Sections 1 and 2 constitute a start on the subject of knots. Later sections introduce more technical topics. The theme of a relationship of knots with physics begins already with the Jones polynomial and the bracket model for the Jones polynomial as discussed in Section 4. Sections 5 and 6 provide an introduction to Vassiliev invariants and the remarkable relationship between Lie algebras and knot theory. The idea for the bracket model and its generalizations is to regard the knot itself as a discrete physical system and to obtain information about its topology by averaging over the states of the system. In the case of the bracket model this summation is finite and purely combinatorial. Transpositions of this idea occur throughout, involving ideas from quantum mechanics (Sections 7 and 8) and quantum field theory (Section 9). In this way knots have become a testing ground not only for topological ideas, but also for the methods of modern theoretical physics.

This article concentrates on the construction of invariants of knots and the relationships of these invariants to other mathematics (such as Lie algebras) and to physical ideas (quantum mechanics and quantum field theory). There is also a rich vein of knot theory that considers a knot as a physical object in three-dimensional space. Then one can put electrical charge on a knot and watch (in a computer) the knot repel itself to form beautiful shapes in three dimensions. Or one can think of a knot as made of thick rope and ask for an "ideal" form of the knot with minimal length-to-diameter ratio. This idea of physical knots is a current topic of research.

I. KNOT TYING AND THE REIDEMEISTER MOVES

A. Introduction

For this section it is recommended that the reader obtain a length of soft rope for the sake of direct experimentation. We begin by making some knots. In particular, we shall look at the bowline, a most useful knot. The bowline is widely used by persons who need to tie a horse to a post or their boat to a dock. It is easy and quick to make, holds exceedingly well, and can be undone in a jiffy. Figure 1 gives instructions for making the bowline. In showing the bowline we have drawn it loosely. To use it, one grabs the lower loop and pulls it tight by the upper line shown in the drawing. It tightens while maintaining the given size of the loop. Nevertheless, the knot is easily undone, as some experimentation will show.

The utility of a schema for drawing a knot is that the schema does not have to indicate all the physical properties of the knot. It is sufficient that the schema should contain the information needed to build the knot. Here is a remarkable use of language. The language of the diagrams for knots implicitly contains all their topological and physical properties, but this information may not be easily available unless the "word is made flesh" in the sense of actually building the knot from rope or cord.

Our aim is to get topological information about knots from their diagrams. Topological information is information about a knot that does not depend upon the material from which it is made and is not changed by stretching or

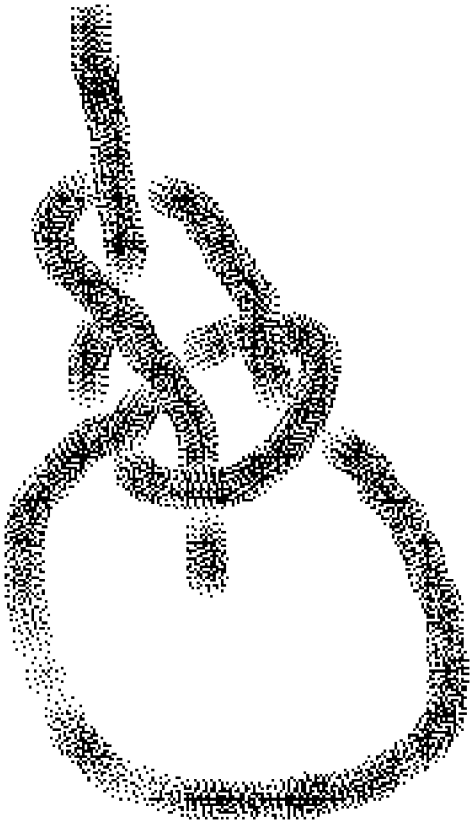


FIGURE 1 The bowline.

bending that material so long as it is not torn in the process. We do not want the knot to disappear in the course of such a stretching process by slipping over one of the ends of the rope. The knot theorist's usual convention for preventing this is to assume that the knot is formed in a closed loop of string. The trefoil knot shown in Fig. 2 is an example of such a closed, knotted loop.

A knot presented in closed-loop form is a robust object, capable of being pushed and twisted into many topologically equivalent forms. For example, the knot shown in Fig. 3 is topologically equivalent to the trefoil shown in Fig. 2.

The existence of innumerable versions of a given knot or link gives rise to a mathematical problem. To state that a loop is knotted is to state that nowhere among the infinity of forms that it can take do we find an unkotted loop. Two loops are said to be (topologically) *equivalent* if it is possible to deform one smoothly into the other so that all the intermediate stages are loops without self-intersections. In this sense a loop is knotted if it is not equivalent to a simple flat loop in the plane.

B. Reidemeister Moves

The key result that makes it possible to begin a (combinatorial) theory of knots is the theorem of Reidemeister (1948, 1932), which states that two diagrams represent equivalent loops if and only if one diagram can be obtained from the other by a finite sequence of special deformations called the *Reidemeister moves*. I shall illustrate these moves in a moment. The upshot of Reidemeister's theorem is that the topological problems about knots can all be formulated in terms of knot diagrams.

There is a famous philosophy of mathematics called "formalism" in which mathematics is considered to be

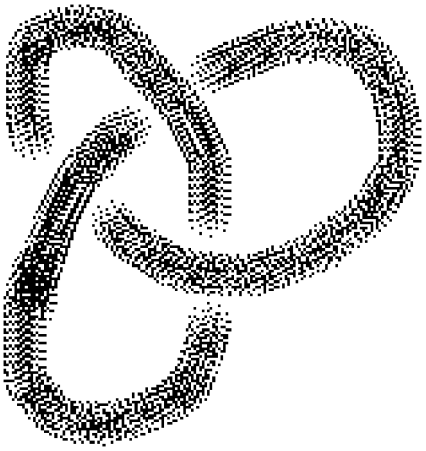


FIGURE 2 The trefoil as closed loop.

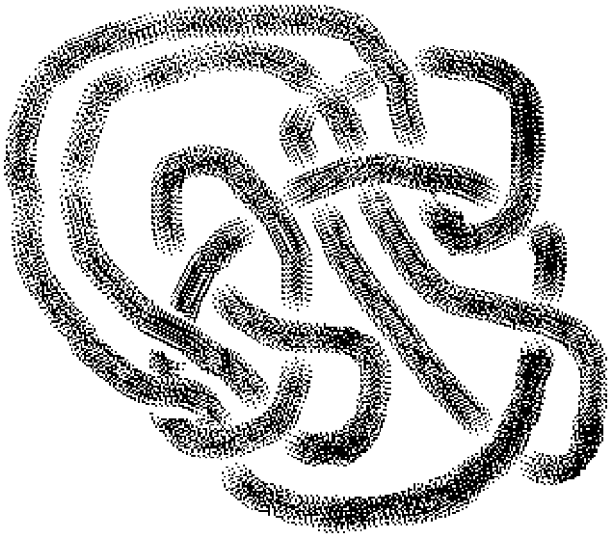


FIGURE 3 Deformed trefoil.

a game played with symbols according to specific rules. Knot theory, done with diagrams, illustrates the formalist idea very well. In the formalist point of view a specific mathematical game (formal system) can itself be an object of study for the mathematician. Each particular game may act as a coordinate system, illuminating key aspects of the subject. One can think about knots through the model of the diagrams. Other models (such as regarding the knots as specific kinds of embeddings in three-dimensional space) are equally useful in other contexts. As we shall see, the diagrams are amazingly useful, allowing us to pivot from knots to other ideas and fields and then back to topology again.

The Reidemeister moves are illustrated in Fig. 4. The moves shown in Fig. 4 are intended to indicate changes that are made in a larger diagram. These changes modify the diagram only locally as shown in the figure. Figure 5 shows a sequence of Reidemeister moves from one diagram for a trefoil knot to another. In this illustration we performed two instances of the second Reidemeister move in the first step and a combination of the second move and the third move in the second step, and used "move zero" (a topological rearrangement that does not change any of the crossing patterns) in the last step. Move zero is as important as the other Reidemeister moves, but since it does not change any essential diagrammatic relationships it is left in the background of the discussion.

C. Knots as Analog Computers

We end this section with one more illustration. This time we take the bowline and close it into a loop. A deformation then reveals that the closed-loop form of the bowline

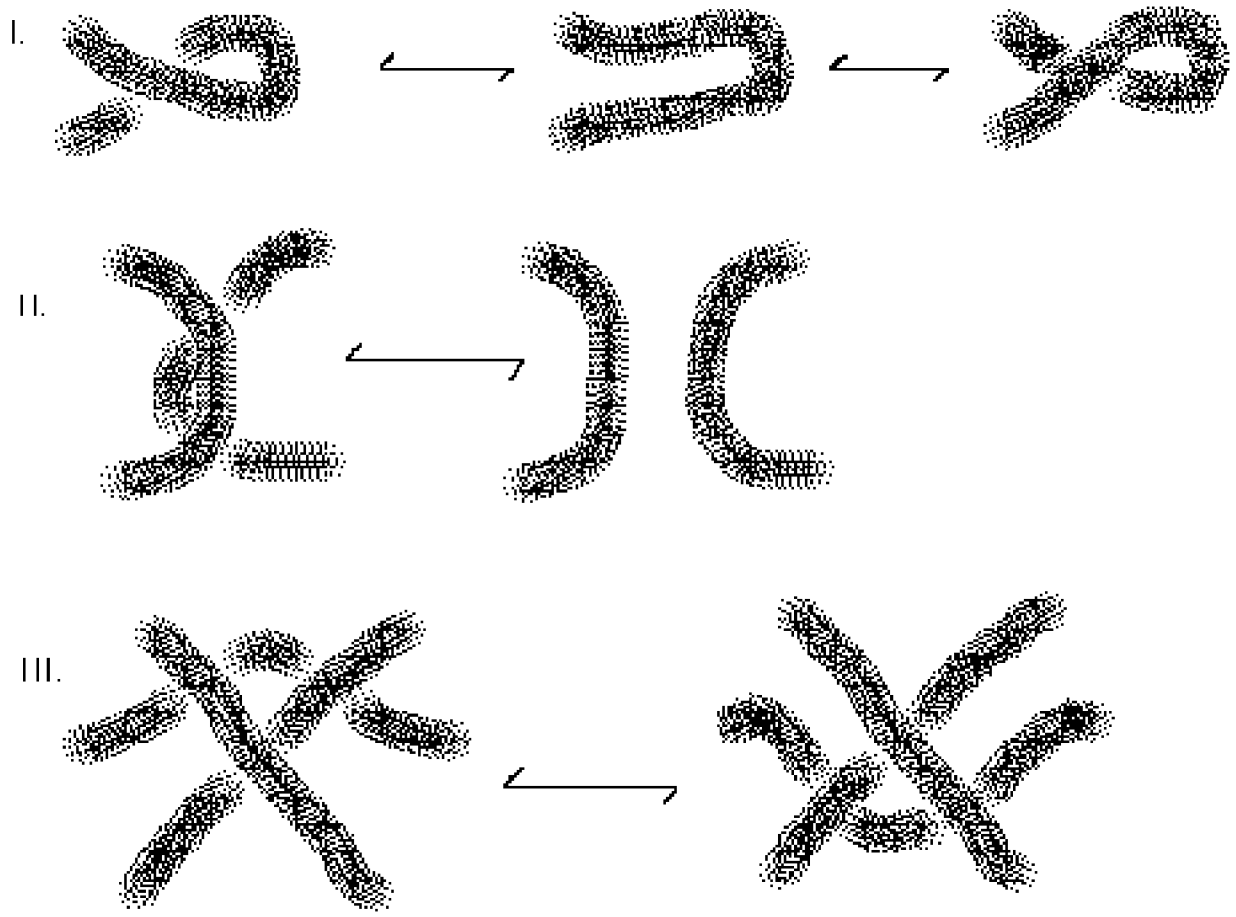


FIGURE 4 Reidemeister moves.

is topologically equivalent to two trefoils clasping one another, as shown in Fig. 6.

This deformation was discovered by making a bowline in a length of rope, closing it into a loop, and fooling about with the rope until the nice pair of clasped trefoils appeared. Note that there is more than one way to close the bowline into a loop. Figure 6 illustrates one choice. After discovering them, it took some time to find a clear pictorial pathway from the closed-loop bowline to the clasped trefoils. The pictorial pathway shown in Fig. 6 can be easily expanded to a full sequence of Reidemeister moves. In this way the model of the the knot in real rope is an analog computer that can help to find sequences of deformations that would otherwise be overlooked.

It is a curious reversal of roles that the original physical object of study becomes a computational aid for getting insight into the mathematics. Of course this is really a two-way street. The very close fit between the mathematical model for knots and the topological properties of actual knotted rope is the key ingredient. Knots are analogous to integers. Just as we believe that objects follow the laws of arithmetic, we believe that the topological properties of knotted rope follow the laws of knot topology.

II. INVARIANTS OF KNOTS AND LINKS, A FIRST PASS

We want to be able to calculate numbers (or bits of algebra such as polynomials) from given link diagrams in such a way that these numbers do not change when the diagrams are changed by Reidemeister moves. Numbers or polynomials of this kind are called *invariants* of the knot or link represented by the diagram. If we produce such invariants, then we are finding topological information about the knot or link. The easiest example of such an invariant is the linking number of two curves, which measures how many times one curve winds around another. In order to calculate the linking number we *orient* the curves. This means that each curve is equipped with a directional arrow, and we keep track of the direction of the arrow when the curve is deformed by the Reidemeister moves. If the

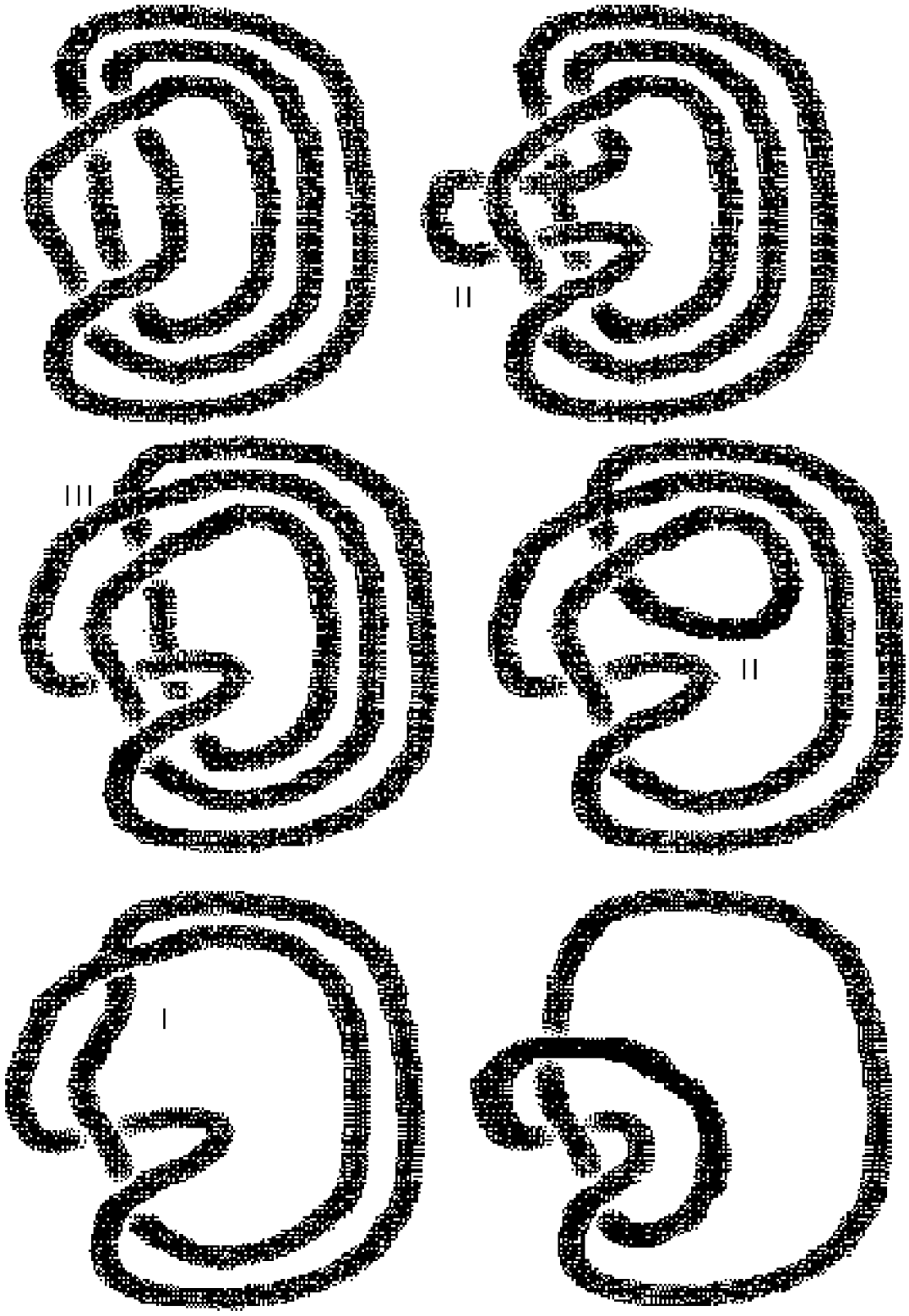


FIGURE 5 Deforming a Braid to the Trefoil Knot.

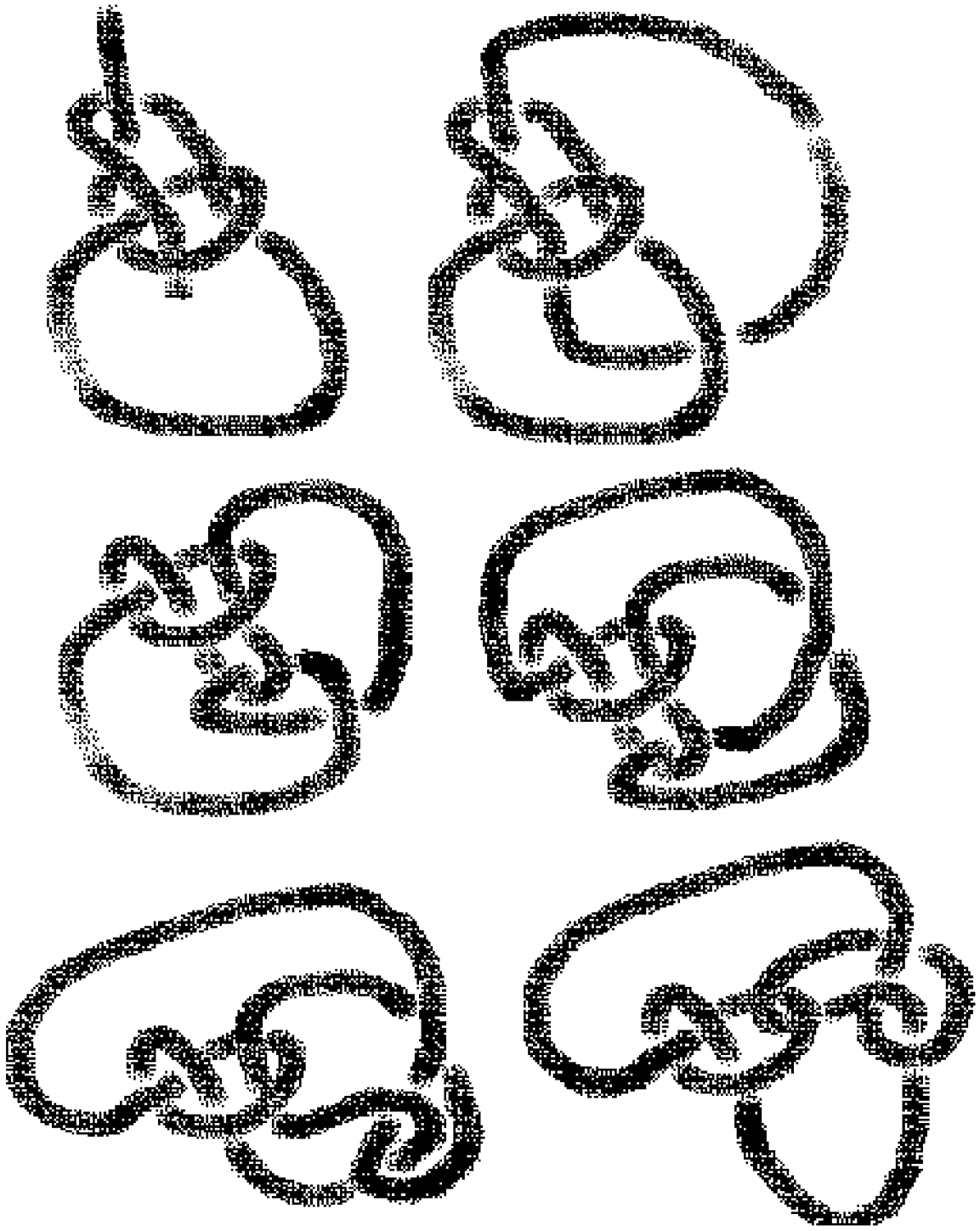


FIGURE 6 The Bowline.

curves A and B are represented by an oriented link diagram with two components, attach a sign ($+1$ or -1) to each crossing as in Fig. 7. Then the linking number $\mathrm{Lk}(A, B)$ is the sum of these signs over all the crossings of A with B.

Of course, two singly linked rings receive linking number equal to $+1$ or -1 as shown in Fig. 8.

It can be shown that the linking number is invariant under the Reidemeister moves. That is, if we take a given diagram D (representing the curves A and B) and change it



FIGURE 7 Crossing Signs.

to a new diagram E by applying one of the Reidemeister moves, then the linking number calculation for D will be the same as the calculation for E. The calculation is unaffected by the first Reidemeister move because self-crossings of a single curve do not figure in the calculation of the linking number. The second Reidemeister move either creates or destroys two crossings of opposite sign, and the third move rearranges a configuration of crossing without changing their signs.

With these observations we have in fact *proved* that the singly linked rings are indeed linked! There is no possible sequence of Reidemeister moves from these rings to two separated rings because the linking number of separated rings is equal to zero, not to plus or minus one.

It may seem a minor accomplishment to prove something as obvious as the inseparability of this simple configuration, but it is the first step in the successful application of algebraic topology to the study of knots and links. The linking number has a long and interesting history, and there are a number of ways to define it, many considerably more complicated than the sum of diagrammatic signs. We shall discuss some of these alternative definitions at the end of this section.

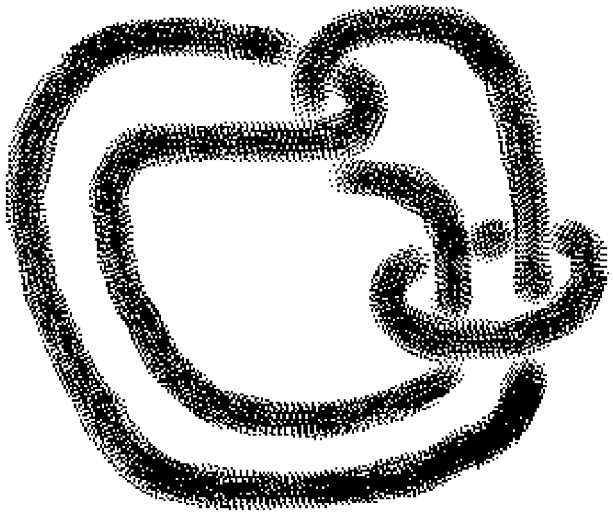


FIGURE 9 The Whitehead link.

One of the most fascinating aspects of the linking number is its limitations as an invariant. Figure 9 shows the Whitehead link, a link of two components with linking number equal to zero. The Whitehead link is indeed linked, but it requires methods more powerful than the linking number to demonstrate this fact.

Another example of this sort is the Borromean (or Ballantine) rings as shown in Fig. 10. These three rings are topologically inseparable, but if any one of them is ignored, then the other two are not linked.

Just in case these last few examples leave the reader pessimistic about the prospects of the linking number, here is a positive application. We shall use the linking number to show that the Möbius strip is not topologically equivalent to its mirror image. The Möbius strip is a circular band with a half twist in it as illustrated in Fig. 11. The Möbius

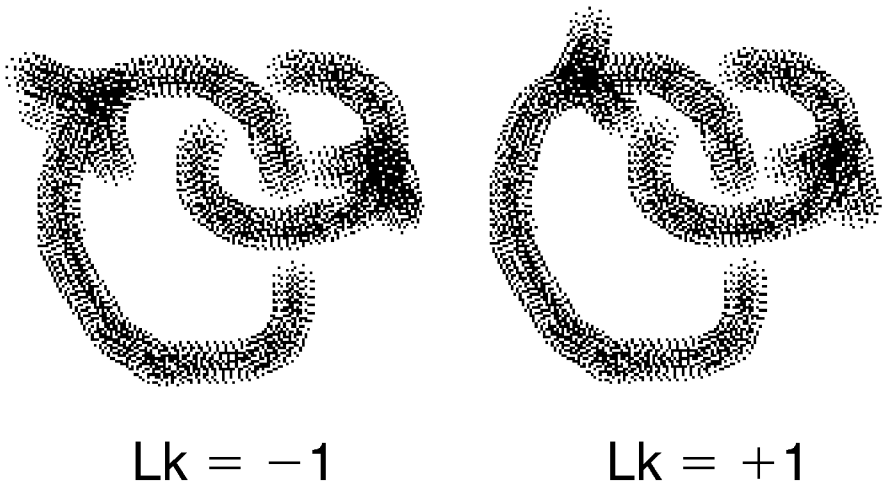


FIGURE 8 Two Linking Numbers.

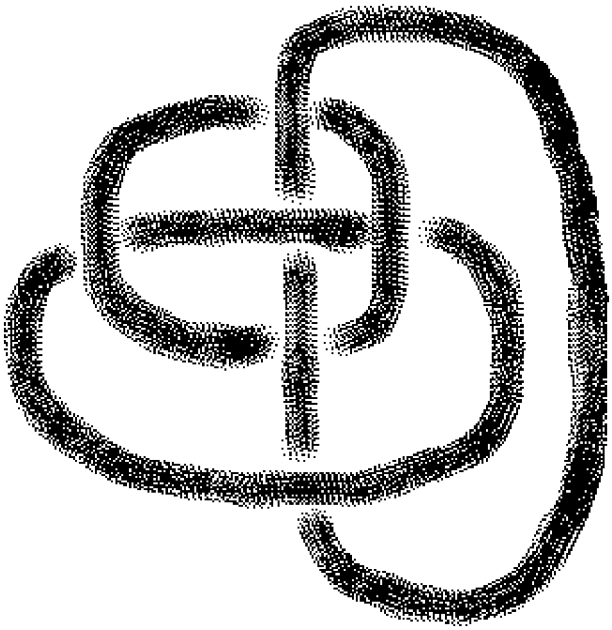


FIGURE 10 The Borromean rings.

is a justly famous example of a surface with only one side and one edge. An observer walking along the surface goes through the half-twist and arrives back where she started only to discover that she is on the other local side of the band! It requires another trip around the band to return to the original local side. As a result there is only one side to the surface in the global sense. It is as though the opposite side of the world were infinitesimally close to us by drilling into the ground, but a full circumnavigation of the globe away by external travel.

To make matters even more surprising, there are actually two Möbius bands, depending on the sense of the half twist. Call them M and M^* as illustrated in Fig. 11. If one makes these two Möbius bands from strips of paper and tries to deform one into the other without tearing the paper, one will fail (Try it!).

How can we understand the topological nature of the handedness of the Möbius band M? Draw a curve C down the center of the band M as shown in Fig. 11. Compare this curve with the space curve formed by the boundary of the band. Orient these curves in parallel and compute the linking number. It is $+1$. The very same calculation for the mirror image band M^* yields the linking number of -1.

If it were possible topologically to deform M to M^*, then the corresponding links (formed by the core curve and the boundary curve of the band) would be topologically equivalent, and hence they would have the same linking number. Since this is not the case, we conclude that M cannot be deformed to M^*.

We have shown that there are two topologically distinct Möbius bands. The two bands are mirror images of one another in the sense that each looks like the image of the other in a reflecting mirror. When an object is topologically inequivalent to its mirror image, it is said to be *chiral*. We have demonstrated the chirality of the Möbius band.

A. Three Coloring a Knot

There is a remarkable proof that the trefoil knot is knotted. This proof goes as follows: Color the three arcs of the trefoil diagram with three distinct colors, say red, blue, and purple (Fig. 12). Note that in the standard trefoil diagram three distinct colors occur at each crossing. Now adopt the following coloring rule:

- Either three colors or exactly one color occur at any crossing in the colored diagram.

Call a diagram *colored* if its arcs are colored and they satisfy this rule. Note that the standard unknot diagram is colored by simply assigning one color to its circle. A coloring does not necessarily have three colors on a given diagram. Call a diagram *three-colored* if it is colored and three colors actually appear on the diagram.

Theorem. *Every diagram that is obtained from the standard trefoil diagram by Reidemeister moves can be three-colored. Hence the trefoil diagram represents a knot.*

Rather than write a formal proof of this theorem, we illustrate the coloring process in Figs. 13 and 14. Each time a Reidemeister move is performed, it is possible to extend the coloring from the original diagram to the diagram that is obtained from the move. These extensions of colorings involve only local changes in the colorings of the original diagrams. The best way to see that this proof works is to do a few experiments. Figures 13 and 14 are a start.

Note that in the case of the second move performed in the simplifying direction, although a color is lost in the arc that disappears under the move, this color must appear elsewhere in the diagram or else it is not possible for the two arcs in the move to have different colors (since there is a path along the knot from one local arc to the other). Thus three-coloration is preserved under Reidemeister moves, whether they make the diagram simpler or more complicated. As a result, every diagram for the trefoil knot can be colored with three colors according to our rules. This proves that the trefoil is knotted, since an unknotted trefoil would have a simple circle among its diagrams, and the simple circle can be colored with only one color.

B. The Quandle and the Determinant of a Knot

There is a wide generalization of this coloring argument. We shall replace the colors by arbitrary labels for the arcs

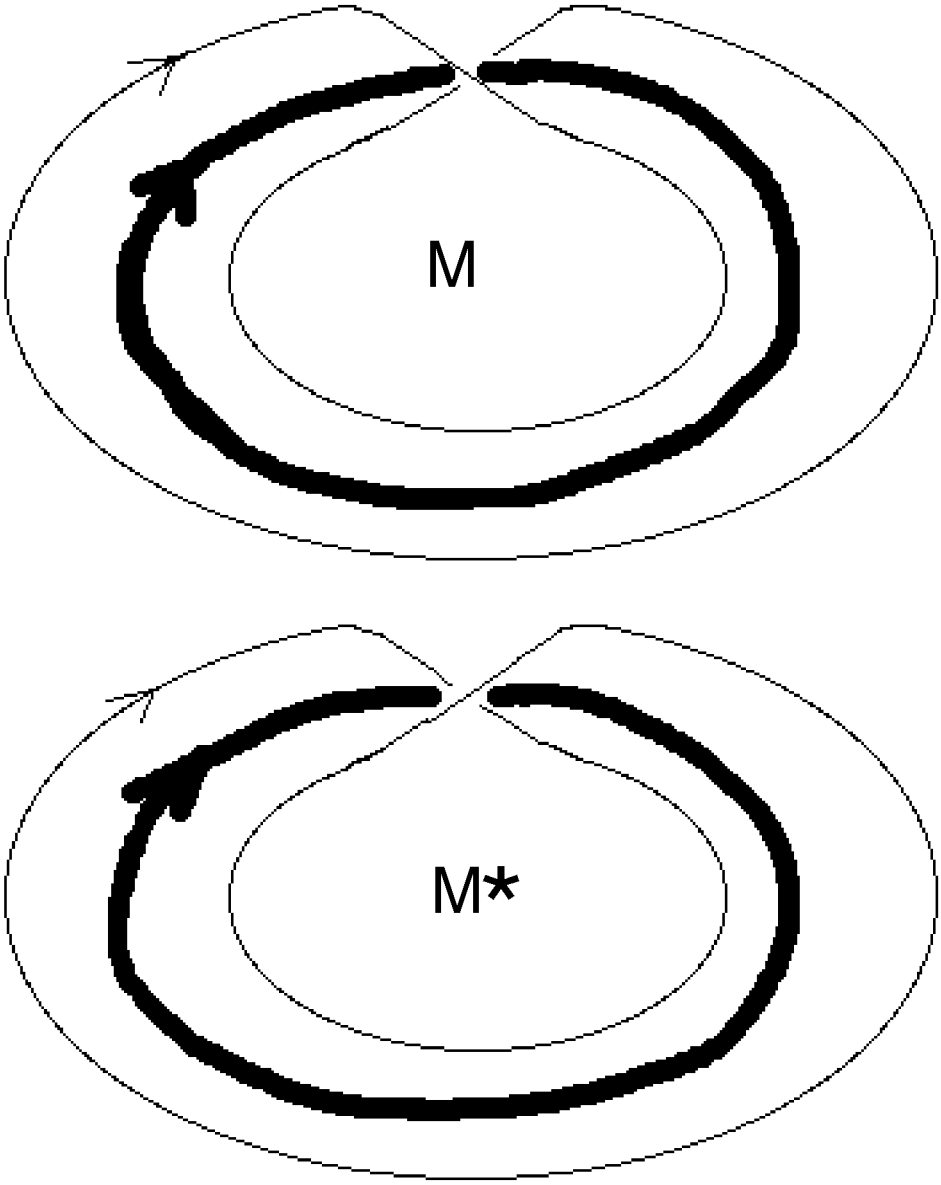


FIGURE 11 Möbius and mirror Möbius.

in the diagram and replace the coloring rule by a method for combining these labels. It turns out that a good way to articulate such a rule of combination is to make the label on one of the undercrossing arcs at a crossing a product (in the sense of this new mode of combination) of the labels of the other two arcs. In fact, we shall assume that this product operation depends upon the orientation of the arcs as shown in Fig. 15.

In Fig. 15 we show how a label A on an undercrossing arc combines with a label B on an overcrossing arc to form $C = A^*B$ or $C = A^{\#}B$ depending upon whether the overcrossing arc is oriented to the left or to the right for an observer facing the overcrossing line and standing on the arc labeled A.

This operation depends upon the orientation of the line labeled B, so that A^*B corresponds to B pointing to the right for an observer approaching the crossing along A, and $A^{\#}B$ corresponds to B pointing to the left for the same observer. All of this is illustrated in Fig. 15.

The binary operations * and $^{\#}$ are not necessarily associative. For example, our original color assignments of R (red), B (blue), and P (purple) for the trefoil knot correspond to products $\mathrm{R}^*\mathrm{R} = \mathrm{R}$, $\mathrm{B}^*\mathrm{B} = \mathrm{B}$, $\mathrm{P}^*\mathrm{P} = \mathrm{P}$, $\mathrm{R}^*\mathrm{B} = \mathrm{P}$,

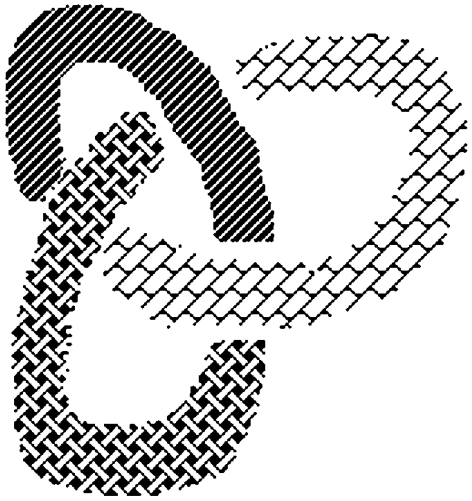


FIGURE 12 The three-colored trefoil.

$B^*P = R$, $P^*R = B$. Then $R^*(B^*P) = R^*R = R$ and $(R^*B)^*P = P^*P = P$.

We shall insist that these operations satisfy a number of identities so that the labeling is compatible with the Reidemeister moves.

Figure 16 illustrates the diagrammatic justification for the following algebraic rules about * and $^\#$. An algebraic system satisfying these rules is called a *quandle*.

1. $A^*A = A$ and $A^\#A = A$ for any label A.
2. $(A^*B)^\#B = A$ and $(A^\#B)^*B = A$ for any labels A and B.
3. $(A^*B)^*C = (A^*C)^*(B^*C)$ and $(A^\#B)^\#C = (A^\#C)^\#(B^\#C)$ for any labels A, B, C.

These rules correspond, respectively, to the Reidemeister moves 1–3. Labelings that obey these rules can

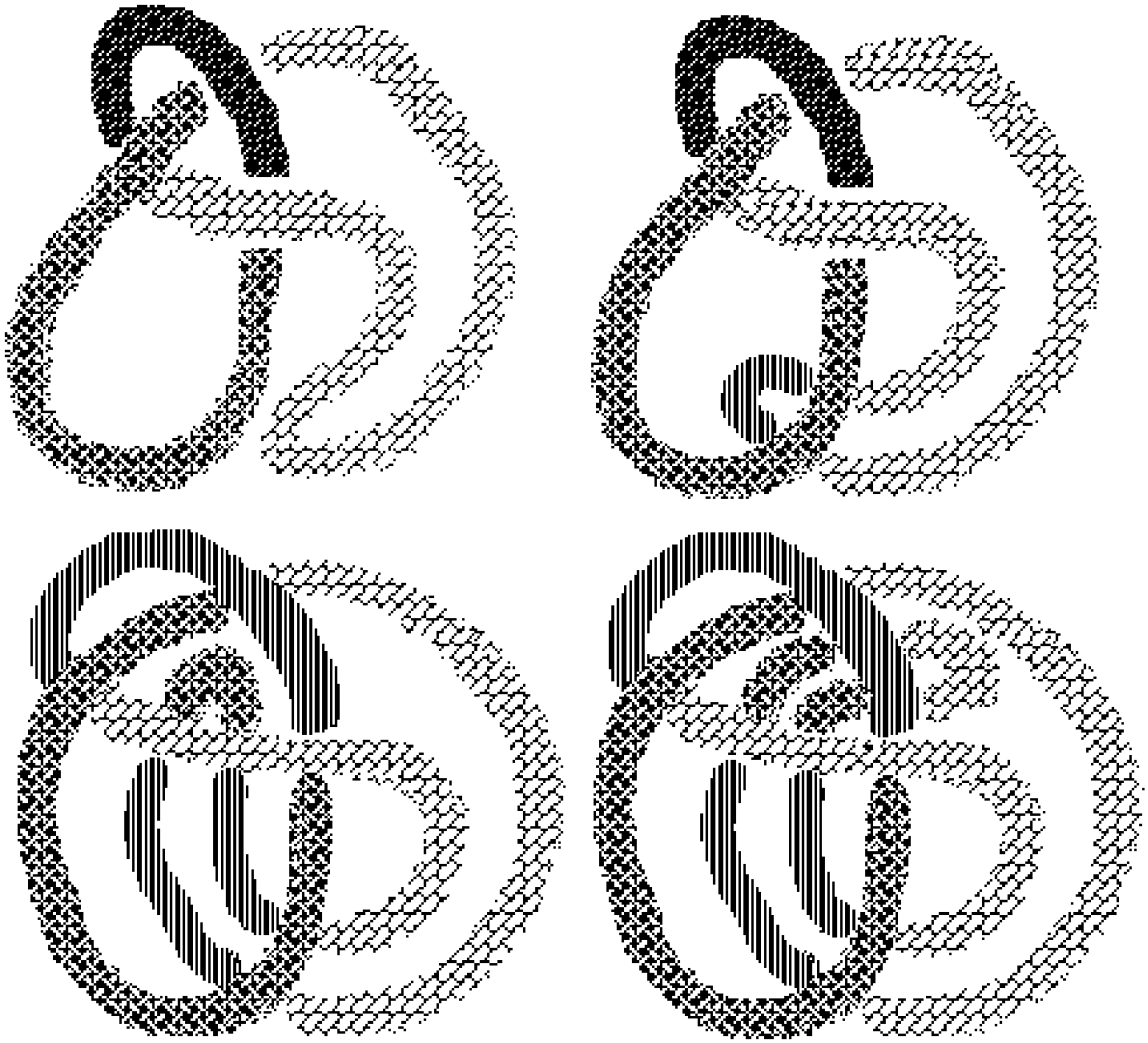


FIGURE 13 Inheriting coloring under the type 2 move.

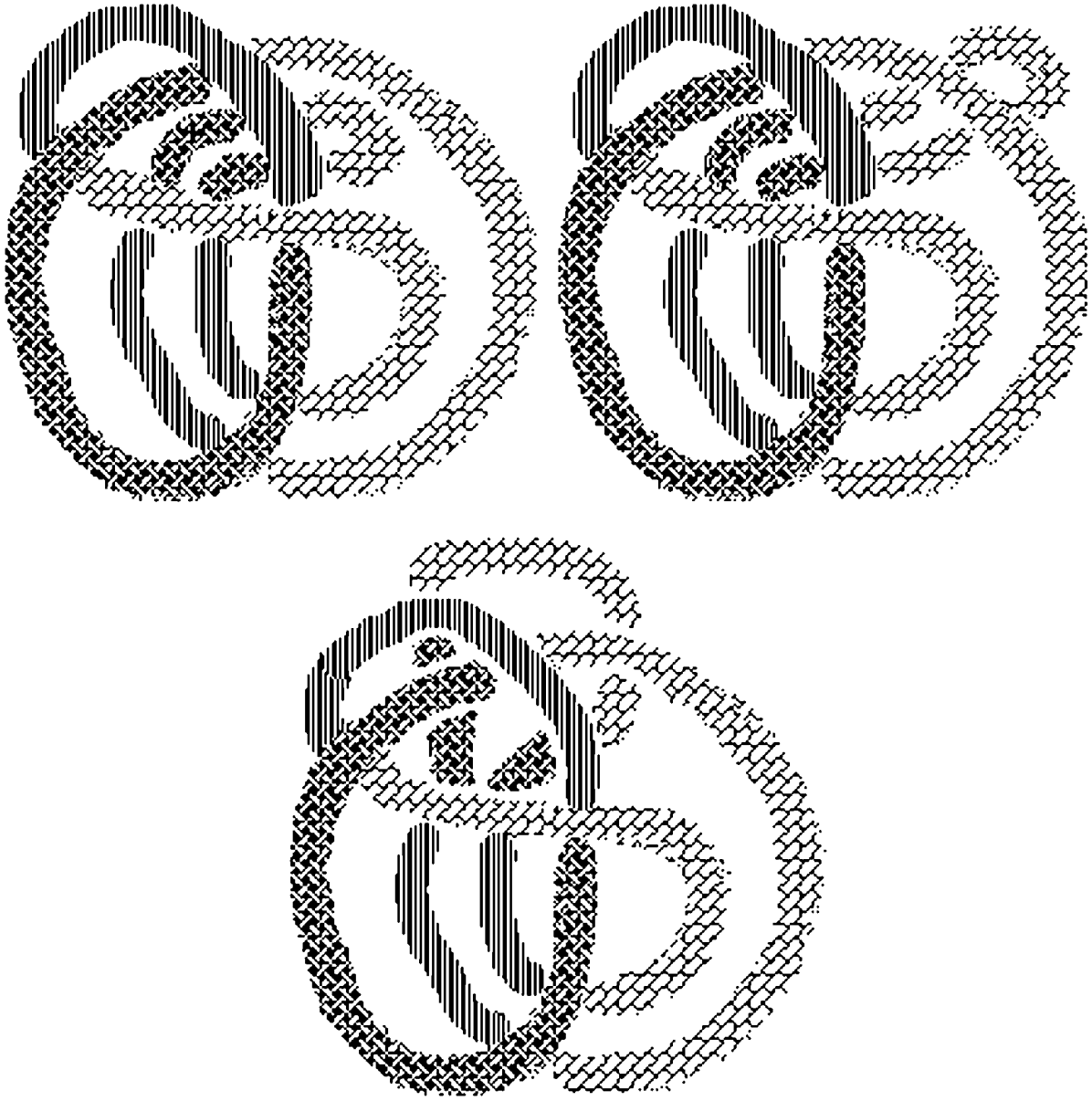


FIGURE 14 Coloring under type 2 and 3 moves.

be handled just like the three-coloring that we have already studied. In particular, a given labeling of a knot diagram means that it is possible to label (satisfying the rules given above for the labels) any diagram that is related to it by a sequence of Reidemeister moves. However, not all the labels will necessarily appear on every related diagram, and for a given coloring scheme and a given knot, certain special restrictions can arise.

To illustrate this, consider the color rule for numbers: $A^*B = A^\#B = 2B - A$. This satisfies the axioms, as is easy to see. Figure 17 shows how, on the trefoil, such a coloring must obey the equations $A^*B = C$, $C^*A = B$, $B^*C = A$. Hence $2B - A = C$, $2A - C = B$, $2C - B = A$. For example, if $A = 0$ and $B = 1$, then $C = 2B - A = 2$ and $A = 2C - B = 4 - 1 = 3$. We need $3 = 0$. Hence this system of equations will be satisfied for appropriate labelings in $\mathrm{Z}/3\mathrm{Z}$, the integers modulo three, a modular number system.

For the reader unfamiliar with the concept of modular number system, consider a standard clock whose dial is labeled with the hours 1, 2, 3, . . . , 11, 12. We ask what time is it 4 hr past the hour of 10? The answer is 2, and one can say that in the arithmetic of this clock $10 + 4 = 2$. In

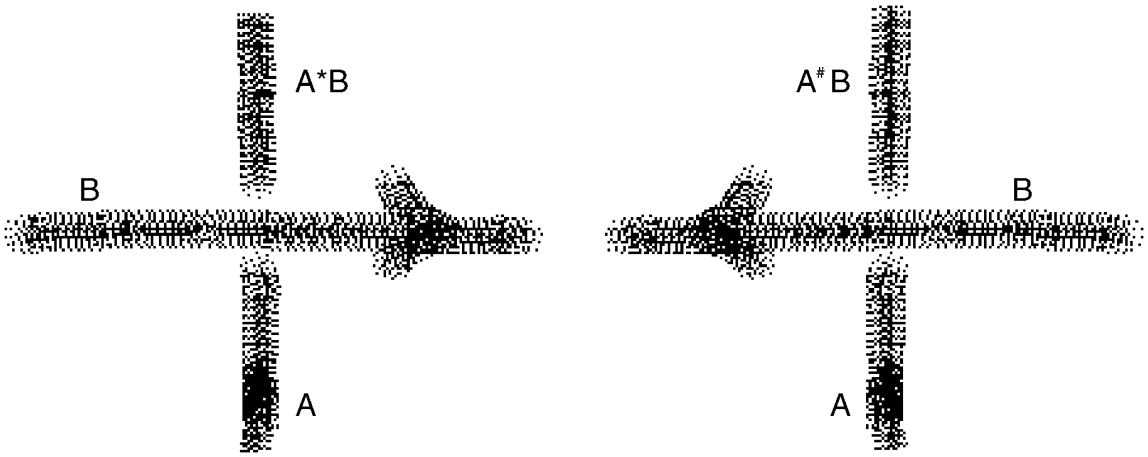


FIGURE 15 The quandle operation.

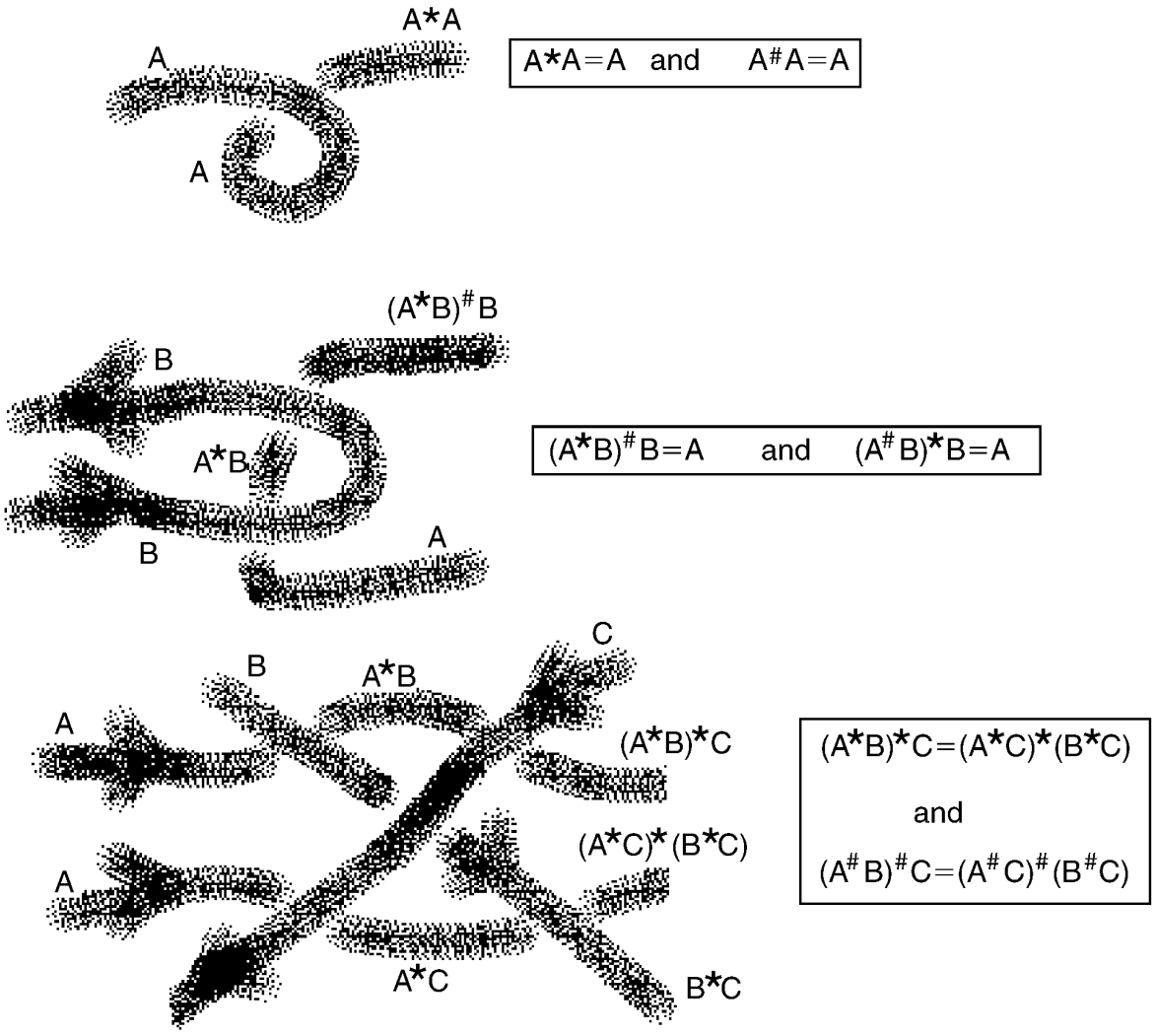


FIGURE 16 Quandle identities.

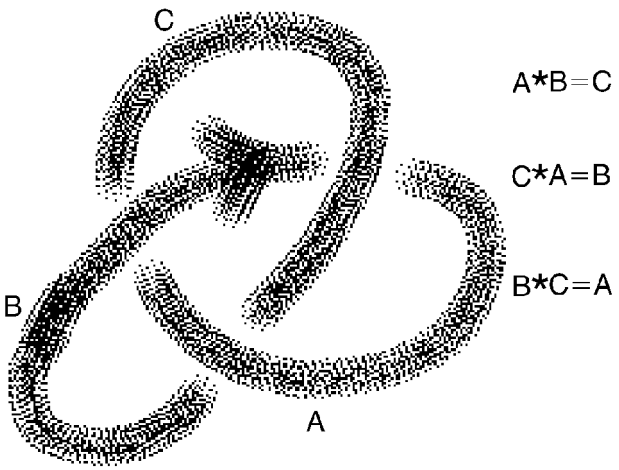


FIGURE 17 Equations for the trefoil knot.

fact $12 = 0$ in this arithmetic because adding 12 hr to the time does not change the time indicated on the clock. We work in clock arithmetic by remembering to set blocks of 12 hr to zero. One can multiply in this arithmetic as well. The square of the present time is 1 o'clock; what time is it? The answer is 7 since 7 squared is 49 and 49 is equal to 1 on the clock.

We say that the clock represents a modular number system Z/12Z with modulus 12. It is convenient in mathematics to think of the elements of Z/12Z as the set $\{0, 1, 2, \ldots, 11\}$. Since $0 = 12$ this takes care of all the hours.

In general we can consider Z/nZ, where n is any positive integer modulus. The resulting modular number system has elements $\{0, 1, 2, \ldots, n-1\}$ and is handled just as though there were a clock with n hours rather than 12.

In such a system one says that $x = y \pmod{n}$ if the difference between x and y is divisible by n. For example, $49 = 1 \pmod{12}$ since $49 - 1 = 48$ is divisible by 12.

The modular number system Z/3Z reproduces exactly the three coloring of the trefoil, and we see that the number 3 emerges as a characteristic of the equations associated with the knot. In fact, *3 is the value of a determinant that is associated with these equations*, and its absolute value is an invariant of the knot. For more about this construction, see Kauffman (1993, Part 1, Chapter 13).

Here is another example: For the figure eight knot E, we have that the modulus is 5. This shows that E is indeed knotted and that it is distinct from the trefoil. We can color (label) the figure eight knot with five "colors" 0, 1, 2, 3, and 4 with the rule $A^*B = 2B - A \pmod{5}$. See Fig. 18.

Note that in coloring the figure eight knot we have only used four of the five available "colors" from the set $\{0, 1, 2, 3, 4\}$. Figure 18 uses the colors 0, 1, 2, and 4. The

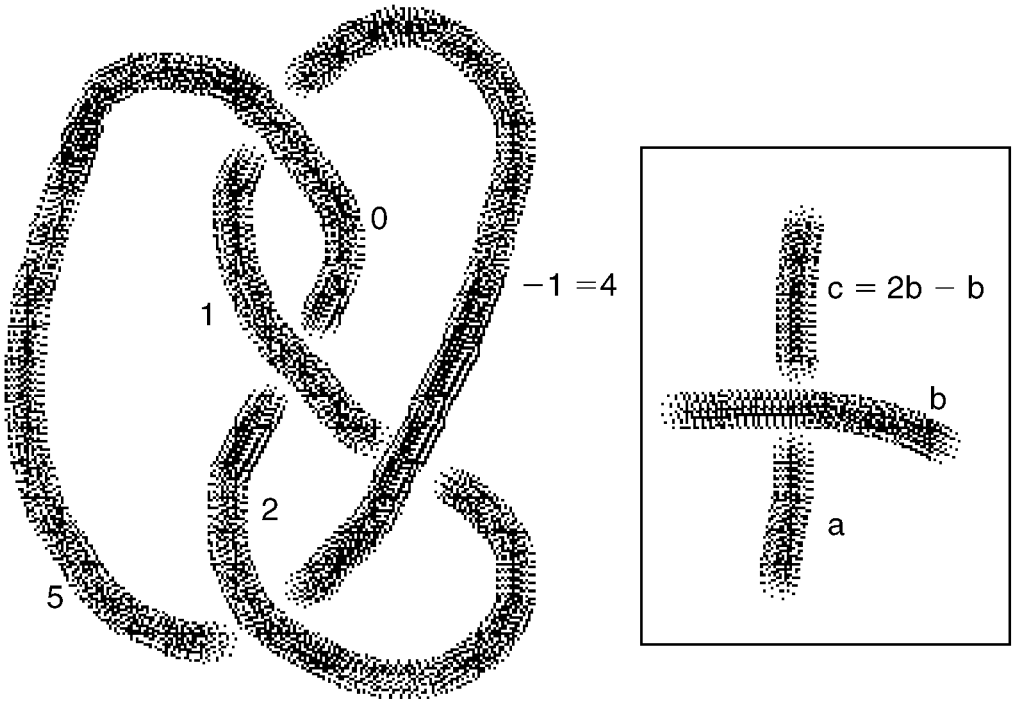


FIGURE 18 Five colors for the figure eight knot.

coloring number of a knot or link K is the least number of colors (greater than 1) needed to color it in the $2B - A$ fashion for *any* diagram of K. It is a nice exercise to verify that the coloring number of the figure eight knot is indeed four. In general the coloring number of knot or link is not easy to determine. This is an example of a topological invariant that has subtle combinatorial properties.

Other knots and links mentioned in this section can be shown to be knotted and linked by the modular method. The reader should try it for the Borromean rings and the Whitehead link.

The coloring (labeling) rules as we have formalized them can be described as axioms for an algebra associated with the knot. This is called the quandle. It has been generalized to the crystal, the interlock algebra, and the rack. The quandle is itself a generalization of the fundamental group of the knot complement.

C. The Alexander Polynomial

The modular labeling method has a marvellous generalization to the Alexander polynomial of the knot. This comes about through generalized coloring rules $A^*B = tA + (1-t)B$ and $A^\#B = t^{-1}A + (1-t^{-1})B$, where t is an indeterminate. It is a nice exercise to verify that these rules satisfy the axioms for the quandle. This algebraic structure is called the Alexander module.

The case $t = -1$ gives the rule $2B - A$ that we have already considered. By coloring diagrams with arbitrary t, we obtain a polynomial that generalizes the modulus. This polynomial is the Alexander polynomial.

Alexander (1923) described it differently in his original paper, and there is a remarkable history to the development of this invariant. See, for example Crowell and Fox (1963) and Kauffman (1987b) for more information. The flavor of this relationship can be seen by doing a little experiment in labeling the trefoil diagram shown in Fig. 19. The circularity inherent in the knot diagram results in relations that must be satisfied by the module action. In Fig. 19 we see directly by labeling the diagram that if arc **1** is labeled 0 and arc **2** is labeled A, then $(t + (1-t)^2)A = 0$. In fact, $t + (1-t)^2 = t^2 - t + 1$ is the Alexander polynomial of the trefoil knot. The Alexander polynomial is an algebraic modulus for the knot.

III. THE JONES POLYNOMIAL

Our next topic describes an invariant of knots and links of quite a different character than the modulus or the Alexander polynomial of the knot. It is a "polynomial" invariant of knots and links discovered by Jones (1985). Jones' invariant, usually denoted $V_K(t)$, is a polynomial in the variable $t^{1/2}$ and its inverse $t^{-1/2}$. One says that $V_K(t)$ is a Laurent polynomial in $t^{1/2}$. Superficially, the Jones polynomial appears to be just another polynomial invariant of knots and links, somewhat similar to the Alexander polynomial. When I say that the Jones polynomial is of a different

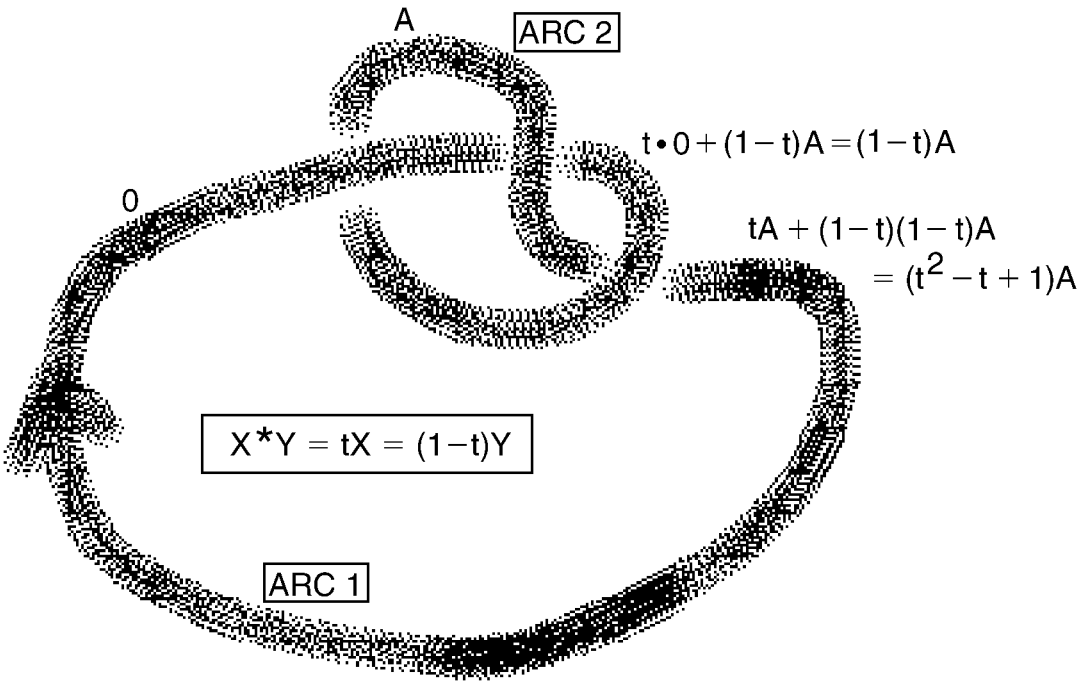


FIGURE 19 Alexander polynomial of the trefoil knot.

character, I mean something deeper, and it will take a little while to explain this difference. A little history will help.

The Alexander polynomial was discovered in the 1920s and until 1984 no one had found another polynomial invariant of knots and links that was not a simple generalization of the Alexander polynomial. Jones discovered a new polynomial invariant of knots and links that had some very remarkable properties. The Alexander polynomial cannot detect the difference between any knot and its mirror image. What made the Jones polynomial such an exciting discovery for knot theorists was the fact that it could detect the difference between many knots and their mirror images. Later, other properties began to emerge. It became a key tool in proving properties of alternating links (and generalizations) that had been conjectured since the last century (see, eg., Murasugi, 1987a, b, Kauffman 1987a and Thistlewaite 1987).

It turns out the the Jones polynomial is intimately related to a number of topics in mathematical physics. Curiously, it is actually easier to define and verify the properties of the Jones polynomial than for any other invariant in the theory of knots (except of course the linking number). We shall devote this section to the defining properties of the Jones polynomial, and later sections to the relationships with physics.

Here is a set of axioms for the Jones polynomial. The polynomial was not discovered in the form of these axioms. The axioms are in a format analogous to the framework that Conway (1970; Kauffman, 1980, 1983) discovered for the Alexander polynomial. I am starting with these axioms because they give a quick access to the polynomial and to sample computations.

1. Axioms for the Jones Polynomial

1. If two oriented links K and K' are ambient isotopic, then $V_K(t) = V_{K'}(t)$.
2. If U is an unknotted loop, then $V_U(t) = 1$.
3. If K_+, K_-, and K_0 are three links with diagrams that differ only as shown in the neighborhood of a single crossing site for K_+ and K_- (see Fig. 20), then $t^{-1}V_{K_+}(t) - tV_{K_-}(t) = (t^{1/2} - t^{-1/2})V_{K_0}(t)$.

The axioms for $V_K(t)$ are a consequence of Jones' original definition of his invariant. He was led to this invariant by a trail that began with the study of von Neumann algebras (a branch of algebra directly related to quantum theory and to statistical mechanics) and ended in braids, knots, and links. The Jones polynomial has a distinctly different flavor from the Conway–Alexander polynomial, even though it can be axiomatized in a very similar way. In fact, this similarity of axiomatics points to a common generalization [the Homfly(Pt) polynomial] and to another generalization (the Kauffman polynomial) and then to further generalizations in the connection with statistical mechanics (see, e.g., Kauffman, 1989).

To this date no one has found a knotted loop that the Jones polynomial does not declare to be knotted. Thus one can make the following conjecture:

Conjecture. If a single-component loop K is knotted, then $V_K(t)$ is not equal to one.

While it is possible that the Jones polynomial is able to detect the property of being knotted, it is not a complete classifier for knots. There are inequivalent pairs of knots that have the same Jones polynomial. Such a pair is shown in Fig. 21. These two knots, the Kinoshita–Terasaka knot and the Conway knot, have the same Jones polynomial but are different topologically. Incidentally these two knots are examples of knots whose knottedness cannot be detected by the Alexander polynomial.

Let us use the axioms to compute the Jones polynomial for the trefoil knot. To this end, there is a useful device called the *skein tree*. A skein tree is obtained from a given knot or link diagram by recording the knots and links obtained from this diagram by smoothing or switching crossings. Each node of the tree is a knot or link. The nodes farthest from the original knot or link are unknotted or unlinked. Such a tree can be produced from a given knot or link by using the fact that any knot or link diagram can

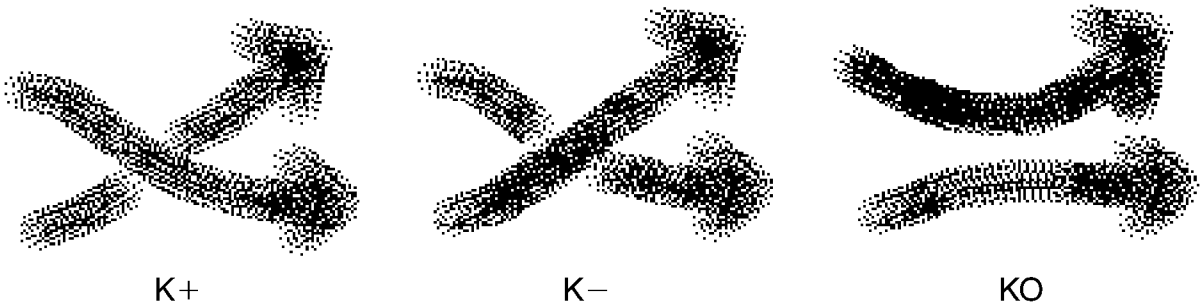


FIGURE 20 Skein Triple.

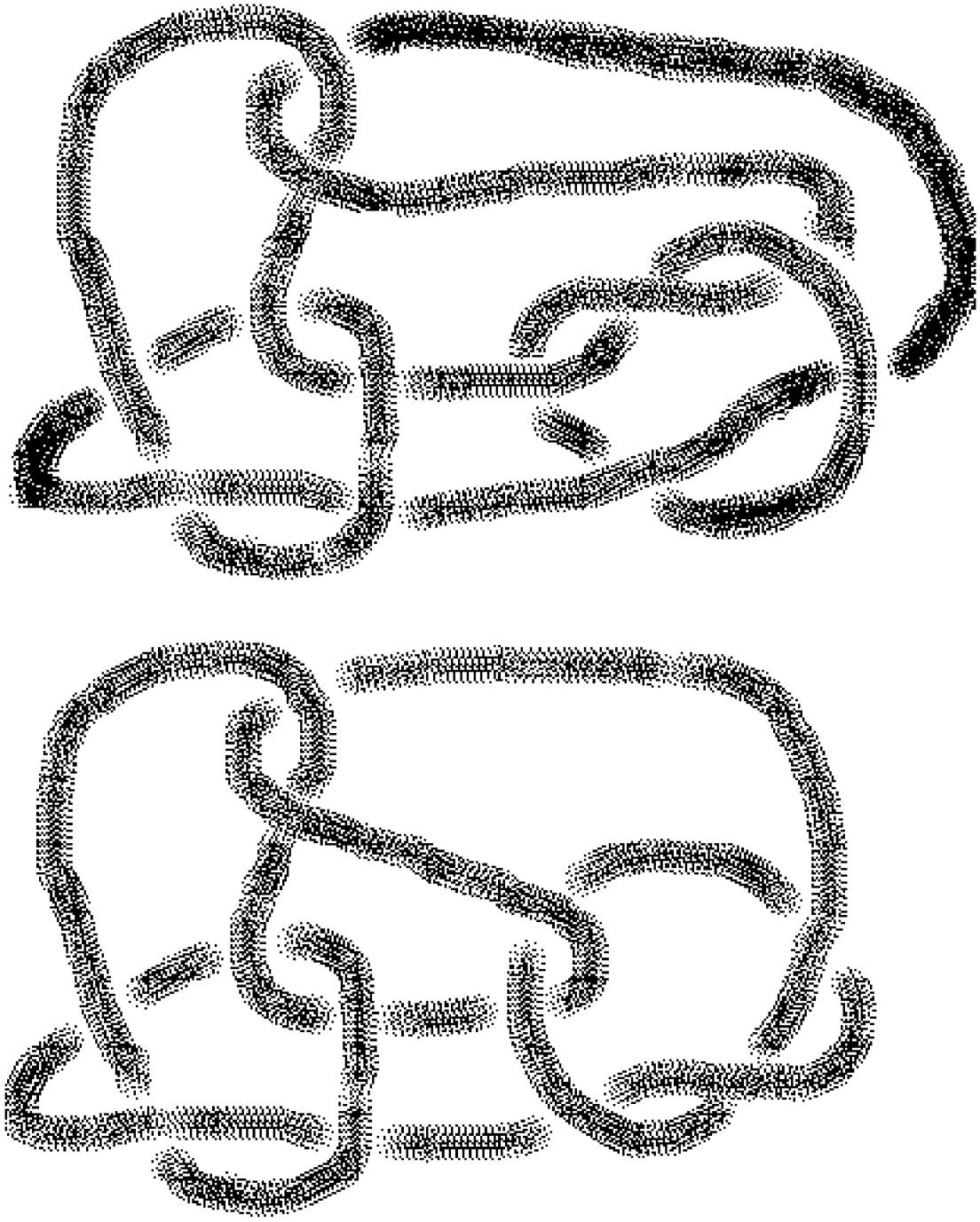


FIGURE 21 (Top) Conway Knot. (Bottom) Kinoshita-Terasaka Knot. Two Knots with trivial Alexander Polynomial.

be transformed into an unknotted (unlinked) diagram by a sequence of crossing switches.

Figure 22 illustrates a "standard unknot diagram." This diagram is drawn by starting at the arrowhead in the figure and tracing the diagram in such a way that one always draws an overcrossing before drawing an undercrossing. This is the easiest possible knot diagram to draw since one never has to make any corrections: one just passes under when one wants to cross an an already created line in the diagram. Standard unknot diagrams are always unknotted. Trying the one in Fig. 22 will show why this is so.

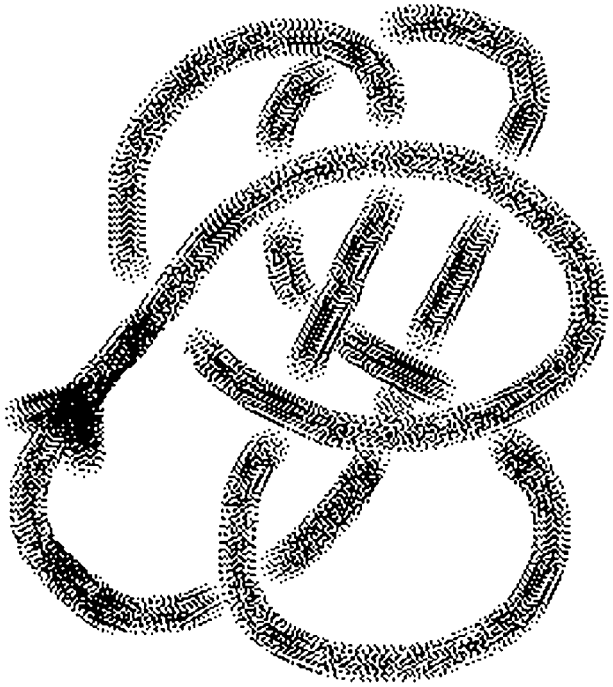


FIGURE 22 A standard unknot.

Using the fact that standard unknot diagrams are available, we can use the difference between a given diagram K and a standard unknot with the same plane projection to give a procedure for switching crossings to unknot the diagram K. This switching procedure can be used to produce a skein tree for calculating the Jones polynomial of K.

Figure 23 shows a skein tree for the computation of the Jones polynomial of the trefoil knot. The tree reduces the calculation of the Jones polynomial of the trefoil diagram to the calculation of certain unknots and unlinks. In order to see how to calculate an unlink it is useful to observe the behavior of the axioms in this case:

$$t^{-1}V_{U_+} - tV_{U_-} = (t^{1/2} - t^{-1/2})V_{U_0}.$$

Here U^+ and U^- denote unknots with single positive and negative twists in them, respectively. U_0, obtained by smoothing the crossing of U_+ or U_-, is an unlinked pair of circles with no twists. See Fig. 24.

Therefore $(t^{-1} - t)1 = (t^{1/2} - t^{-1/2})V_{U_0}$.

Hence $d = V_{U_0} = (t^{-1} - t)/(t^{1/2} - t^{-1/2}) = -(t^{1/2} + t^{-1/2})$.

Thus, we see that an extra unknotted component of the link multiplies the invariant by $d = -(t^{1/2} + t^{-1/2})$. For T the trefoil knot, U the unknot, and L the link of two unknotted circles as shown in Fig. 23, we find that

$$t^{-1}V_T - tV_U = (t^{1/2} - t^{-1/2})V_L,$$
$$t^{-1}V_L - td = (t^{1/2} - t^{-1/2})V_U.$$

Thus, $V_L = t(td + (t^{1/2} - t^{-1/2})) = -t^{5/2} - t^{1/2}$.
Hence

$$\begin{aligned} V_T &= t(t + (t^{1/2} - t^{-1/2})V_L) \\ &= t(t + (t^{1/2} - t^{-1/2})(-t^{5/2} - t^{1/2})) \\ &= t(t - t^3 - t + t^2 + 1) = t(-t^3 + t^2 + 1) \\ &= -t^4 + t^3 + t. \end{aligned}$$

The same calculation applied to the mirror image T^* (obtained by reversing all the crossings of T) of the trefoil yields the invariant $V_{T^*} = -t^{-4} + t^{-3} + t^{-1}$. This shows how the Jones polynomial discriminates between the trefoil and its mirror image, thereby proving that there is no ambient isotopy from T to T^*.

This method of calculating the Jones polynomial from its axioms does not tell us why the invariant works. It is possible to analyze this method of calculation and show that it does not depend upon the choices that one makes in the process and that it gives topological information about the knot or link in question. There is a different way to proceed that leads to a very nice formula for the Jones polynomial as a sum over "states" of the diagram. In this formulation, the polynomial is well defined from the beginning, and we can see the topological invariance arise in the course of adjusting certain parameters of a well-defined function. Our next topic is this state summation model for the Jones polynomial.

IV. THE BRACKET STATE SUM

In the last section we gave axioms for the Jones polynomial and showed how to compute it by skein calculations from these axioms. In this section we shall show one way to prove that the Jones polynomial is well defined by these axioms and that it is an invariant of ambient isotopy of links in three-dimensional space. In order to accomplish this aim, we shall give a different definition of the polynomial as a certain summation over combinatorial configurations associated with the given link diagram. This summation will be called a *state summation model* for the Jones polynomial.

In fact, we shall first construct a state summation called the *bracket polynomial* and then explain how to modify the bracket polynomial to obtain the Jones polynomial. The bracket polynomial has a rather natural development, and is defined for unoriented link diagrams.

We work with diagrams for unoriented knots and links. To each crossing in the diagram assign two *local states*

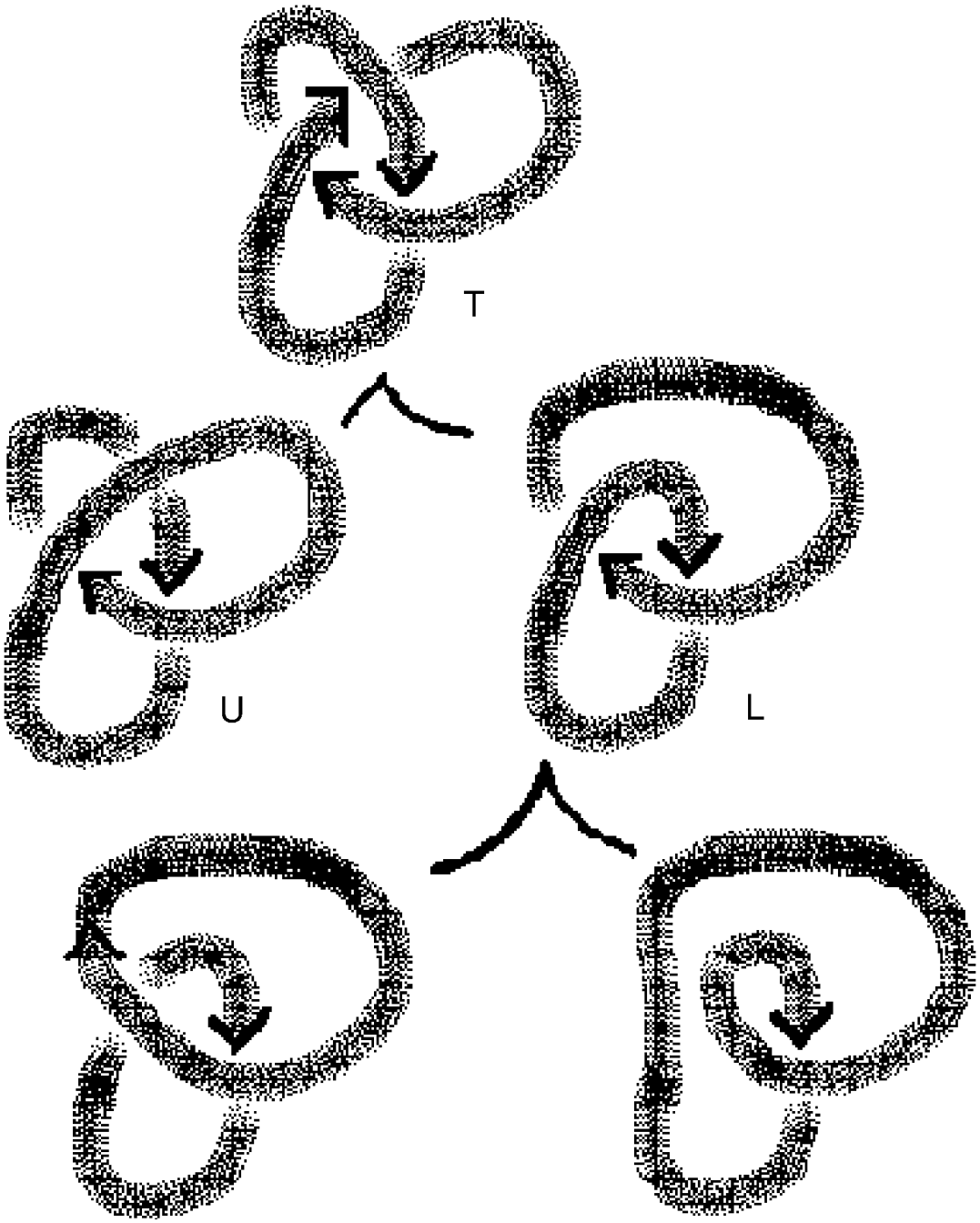


FIGURE 23 Trefoil skein tree.

with labels A and B as shown in Fig. 25. (In the A state the regions swept out by a counterclockwise turn of the overcrossing line are joined. In the B state the regions swept out by a clockwise turn of the overcrossing line are joined.)

A state S of a diagram K consists in a choice of local state for each crossing of K. Thus a diagram with N crossings will have $2U$ states. Two states S and S' of the trefoil diagram are indicated in Fig. 26. States are evaluated in two ways. These ways are denoted by $\langle K \mid S \rangle$ and by $\|S\|$. The *norm* of the state S, $\|S\|$, is defined to be one less than the number of closed curves in the plane described by S. In the example in Fig. 26, we have $\|S\| = 1$ and $\|S'\| = 0$. The evaluation $\langle K \mid S \rangle$ is defined to be the product of all the state labels (A and B) in the state. Thus, in Fig. 26, we have, $\langle K \mid S \rangle = A^3$ and $\langle K \mid S' \rangle = A^2 B$.

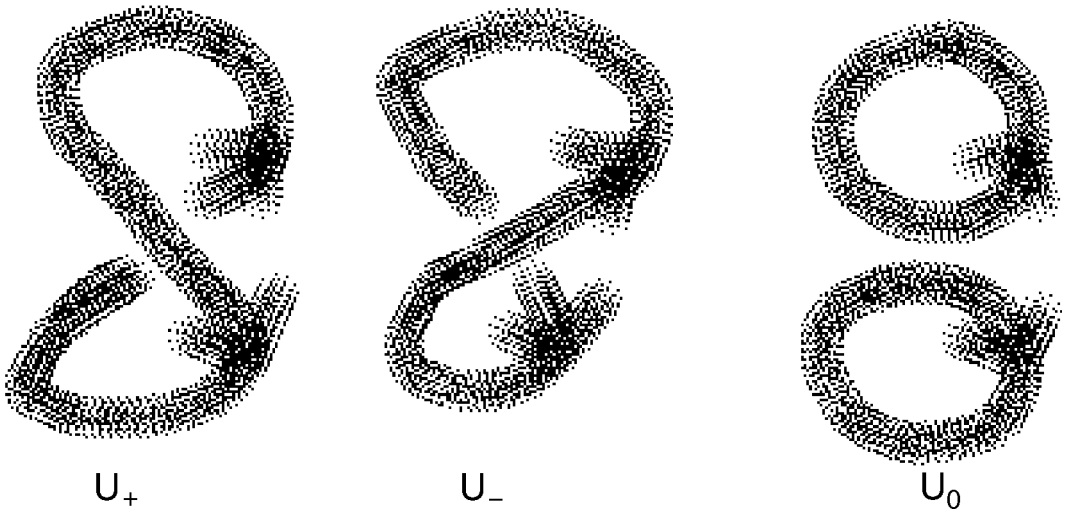


FIGURE 24 Skein Triple for Au Unknot.

Taking variables A, B, and d, we define the state summation associated with a given diagram K by the formula

$$\langle K \rangle = \sum \langle K \mid S \rangle d^{\|S\|}.$$

In other words, for each state we take the product of the labels for that state multiplied by d raised to the number of loops in the state. $\langle K \rangle$ is the summation of this state evaluation over all the states in the diagram for K. (The notation $\sum$ means "sum over all instances of S.") We will show that the state summation $\langle K \rangle$ is invariant under the second and third Reidemeister moves if we take $B = 1/A$ and $d = -(A^2 + B^2)$. A normalization then enables us to obtain invariance under all three Reidemeister moves, and hence topological information about knots and links. [See Kauffman (1987a) for more information about the bracket and its relationship with the Jones polynomial.] There is a great deal of topological information in the calculations that ensue from the bracket polynomial. In particular, one can distinguish many knots from their mirror images, and it is possible that the bracket calculation can detect whether a given diagram is actually knotted.

A. First Steps in Bracketology

The first constructions related to the bracket polynomial are quite elementary. There are two basic formulas that are reminiscent of the exchange relations we have already seen for the Jones polynomial. These formulas are

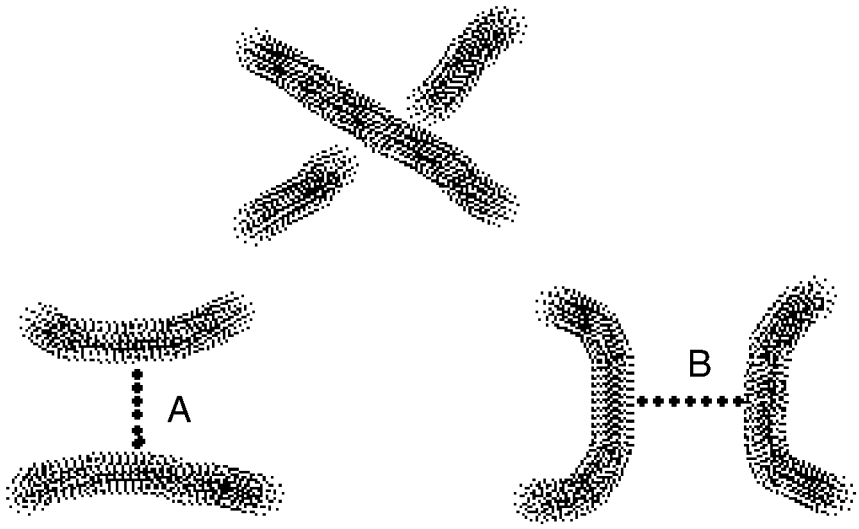


FIGURE 25 States of a Crossing.

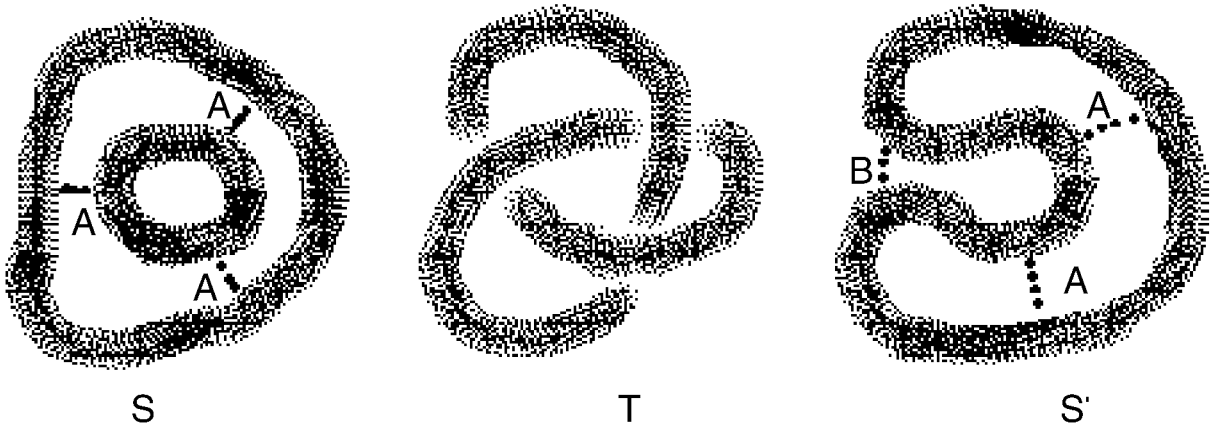


FIGURE 26 Two States of the Trefoil Knot.

as shown in Fig. 27. Here the small diagrams indicate parts of larger diagrams that are otherwise identical. Formula 1 just says that the state summation breaks up into two sums with respect to a given crossing in the diagram. In one sum, we have made a smooothing of type A at the crossing, while in the other sum we have made a smoothing of type B. The factors of A and B indicated in the formula are the contributions to the product of vertex weights from this crossing. All the rest of the two partial sums can be interpreted as bracket evaluations of the smoothed diagrams.

Formula 2 in Fig. 27 just states that an extra, simple closed curve in a diagram multiplies its bracket evaluation by the loop value d. Note that a single loop receives the value 1.

With the help of these two formulas, we can compute some basic bracket evaluations. Note that we have not yet specialized the variables A, B, and d. We shall analyze just what specialization will produce an invariant of knots and links. The advantage to having set up the definition of the bracket polynomial in this way is exactly that we have a method of labeling link diagrams with algebra, and it is possible to then adjust the evaluation so that it is invariant under Reidemeister moves. To this end, the next lemma tells us how the general bracket behaves under a Reidemeister move of type two. Essential diagrams for this lemma are in Fig. 28.

Lemma. Let K be a given link diagram, and let K' denote a diagram that is obtained from K by performing a type 2 Reidemeister move in the simplifying direction (eliminating two crossings from K). Let K'' be the diagram obtained from K by replacing the site of the type 2 move by two arcs in the opposite

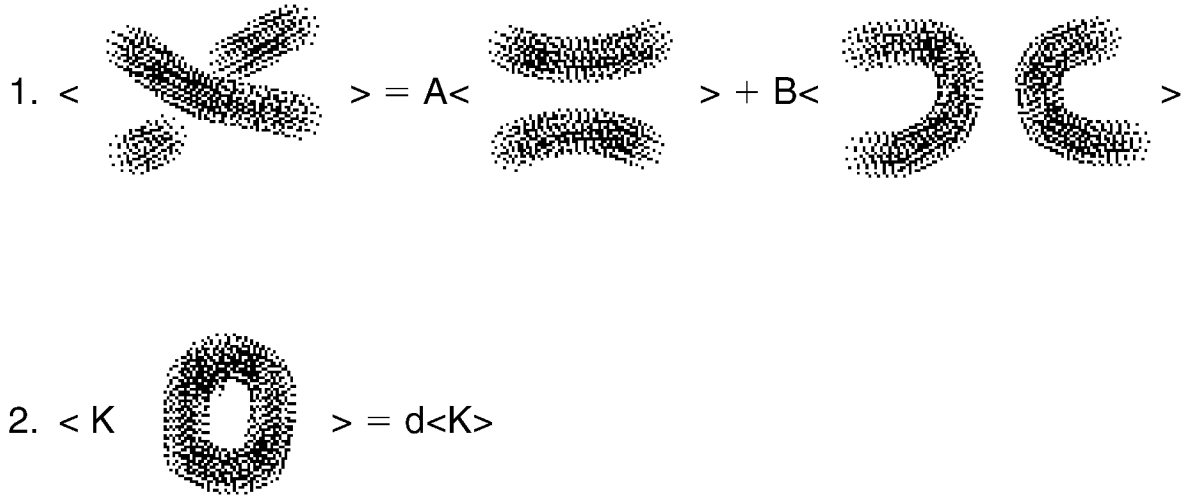


FIGURE 27 Bracket equations.

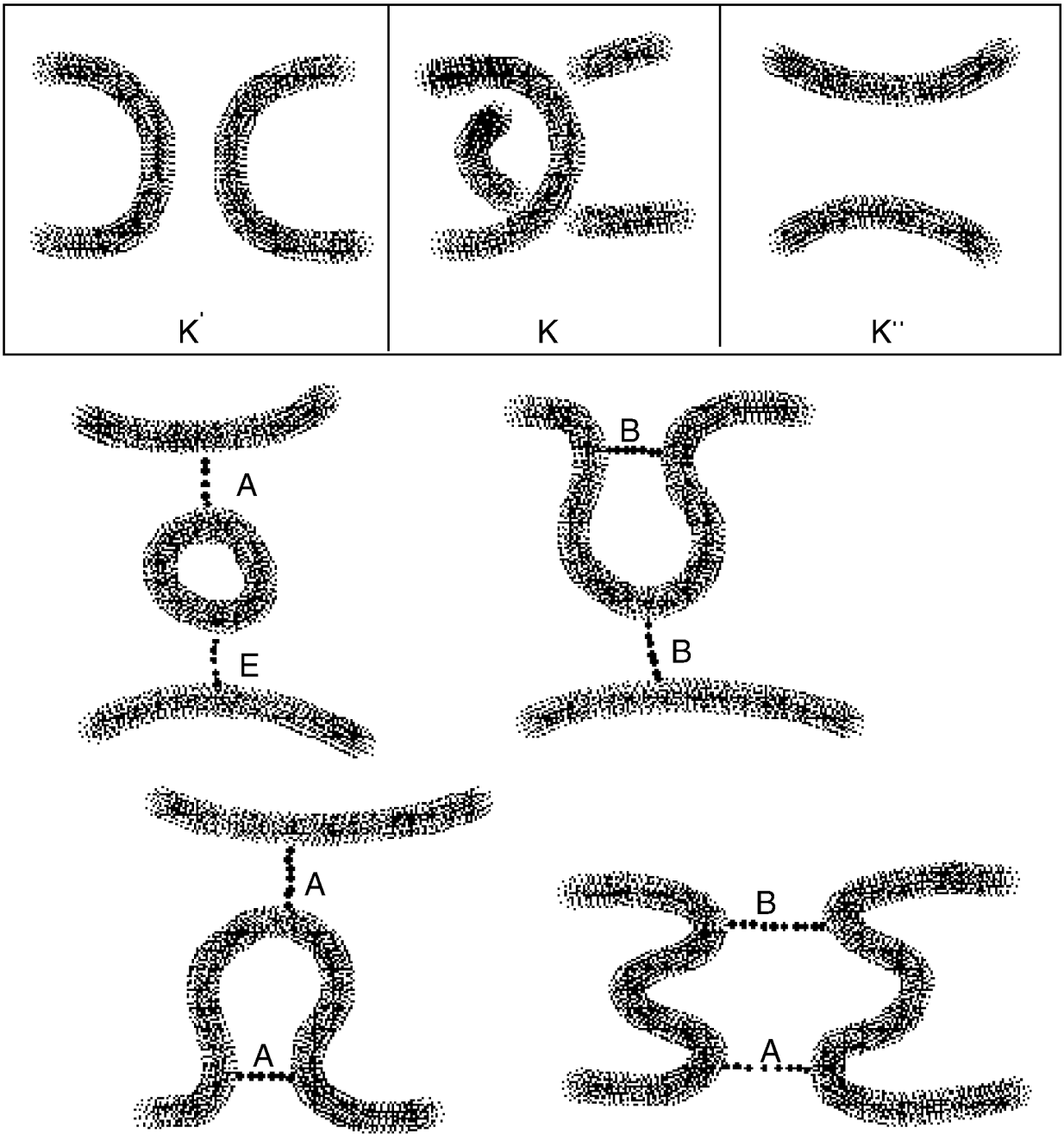


FIGURE 28 Type 2 changes.

pattern to the form of the simplified site in K'. (The diagrams in Fig. 28 illustrate this construction.) Then $\langle K\rangle = AB\langle K'\rangle + (ABd + A^2 + B^2)\langle K''\rangle$.

Proof. Consider the four local state configurations that are obtained from the diagram K on the left-hand side of the equation, as illustrated in Fig. 28. The formula follows from the fact that one of these states has coefficient AB and the other three have the same underlying diagram and respective coefficients ABd (after converting the loop to a value d), A^2, and B^2. This completes the proof of the Lemma.

With the help of this lemma it is now obvious that if we choose $B = 1/A$ and $d + A^2 + B^2 = 0$, then $\langle K\rangle$ is invariant under the second Reidemeister move.

Once this choice is made, the resulting specialized bracket is invariant under the third Reidemeister move, as illustrated in Fig. 29.

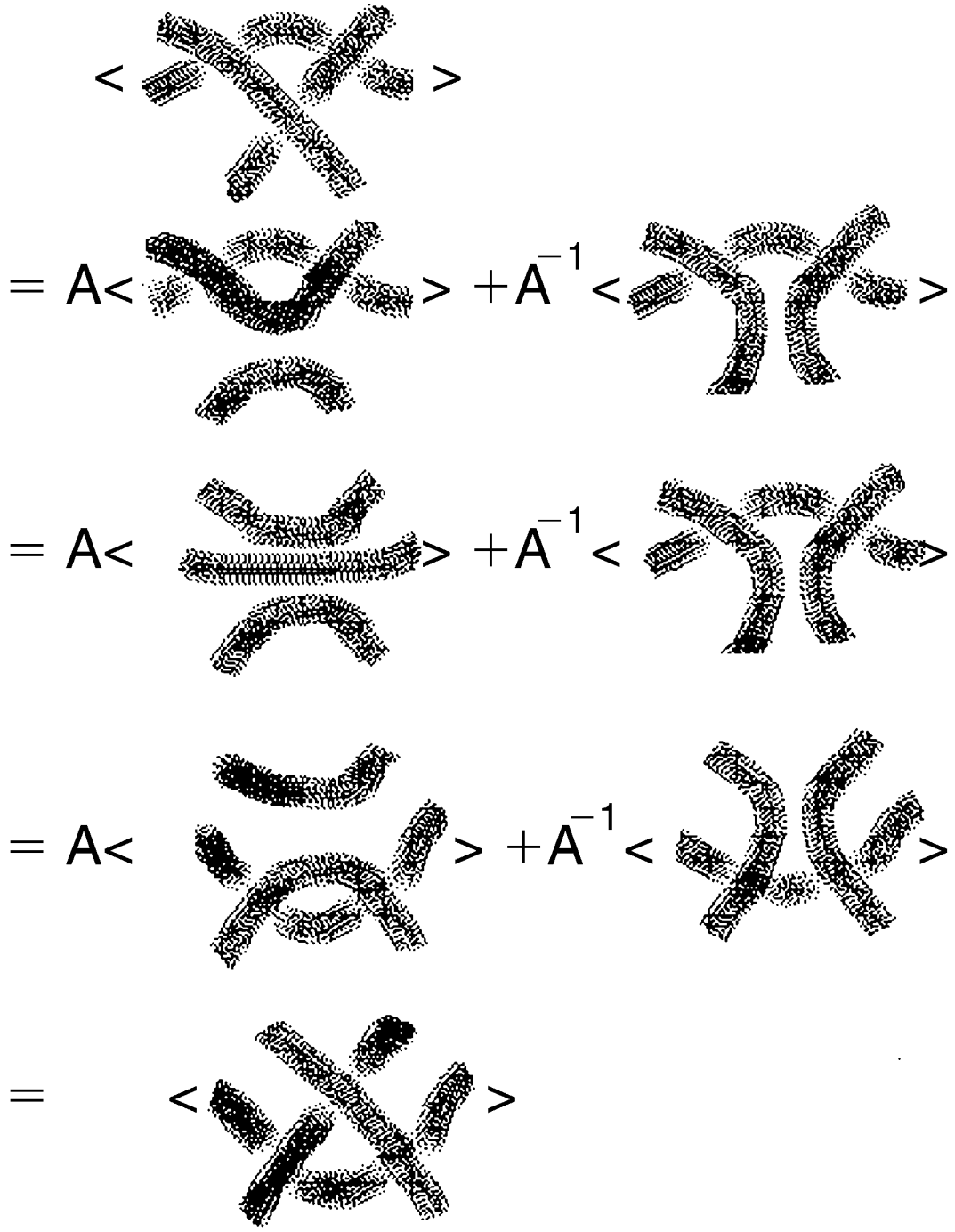


FIGURE 29 Type 3 invariance of the bracket.

Finally, we can investigate bracket behavior under the first Reidemeister move.

Lemma. Let $\langle K \rangle$ denote the bracket state sum with $B = A^{-1}$ and $d = -A^2 - A^{-2}$. Then $\langle K \rangle$ is invariant under the Reidemeister moves 2 and 3 and on move 1 behaves as shown:

$$\langle K(+) \rangle = (-A^3)\langle K \rangle,$$
$$\langle K(-) \rangle = (-A^{-3})\langle K \rangle.$$

Here $K(+)$ denotes a diagram with a simplifying move of type 1 available where the crossing that is to be removed has type $+1$. K is the diagram obtained from $K(+)$ by doing the type 1 move. Similarly, $K(-)$ denotes a diagram

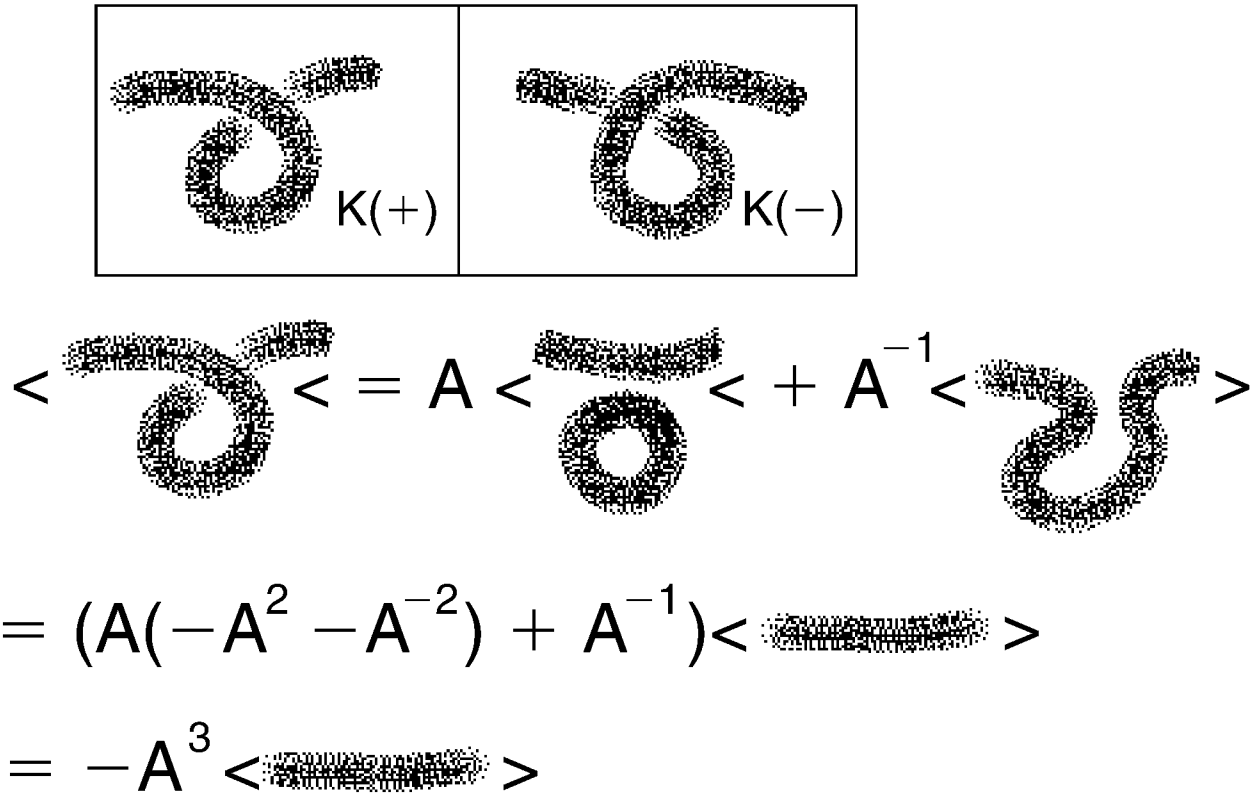


FIGURE 30 Bracket under type 1 move.

with a simplifying move of type 1 available where the crossing that is to be removed has type -1. Figure 30 illustrates the diagrams for $K(+)$ and $K(-)$.

Proof. See Fig. 30 for the behavior under type 1 moves. We have already verified the other statements in this lemma.

B. Framing Philosophy, Twist and Writhe

Is it unfortunate that the bracket is not invariant under the first Reidemeister move? No, it is fortunate! First of all, the matter is easy to fix by a little adjustment: Let K be an oriented knot or link, and define the *writhe* of K, denoted $w(K)$, to be the sum of the signs of all the crossings in K. Thus, the writhe of the right-handed trefoil knot is three. The writhe has the following behavior under Reidemeister moves:

(i) $w(K)$ is invariant under the second and third Reidemeister moves.

(ii) $w(K)$ changes by plus or minus one under the first Reidemeister move:

$$w(K(+)) = w(K) + 1$$
$$w(K(-)) = w(K) - 1.$$

(Here we use the notation of the previous lemma as shown in Fig. 30.)

Thus the writhe behaves in a parallel way to the bracket on the type 1 moves, and we can combine writhe and bracket to make a new calculation that is invariant under all three Reidemeister moves. We call the fully invariant calculation the "f-polynomial" and define it by the equation

$$f_K(A) = (-A^3)^{-w(K)}\langle K\rangle(A).$$

Up to this normalization, the bracket gives a model for the original Jones polynomial. The precise relationship is that $V_K(t) = f_K(t^{-1/4})$, where $w(K)$ is the sum of the crossing signs of the oriented link K, and $\langle K\rangle$ is the bracket polynomial obtained by ignoring the orientation of K.

We shall return to this relationship with the Jones polynomial in a moment, but first a little extra mathematical philosophy: Another way to view the fact of the bracket's lack of invariance under the first Reidemeister move is to see that the bracket is an invariant of knotted and linked bands embedded in three-dimensional space. Regard a link diagram as shorthand for an embedding of bands as shown in Fig. 31.

Figure 31 illustrates a link diagram for the trefoil knot in a thick, dark mode of drawing. This diagram is juxtaposed

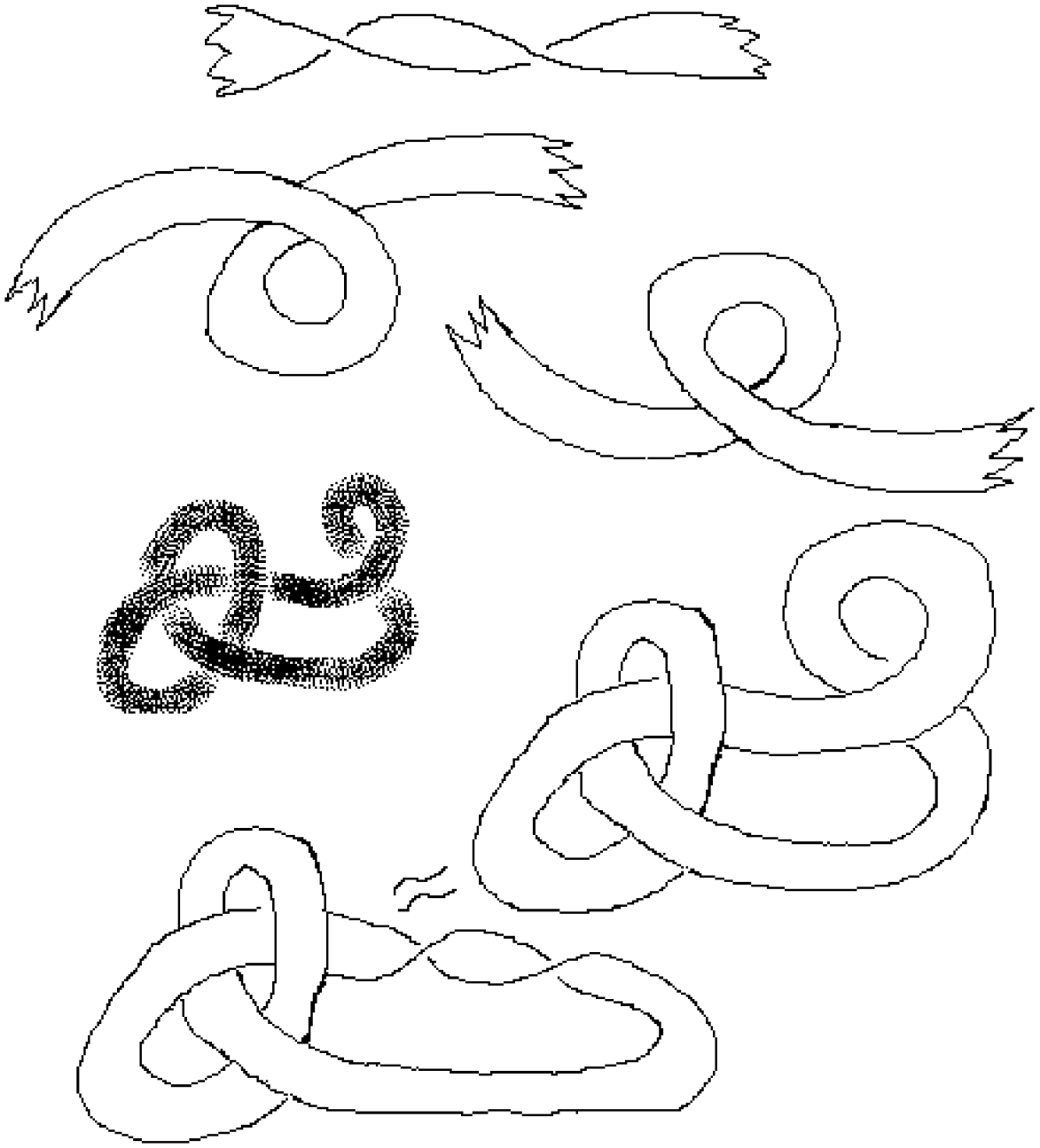


FIGURE 31 Bands and twists.

with a drawing of a knotted band that parallels that knot diagram. The band has two boundary components that proceed (mostly in the plane) parallel to one another. The curl in the knot diagram becomes a flat curl in the band that is ambient isotopic to a full twist (two half-twists) in the band. This isotopy is indicated in Fig. 31. The top of Fig. 31 shows a full twist in a band and two flat curls that both give rise to this same full twist by ambient isotopy that leaves their ends fixed. Each component of a link diagram is replaced by a paralleled version: the analog of a ribbbon-like strip of paper attached to itself with an even number of half-twists. The first Reidemeister move no longer applies to this shorthand since we can, at best, replace a curl by a twist as shown in Fig. 31.

In fact, as Fig. 31 shows, there are two distinct curls corresponding to a single full twist of a band. The bracket (and the writhe) behave the same way on both of these twists. This means that we can reinterpret the bracket as an invariant of the topological embeddings of knotted, linked, and twisted bands in three-dimensional space. This means

that the bracket has a fully three-dimensional interpretation, although its definition depends upon the use of planar projections.

C. Calculating the Bracket

Figure 32 shows a tree for calculating the bracket polynomial of the trefoil knot T.

It follows at once from the behavior of the bracket on curls that the contributions of the three (farthest from the trefoil itself) branches of this tree add to give the bracket polynomial of the trefoil:

$$\begin{aligned}\langle T\rangle &= A^2(-A^3) + AA^{-1}(-A^{-3}) + A^{-1}(-A^{-3})^2\\ &= -A^5 - A^{-3} + A^{-7}.\end{aligned}$$

Hence,

$$f_T = (-A^3)^{-w(T)}\langle T\rangle = A^{-4} + A^{-12} - A^{-16}.$$

Note that we managed only three branches in the tree for this calculation rather than the full expansion into the eight states. A savings like this is always possible because we know how the bracket behaves on curls. The resulting expansion gives a sum of monomials and is useful for thinking about the properties of the invariant.

D. Mirror Mirror

The knot K^* obtained by reversing all the crossings of K is called the mirror image of K. The knot K^* is the mirror

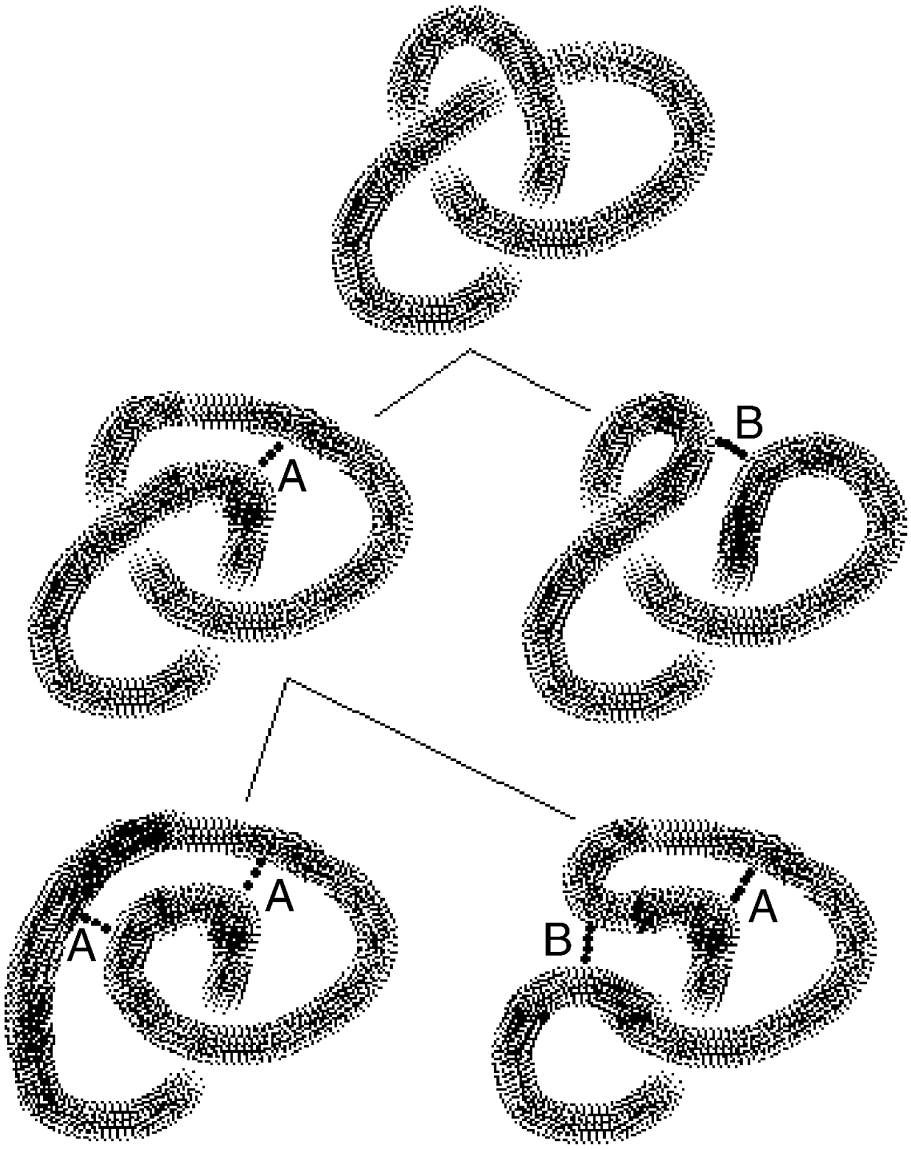


FIGURE 32 Tree for bracket of trefoil.

image of the knot that would ensue if the plane on which the knot is drawn were a mirror. It is easy to see that $\langle K^* \rangle(A) = \langle K \rangle(A^{-1})$ and that $f_{K^*}(A) = f_K(A^{-1})$. Thus, if K is ambient isotopic to K^* (all three Reidemeister moves allowed), then

$$f_K(A) = f_{K^*}(A) = f_K(A^{-1}).$$

Returning to the evaluation of the f-invariant for the trefoil, note that $f_T(A^{-1})$ is not equal to $f_T(A)$. Therefore, the trefoil knot T and its mirror image T^* are topologically distinct. The proof that we have given for it is the simplest proof known to this author. Note that we have given a complete proof of this fact, starting with the Reidemeister moves, constructing and applying the bracket invariant.

A knot is said to be *chiral* if it is not ambient isotopic to its mirror image. The words chiral and *chirality* come from physical chemistry and natural science. A knot that is equivalent to its mirror image is said to be *achiral*. (or *amphicheiral* in the speech of knot theorists). Many knots are achiral. The reader may enjoy verifying that the figure eight knot shown in Fig. 33 is ambient isotopic to its mirror image.

A complete understanding of the problem of determining whether a knot is chiral remains in the far distance. The new invariants of knots and links have enhanced our understanding of this difficult question.

E. Return to the Jones Polynomial

Now let us verify that the bracket does indeed give a model for the Jones polynomial. To see this, consider $f_K(A) = (-A^3)^{-w(K)} \langle K \rangle(A)$. Since the writhe $w(K)$ is obtained by summing signs over all the crossings of K, we can interpret the factor $(-A^3)^{-w(K)}$ as the product of contributions of $(-A^3)$ or $(-A^3)^{-1}$, one from each crossing and depending upon the sign of the crossing. Thus we can write an oriented state expansion formula for f_K as shown below, where K_+ and K_- denote links with corresponding sites with oriented crossings, K_0 is the result of smoothing the crossing in an oriented fashion, and $K_\&$ is the result of smoothing the crossing against the orientation:

$$f_{K+} = (-A^3)^{-1} A f_{K_0} + (-A^3)^{-1} A^{-1} f_{K\&}.$$

Hence,

$$f_{K_+} = -A^{-2} f_{K_0} - A^{-4} f_{K_\&}$$

and similarly, for a negative crossing,

$$f_{K_-} = -A^2 f_{K_0} - A^4 f_{K_\&}.$$

Letting $V_K(t) = f_K(t^{-1/4})$, we have

$$V_{K_+} = -t^{1/2} V_{K_0} - t V_{K_\&}$$
$$V_{K_-} = -t^{-1/2} V_{K_0} - t^{-1} V_{K_\&}.$$

Therefore,

$$t^{-1} V_{K_+} - t V_{K_-} = (t^{1/2} - t^{-1/2}) V_{K_0}.$$

We leave the rest of the verification that $V_K(t)$ is the Jones polynomial (see Section 3) to the reader (check that it has the right behavior on unknotted loops).

V. VASSILIEV INVARIANTS

We have seen how it is fundamental to take the difference of an invariant at a positive crossing and at a negative crossing, leaving the rest of the diagram alone. The earliest

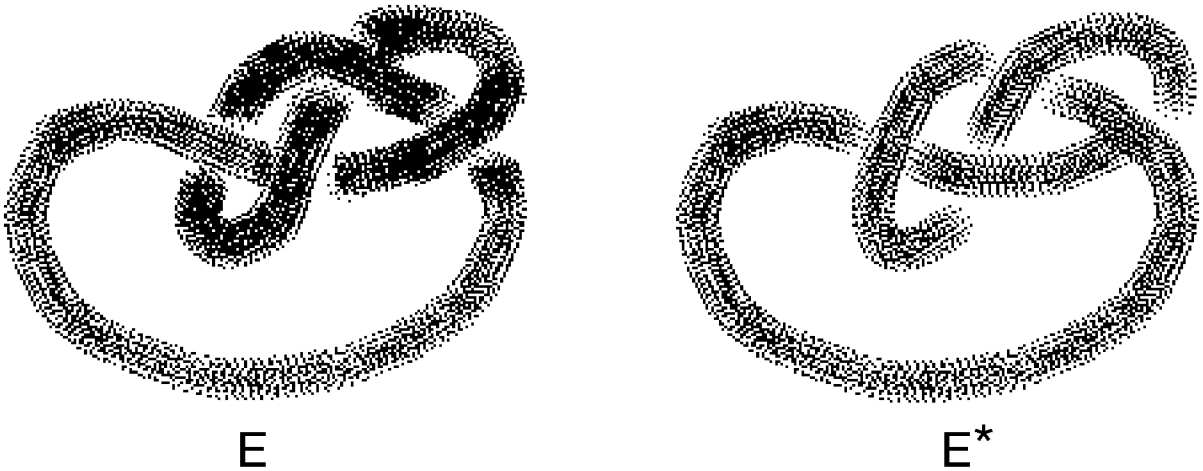


FIGURE 33 Figure eight knot and its mirror image.

instance of this is the Conway polynomial $C_K(z)$ with its exchange identity

$$C_{K_+} - C_{K_-} = zC_{K_0}.$$

Vassiliev (1990) gave new meaning to this sort of identity by thinking of the structure of the entire space of all mappings of a circle into three-dimensional space. This space of mappings includes mappings with singularities where two points on a curve touch. He interpreted the equation

$$Z_{K_+} - Z_{K_-} = Z_{K_\#}$$

as describing the difference of values across a singular embedding $K_\#$, where $K_\#$ has a transverse singularity in the knot space as illustrated in Fig. 34. (In a transverse singularity the curve touches itself along two different directions.)

The Vassiliev formula serves to define the value of the invariant on a singular embedding in terms of the the values on two knots "on either side" of this embedding. This Vassiliev formula serves to describe a method of extending a given invariant of knots to a corresponding invariant of embedded graphs with controlled singularities of this transverse type. This idea had been considered before Vassiliev. Vassiliev carried out his program of analyzing the singular knot space using techniques of algebraic topology, and in the course of this investigation he discovered a key concept that had been completely overlooked in the context of graph invariants. That concept is the idea of an invariant of *finite type*.

Definition. We shall say that Z_G is an invariant of *finite type i* if Z_G vanishes for all graphs with greater than i nodes.

This concept was extracted from Vassiliev's work by Birman and Lin (1993). A (rigid vertex) invariant of knotted graphs is a *Vassiliev invariant of finite type i* if it satisfies the identity

$$Z_{K_+} - Z_{K_-} = Z_{K_\#}$$

and it is of finite type i. In *rigid vertex isotopy* the cyclic order at the vertex is preserved, so that the vertex behaves like a rigid disk with flexible strings attached to it at specific points.

Vassiliev invariants form an extraordinary class of knot invariants. It is an open problem whether the Vassiliev invariants are sufficient to distinguish knots that are topologically distinct.

Vassiliev began an analysis of the combinatorial conditions on graph evaluations that could support such invariants. The key observation is the following result:

Lemma. If Z_G is a Vassiliev invariant of finite type i, then Z_G is independent of the embedding of the graph G when G has i vertices.

Proof. Suppose that G is an embedded graph G with i nodes. If we switch a crossing in G to form G', then the exchange relation for the Vassiliev invariant says that $Z_G - Z_{G'} = Z_{G''}$, where G'' has one more node than G or G'. But then G'' has $i+1$ nodes and hence $Z_{G''} = 0$. Therefore $Z_G = Z_{G'}$. This shows that we can switch crossings in any embedding of G without changing the value of Z_G. It follows from this that Z_G is independent of the embedding and depends only on the graph G. This completes the proof of the lemma.

For a Vassiliev invariant of type i, there is important information in the values it takes on graphs with exactly i nodes. These evaluations do not depend upon the embedding type of the graph. However, not just any such graphical evaluation will extend to give a topological invariant of knots and graphs. There are necessary conditions.

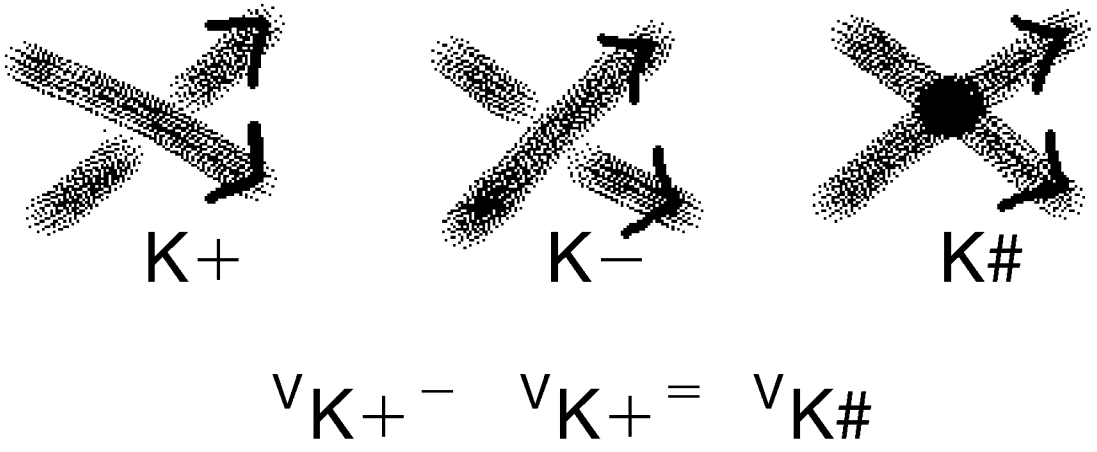


FIGURE 34 Difference equation.

Vassiliev found a version of these conditions through his analysis of the knot space and Stanford (1996) discovered the beautiful topological meaning of these conditions in relation to the switching identity. Stanford's argument goes as follows: Consider a singular crossing that has an arc from the diagram passing underneath it as shown in Fig. 35. Four crossing switches will take that arc above the singular crossing and return the diagram to a position that is topologically equivalent to its original position.

Each crossing switch gives an equation. There are four equations. Add them up and one gets an identity among the values of the invariant on four diagrams. Call this *the four-term relation*. This identity is illustrated in the second box in Fig. 35.

Now recall from the lemma we proved above that for a Vassiliev invariant of type i, the graphs with i nodes have values that are independent of their embeddings in three-dimensional space. This means that at the *top level* (the i-noded graphs for a Vassiliev invariant of type i will be called the top level) the four-term relations will be relations among the evaluations of abstract graphs. At the top level the four-term relations will be purely combinatorial conditions related to the topology.

How shall we think of abstract four-valent graphs corresponding to singular embeddings of a knot? An abstract knot is just a circle. An abstract singular knot is a circle with pairs of points marked that become the singular points in the embedding. Indicate these paired points by arcs between them. Call the resulting structure a *chord diagram*. See the example at the beginning of Fig. 36.

In the language of the chord diagrams the four-term relation at the top level (see the discussion of the top level in the paragraph above) becomes the equation shown in Fig. 36. This can be seen by translating the relation in Fig. 35 into the language of chord diagrams. In Fig. 36 we indicated parts of the chord diagram that are neighbors by showing an outer bracket connecting them. Those sites that are neighbors can have no other chords between them. Otherwise there can be many chords in these diagrams that are not indicated, just so long as the diagrams in the equation for the four-term relation differ only as shown in the figure.

If one can write down a top-level evaluation of chord diagrams that satisfies the four-term relation, then one has the raw data for a Vassiliev invariant. Such an evaluation of chord diagrams is called a *weight system* for a Vassiliev invariant. By the theorems of Kontsevich (1994) and Bar-Natan (1995), these raw data guarantee the existence of at least one invariant that satisfies the top-level evaluation.

The world is rife with Vassiliev invariants. Birman and Lin (1993) showed directly that the Jones polynomial and its generalizations give rise to Vassiliev invariants. In the case of the Jones polynomial here is an easy proof of their result:

Theorem. Let $V_G(t)$ denote the Jones polynomial extended to rigid vertex 4-valent graphs by the formula $V_{K_+} - V_{K_-} = V_{K_\#}$. Let $v_i(G)$ denote the coefficient of x^i in the expansion of $V_G(\exp(x))$. Then $v_i(G)$ is a Vassiliev invariant of type i.

Proof. Use the identities from the end of Section 4:

$$V_{K_+} = -t^{1/2}V_{K_0} - tV_{K_\&},$$
$$V_{K_-} = -t^{-1/2}V_{K_0} - t^{-1}V_{K_\&}.$$

Substitute $t = \exp(x)$. It follows at once that $V_{K_\#} = V_{K_+} - V_{K_-}$ is divisible by x. Hence V_G is divisible by x^i when G has i nodes. This implies that the coefficients $v_i(G) = 0$ if G has more than i nodes. Hence the coefficients $v_i(G)$ are of finite type, proving the theorem.

With the help of theorems of this type it is possible to study Vassiliev invariants by studying the structure of known invariants of knots and links. In particular it is possible to justify the structure of many weight systems in terms of known invariants. We shall not go into these sorts of investigations in this exposition. The next section shows how the algebraic study of Lie algebras is directly related to the construction of Vassiliev invariants. This is one beginning of a whole world of relationships between knot theory and algebra.

VI. VASSILIEV INVARIANTS AND LIE ALGEBRAS

The subject of Lie algebras is an algebraic study with a remarkable connection with the topology of knots and links. The purpose of this section is to first give a brief introduction to the concept of a Lie algebra and then to show the deep connection between these algebras and the structure of Vassiliev invariants for knots and links, as described in the previous section.

In order to understand the idea behind a Lie algebra it is helpful to first consider the concept of a group. A set G is said to be a *group* if it has a single binary operation * such that:

1. Given a and b in G, then a^*b is also in G.
2. If a, b, c are in G, then $(a^*b)^*c = a^*(b^*c)$.
3. There is an element E in G such that $E^*a = a^*E = a$ for all a in G.
4. Given a in G, there exists an element a^{-1} in G such that $a^*a^{-1} = a^{-1*}a = E$.

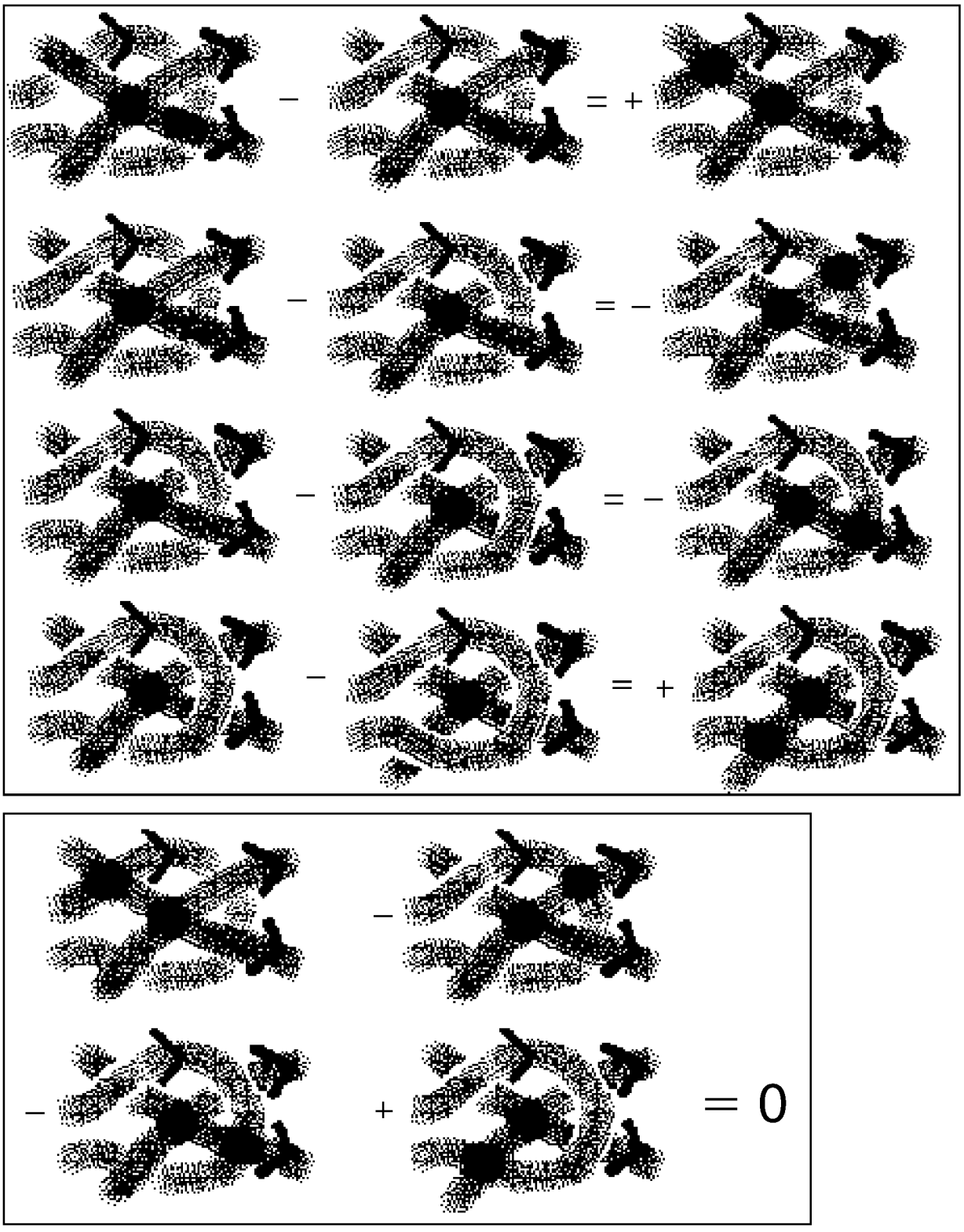


FIGURE 35 Embedded four-term relation.

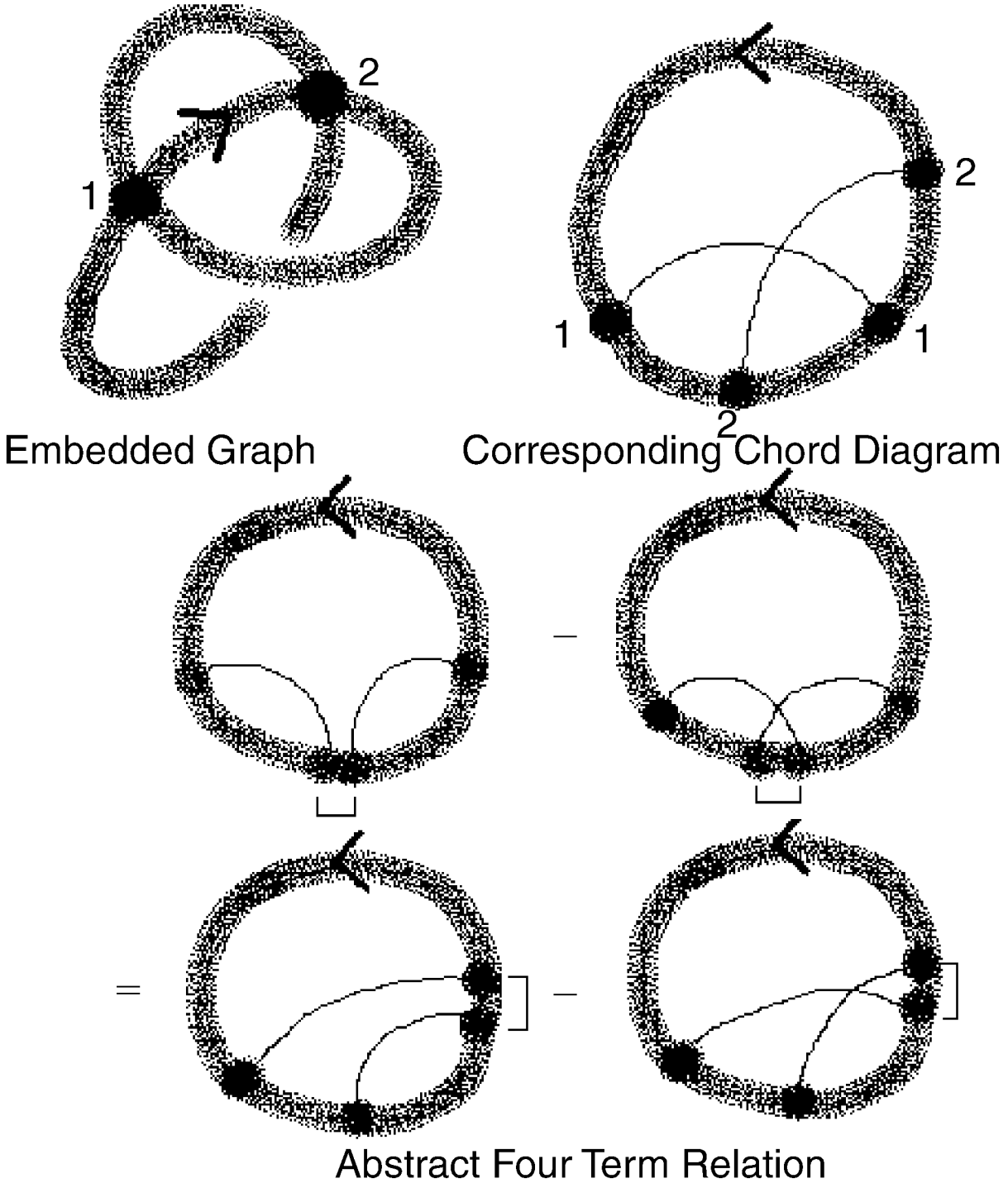


FIGURE 36 Abstract four-term relation via chord diagrams.

One of the most fertile sources of groups is matrix algebra. Recall that an $n \times n$ matrix A is an array of numbers A_{ij} (real or complex) $A = (A_{ij})$, where i and j range in value from 1 to n. One defines the product of two matrices by the formula $(AB)_{ij} = \sum_k A_{ik} B_{kj}$, where k runs from 1 to n in this summation.

For our purposes it is essential to have a diagrammatic representation for matrix multiplication. This

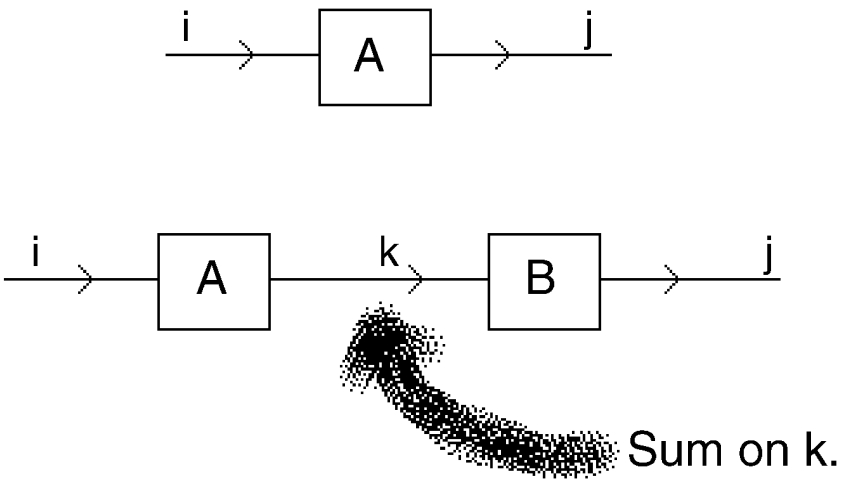


FIGURE 37

representation is illustrated in Fig. 37. Each matrix is represented by a labeled box with one arrow that enters the box and one arrow that leaves the box. The entering arrow corresponds to the left index i in A_{ij} and the right arrow corresponds to the right index j. In multiplying two matrices A and B together to form AB we tie the outgoing arrow of A to the ingoing arrow of B. By convention, an arrow that has no free ends connotes the summation over all possible choices of index for that arrow.

Many facts about matrices become quite transparent in this notation. For example, the *trace* of A, denoted tr(A), is the sum of the diagonal entries A_{ii} where i ranges from one to n. The diagrammatic proof of the basic formula $\mathrm{tr}(AB) = tr(BA)$ is illustrated in Fig. 38.

For a given value of n, we let $M_n(R)$ denote the set of all $n \times n$ matrices with coefficients in the real numbers R. We let $A^*B = AB$ denote the product of matrices and we let E denote the matrix whose entries are given by $E_{ii} = 1$ for all i and $E_{ij} = 0$ if i is not equal to j. With this choice of multiplication and identity element E, $M_n(R)$ satisfies the first three axioms for a group. However, there are matrices A that have no inverse (A^{-1} so that $AA^{-1} = E$). For

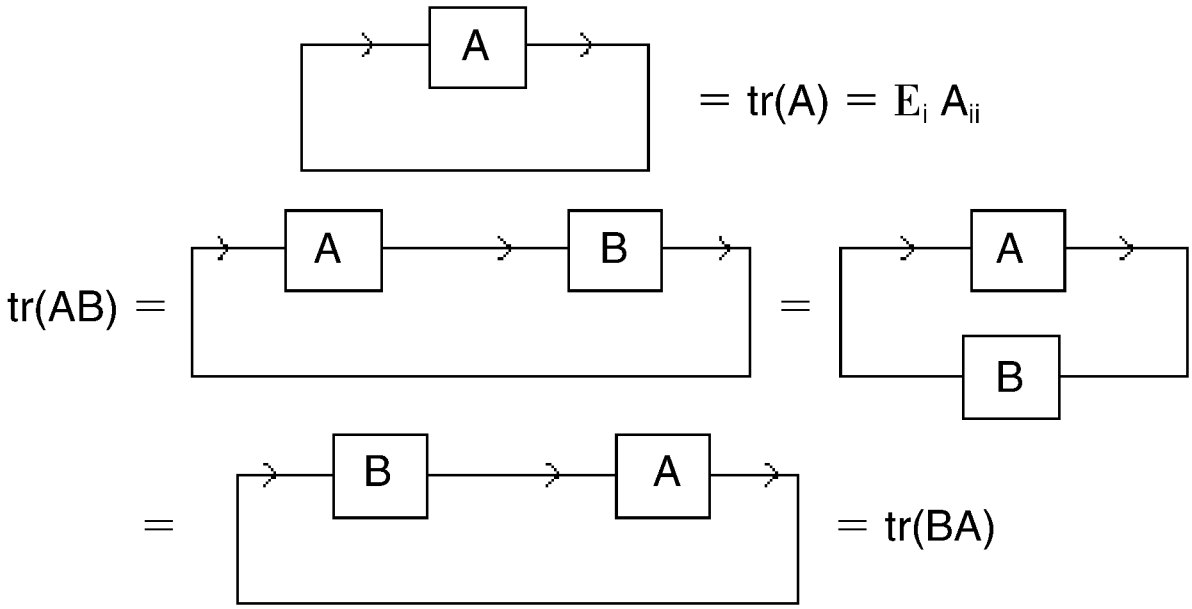


FIGURE 38 Diagrams for Matrix Trace.

example the matrix 0, all of whose entries are zero, is a matrix without an inverse. Thus $M_n(R)$ is not itself a group.

There is a criterion for a matrix to have an inverse. This is simply that the determinant $\mathrm{Det}(A)$ should be nonzero. Thus the largest group of matrices of size $n \times n$ that we can devise is the set of all matrices A such that $\mathrm{Det}(A)$ is nonzero. This is called the *general linear group* and is denoted by $GL_n(R)$. There are many interesting subgroups of this large group of matrices. One example is the group $Sl_{(n)}$ of all matrices with determinant equal to one. We may also restrict to *orthogonal* matrices A over R. These are invertible matrices A such that $A^t = A^{-1}$, where A^t denotes the transpose of the matrix A: $A^t_{ij} = A_{ji}$. The group of orthogonal matrices is denoted by $O(n)$.

The intersection of $O(n)$ and $Sl_{(n)}$ is denoted $SO(n)$. $SO(n)$ consists of the orthogonal matrices of determinant equal to one. In the case $n = 2$, $SO(2)$ consists of rotations of the plane that fix the origin, and in the case of $n = 3$, $SO(3)$ consists of rotations of three-dimensional space about specified axes.

$SO(3)$ has a fascinating collection of finite subgroups including the symmetries of the classical regular solids: the tetrahedron, the cube, the octahedron, the dodecahedron, and the icosahedron. Ultimately, the matrix groups become a language for the precise expression of symmetry.

We now ask when a matrix A can be written in the form

$$A = e^B = E + (1/1!)B + (1/2!)B^2 + (1/3!)B^3 + \cdots$$

for some other matrix B. Since $e^B = \mathrm{limit}(E + B/m)^m$, where the limit is taken as m approaches infinity, we can regard $(E + B/m)$, for m large, as an "infinitesimal" version of the matrix A, and one refers to B as an "infinitesimal generator" for A. It is interesting and mathematically significant to compare the algebraic properties of A and B. The key property for this comparison is the determinant equation

$$\mathrm{Det}(e^B) = e^{\,\mathrm{tr}(B)},$$

where $\mathrm{tr}(B)$ denotes the trace of B. (One way to prove this identity is to use the Jordan canonical form for the matrix and the fact that similar matrices have the same trace and determinant.)

For example, if $\mathrm{Det}(e^B) = 1$, then we need that $\mathrm{tr}(B) = 0$. This means that elements of $Sl_{(n)}$ are the exponentials of matrices with trace equal to zero.

Let $sl_{(n)}$ denote the set of $n \times n$ matrices with trace equal to zero. The set $sl_{(n)}$ is not closed under matrix multiplication, but it is closed under the *Lie bracket* (or *commutator*) operation $[B, C] = BC - CB$.

$$\begin{aligned} &\text{If } \mathrm{tr}(B) = \mathrm{tr}(C) = 0, \text{ then} \\ &\mathrm{tr}[B, C] = \mathrm{tr}(BC - CB) = \mathrm{tr}(BC) - \mathrm{tr}(CB) \\ &\qquad = \mathrm{tr}(BC) - \mathrm{tr}(BC) = 0 \end{aligned}$$

(since $\mathrm{tr}(BC) = \mathrm{tr}(CB)$ for any matrices B and C). Thus, if B and C belong to $sl_{(n)}$, then $[B, C]$ also belongs to $sl_{(n)}$. This closure under the bracket operation leads directly to the notion of a Lie algebra.

Definition. A Lie algebra is a vector space L over a field F that is closed under a binary operation, called the *Lie bracket* and denoted by $[B, C]$ for B and C in L. The bracket is assumed to satisfy the following axioms.

1. $[X, Y] = -[Y, X]$ for all X and Y in L.
2. $[aX + bY, Z] = a[X, Z] + b[Y, Z]$ for all a and b in F and X, Y, Z in L.
3. $[X, [Y, Z]] + [Z, [X, Y]] + [Y, [Z, X]] = 0$.

This last identity is called the *Jacobi identity*. It is easy to verify that the bracket operation $[B, C] = BC - CB$ on the vector space of all $n \times n$ matrices over F (e.g., $F = R$) satisfies the axioms given above. Thus, we have so far seen that $sl_{(n)}$ is a Lie algebra that is naturally associated with the group of matrices $SL_{(n)}$. In fact, $sl(n)$ generates $SL_{(n)}$ by exponentiation.

There is a general pattern. Each matrix group has its corresponding Lie algebra. The classification of matrix groups is simplified by a corresponding classification of Lie algebras. As a result, the Lie algebras are a subject in their own right. It has often happened that Lie algebras are connected mathematically with subjects different from their original roots in group theory.

In our context the Lie algebras turn out to be related to the formation of weight systems for Vassiliev invariants. One way to see this is to just take the case of matrix Lie algebras with commutator brackets and interpret diagrammatically the formula that states that the Lie algebra is closed under the bracket operation. This formula states that there is a basis $\{T^1, T^2, \ldots, T^m\}$ for the Lie algebra as a vector space over F such that each T^a is an $n \times n$ matrix and such that

$$T^a T^b - T^b T^a = f_{abc} T^c,$$

where f_{abc} is a set of constants in F depending on the three indices a, b, c (running from 1 to n). The right-hand side of this equation connotes a summation over all values of the index $c = 1, \ldots, n$. The left-hand side is the commutator of T^a and T^b for any given choice of a and b.

In Fig. 39 we diagram this equation using the conventions for diagrammatic matrix multiplication explained in this section. The structure constants f_{abc} are represented by a graphical vertex with three lines attached to it, one for a, one for b, and one for c. For the purpose of discussion,

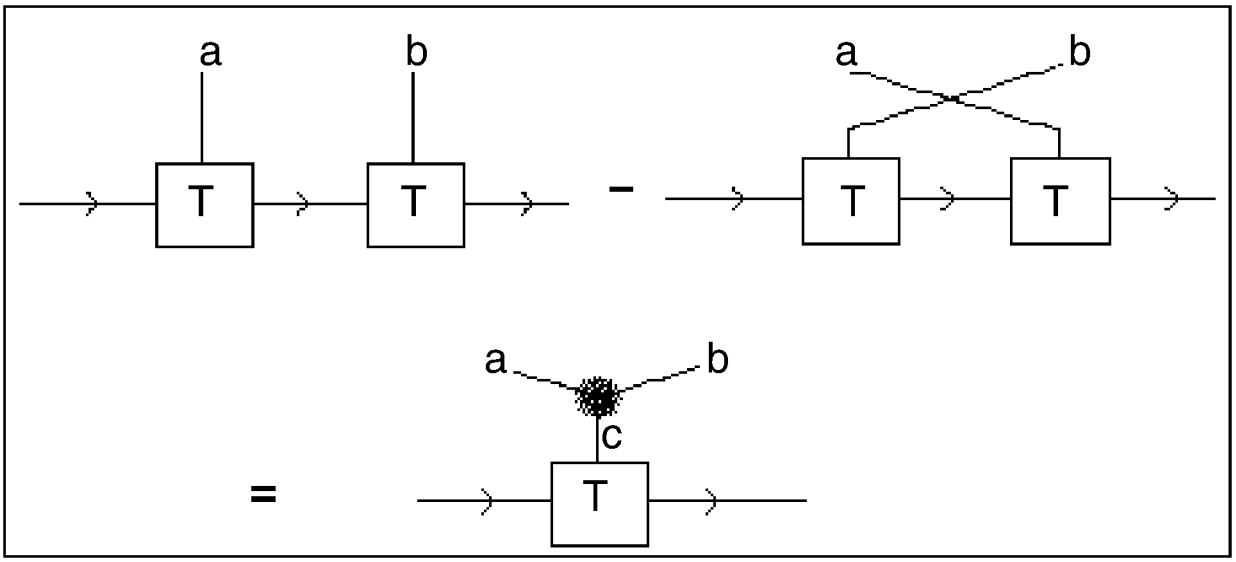


FIGURE 39 Diagrammatic Lic Algebra.

we shall assume that f_c^{ab} is dependent only on the cyclic order of *abc*. It is convenient to regard the graphical vertex as representing a "tensor" that has this cyclic invariance since this means that we can slide the diagram for the structure constant tensor around in the plane so long as we keep the cyclic order of its legs unchanged. Such bases can be obtained in many cases of matrix Lie algebras, and the results that we outline can be generalized in any case.

Figure 40 shows a formal version of the commutator relation of Fig. 39, except that the labels and indices have been removed and the boxes for matrix elements have been replaced by graphical vertices. Imagine that the terms in this formal version of the commutator relation are parts of chord diagrams as illustrated with examples in this figure. In other words, recall the method of chord diagrams from the last section and imagine that along with the chords there are also trivalent graphical vertices among the chords, and that these vertices are related to commutators as shown in the figure. Finally, Fig. 41 shows a formal derivation of the four-term relation for chord diagrams from the diagrammatic commutator identity. This means that the four-term relation that we derived from topological considerations in the last section is intimately related to the basic structure of a Lie algebra. This is the essence of the relationship of Vassiliev invariants with Lie algebras and their generalizations.

Concretely, the relationship we have just described means that it is possible to construct weight systems for Vassiliev invariants by using matrix Lie algebras. To see how this works see Fig. 42. Here we indicated a chord diagram D and a corresponding diagram involving matrices T^a from a Lie algebra basis. The second diagram represents the sum of traces $\mathrm{wt}(D) = \sum \mathrm{tr}(T^a T^b T^c T^a T^b T^c)$, where we are summing over all values for the indices a, b, and c. This second diagram represents the weight $\mathrm{wt}(D)$ that is assigned to the first diagram. It follows from our considerations that this weight system satisfies the four-term relation and hence, by the theorem of Kontsevich, is the top-row evaluation for a Vassiliev invariant.

This section has sketched the amazing and deep connection between Lie algebras and invariants of knots and links. The territory is even more surprising as one explores it further. First of all, it should be clear from what we have said that what is really needed here is an appropriate generalization of Lie algebras. In fact, prior to the discovery of the Vassiliev invariants, a very remarkable such generalization called "quantum groups" was discovered through work in statistical mechanics and was applied to knot theory. It was already known that quantum groups provided a strong connection between Lie algebras and their generalizations and invariants of knots and links. Now the matter of finding all weight systems challenges the resources of

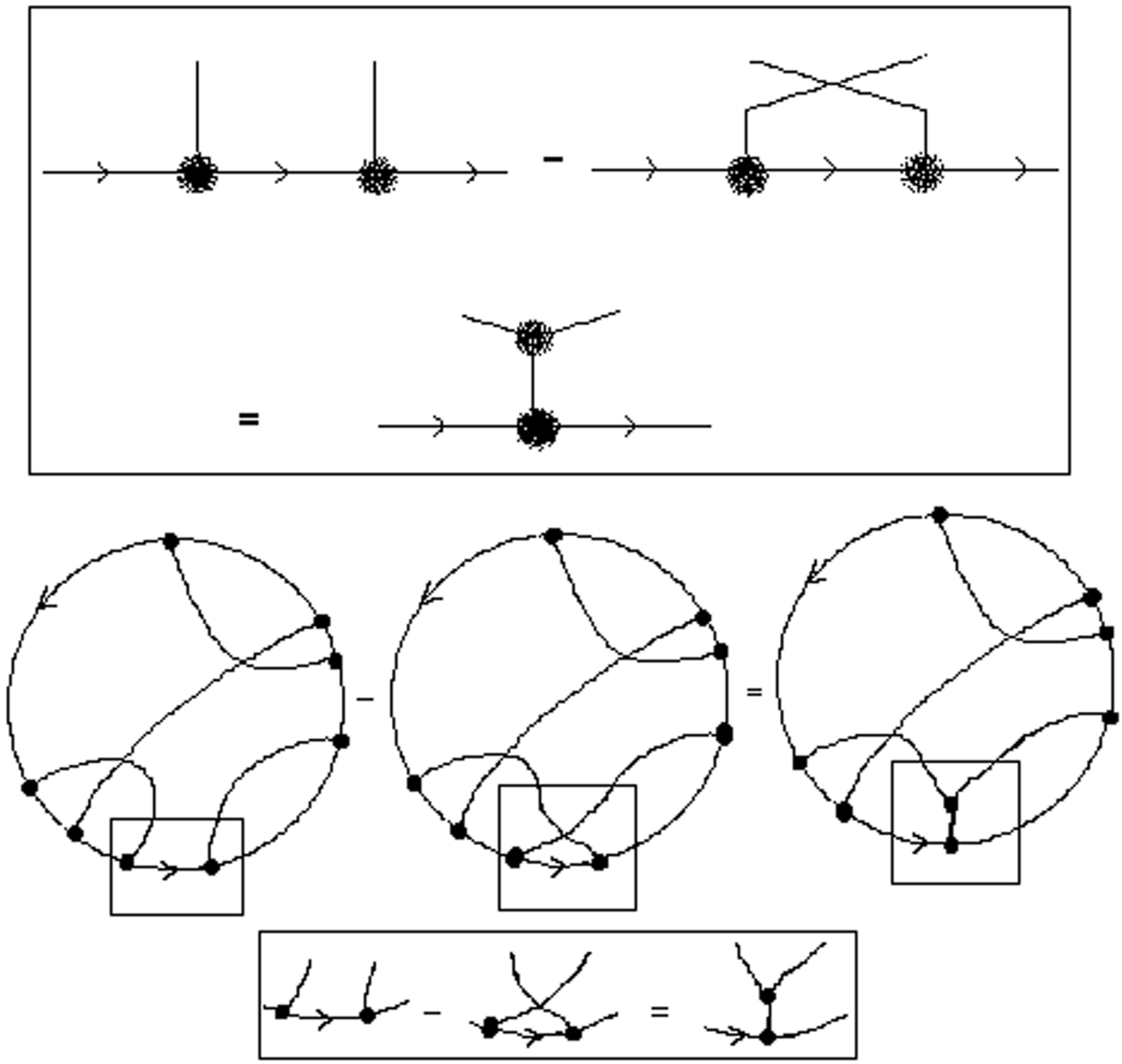


FIGURE 40 Lic Algebra and Chord Diagrams.

quantum groups and it is not known if all Vassiliev invariants can be built through the quantum groups.

In the next few sections we shall discuss the physical background behind many of the mathematical ideas discussed so far in this introduction to knot invariants.

VII. A QUICK REVIEW OF QUANTUM MECHANICS

To recall principles of quantum mechanics it is useful to have a quick historical recapitulation. Quantum mechanics really got started when Louis de Broglie introduced the fantastic notion that matter (such as an electron) is accompanied by a wave that guides its motion and produces interference phenomena just like the waves on the surface of the ocean or the diffraction effects of light going through a small aperture.

de Broglie's idea was successful in explaining the properties of atomic spectra. In this domain, his wave hypothesis led to the correct orbits and spectra of atoms, formally solving a puzzle that had been only described in ad hoc terms by the preceding theory of Niels Bohr. In Bohr's theory of the atom, the electrons are restricted to move only in certain elliptical orbits. These restrictions are placed in the theory to get agreement with the known atomic spectra,

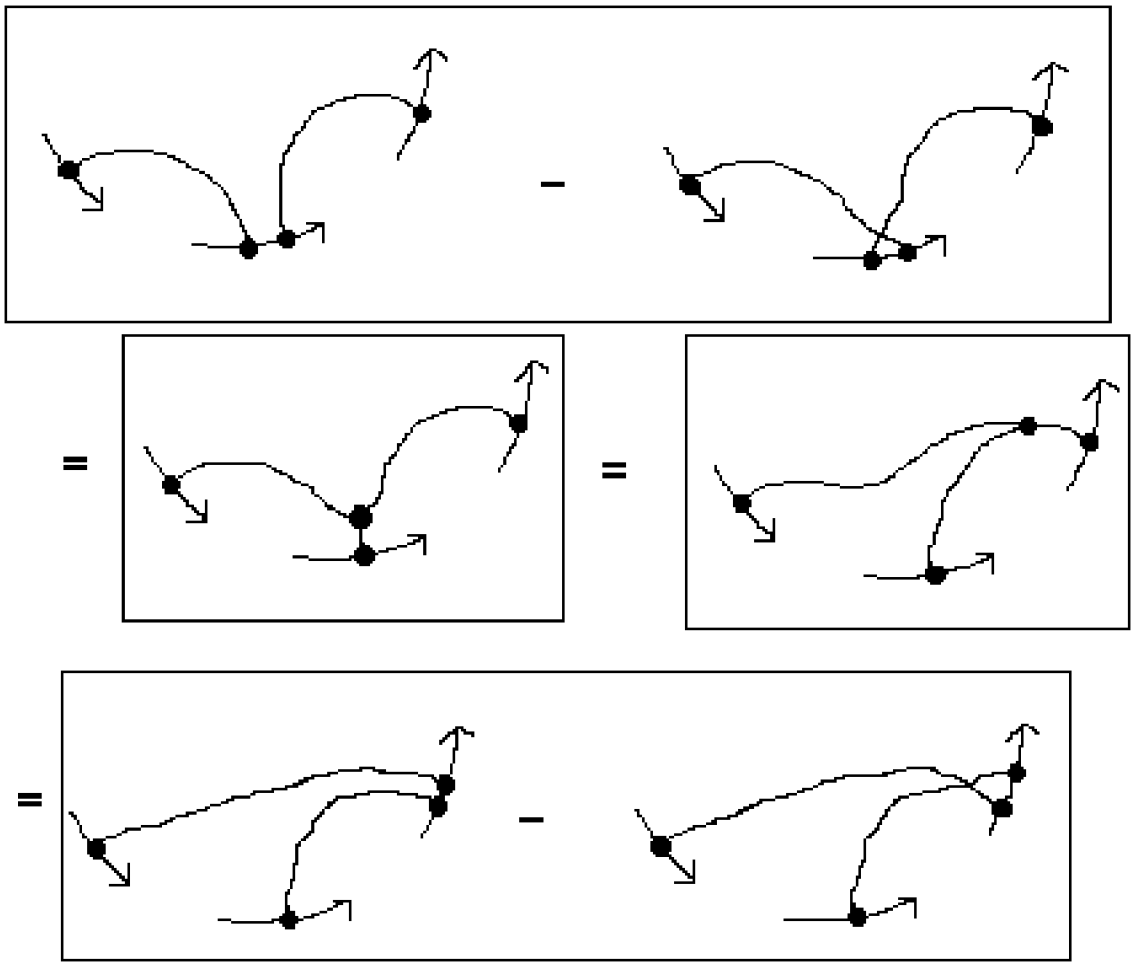


FIGURE 41 Proof of Four Term—Relation.

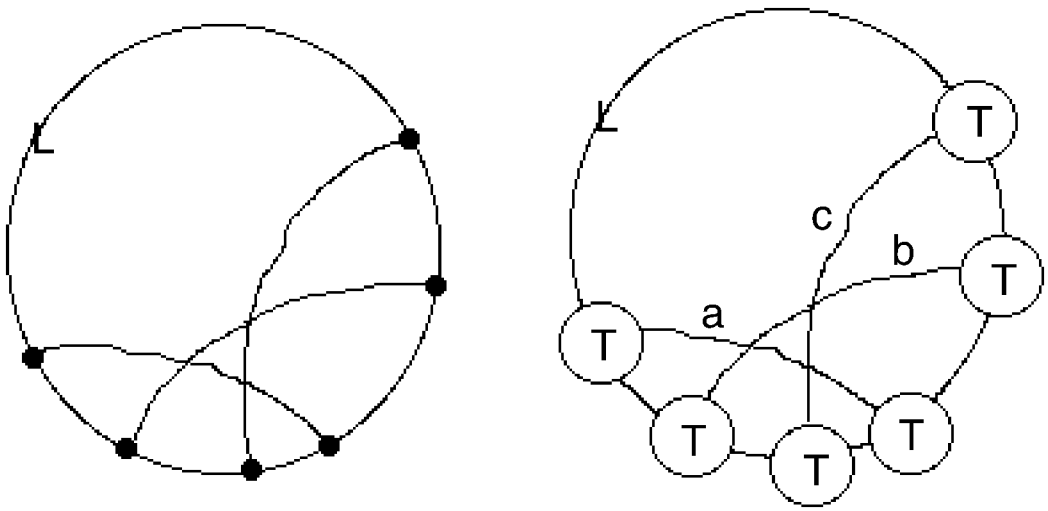


FIGURE 42 Lic Algebra Insertion in a Chord Diagram.

and to avoid a paradox! The paradox arises if one thinks of the electron as a classical particle orbiting the nucleus of the atom. Such a particle is undergoing acceleration in order to move in its orbit. Accelerated charged particles emit radiation. Therefore the electron should radiate away its energy and spiral into the nucleus! Bohr commanded the electron to only occupy certain orbits and thereby avoided the spiral death of the atom—at the expense of logical consistency.

de Broglie hypothesized a wave associated with the electron and he said that an integral multiple of the length of this wave must match the circumference of the electron orbit. Thus, not all orbits are possible, only those where the wave pattern can "bite its own tail." The mathematics works out, providing an alternative to Bohr's picture.

de Broglie had waves, but he did not have an equation describing the spatial distribution and temporal evolution of these waves. Such an equation was discovered by Erwin Schrödinger. Schrödinger relied on inspired guesswork, based on de Broglie's hypothesis, and produced a wave equation, known ever since as the Schrödinger equation. Schrödinger's equation was enormously successful, predicting fine structure of the spectrum of hydrogen and many other aspects of physics. Suddenly a new physics, *quantum mechanics*, was born from this musical hypothesis of de Broglie.

Along with the successes of quantum mechanics came a host of extraordinary problems of interpretation. What is the status of this wavefunction of Schrödinger and de Broglie. Does it connote a new element of physical reality? Is matter "nothing but" the patterning of waves in a continuum? How can the electron be a wave and still have the capacity to instantiate a very specific event at one place and one time (such as causing a bit of phosphor to glow *there* on a television screen)? Max Born developed a statistical interpretation of the wavefunction wherein the wave determines a probability for the appearance of the localized particulate phenomenon that one wanted to call an "electron." In this story the wavefunction ψ takes values in the complex numbers and the associated probability is $\psi^*\psi$, where ψ^* denotes the complex conjugate of ψ. Mathematically, this is a satisfactory recipe for dealing with the theory, but it leads to further questions about the exact character of the statistics. If quantum theory is inherently statistical, then it can give no complete information about the motion of the electron. In fact, there may be no such complete information available even in principle. Electrons manifest as particles when they are observed in a certain manner and as waves when they are observed in another, complementary manner. This is a capsule summary of the view taken by Bohr, Heisenberg, and Born. Others, including de Broglie, Einstein, and Schrödinger, hoped for a more direct and deterministic theory of nature.

As we shall see, the statistical nature of quantum theory has a formal side that can be exploited to understand the topological properties of such mundane objects as knotted ropes in space and spaces constructed by identifying the sides of polyhedra. These topological applications of quantum mechanical ideas are exciting in their own right. They may shed light on the nature of quantum theory itself.

In this section we review a bit of the mathematics of quantum theory. Recall the equation for a wave

$$f(x,t) = \sin[(2\pi/l)(x - ct)].$$

With x interpreted as the position and t and as the time, this function describes a sinusoidal wave traveling with velocity c. We define the wave number $k = 2\pi/l$ and the frequency $w = (2\pi c/l)$, where l is the wavelength. Thus we can write $f(x,t) = \sin(kx - wt)$. Note that the velocity c of the wave is given by the ratio of frequency to wave number, $c = w/k$.

de Broglie hypothesized two fundamental relationships: between energy and frequency, and between momentum and wave number. These relationships are summarized in the equations

$$E = \hbar w, \qquad p = \hbar k,$$

where E denotes the energy associated with a wave and p denotes the momentum associated with the wave. Here $\hbar = h/2\pi$, where h is Planck's constant.

For de Broglie the discrete energy levels of the orbits of electrons in an atom of hydrogen could be explained by restrictions on the vibrational modes of waves associated with the motion of the electron. His choices for the energy and the momentum in relation to a wave are not arbitrary. They are designed to be consistent with the notion that the wave or wave packet moves along with the electron. That is, the velocity of the wave packet is designed to be the velocity of the "corresponding" material particle.

It is worth illustrating how de Broglie's idea works. Consider two waves whose frequencies are very nearly the same. If we superimpose them (as a piano tuner superimposes a tuning fork with the vibration of the piano string), then there will be a new wave produced by the interference of the original waves. This new wave pattern will move at its own velocity, different (and generally smaller) than the velocity of the original waves. To be specific, let

$$f(x,t) = \sin(kx - wt), \qquad g(x,t) = \sin(k'x - w't).$$

Let

$$\begin{aligned} h(x,t) &= \sin(kx - wt) + \sin(k'x - w't) \\ &= f(x,t) + g(x,t). \end{aligned}$$

A little trigonometry shows that

$$h(x,t) = \cos\{[(k-k')/2]x - [(w-w')/2]t\} \times \sin\{[(k+k')/2]x - [(w+w')/2]t\}.$$

If we assume that k and k' are very close and that w and w' are very close, then $(k+k')/2$ is approximately k, and $(w+w')/2$ is approximately w. Thus $h(x,t)$ can be represented by

$$H(x,t) = \cos[(\delta k/2)x - (\delta w/2)t]\, f(x,t),$$

where $\delta k = (k-k')/2$ and $dw = (w-w')/2$. This means that the superposition $H(x,t)$ behaves as the waveform $f(x,t)$ carrying a slower moving "wave packet" $G(x,t) = \cos[(\delta k/2)x - (\delta w/2)t]$. See Fig. 43.

Since the wave packet (seen as the clumped oscillations in Fig. 43) has the equation $G(x,t) = \cos[(\delta k/2)x - (\delta w/2)t]$, we see that that the velocity of this wave packet is $v_g = \delta w/\delta k$. Recall that wave velocity is the ratio of frequency to wave number. Now according to de Broglie, $E = \hbar w$ and $p = \hbar k$, where E and p are the energy and momentum, respectively, associated with this wave packet. Thus we get the formula $v_g = dE/dp$. In other words, the velocity of the wave packet is the rate of change of its energy with respect to its momentum. Now this is exactly in accord with the well-known classical laws for a material particle! For such a particle, $E = mv^2/2$ and $p = mv$. Thus $E = p^2/2m$ and $dE/dp = p/m = v$.

It is this astonishing concordance between the simple wave model and the classical notions of energy and momentum that initiated the beginnings of quantum theory.

A. Schrödinger's Equation

Schrödinger answered the question, Where is the wave equation for de Broglie's waves? Writing an elementary wave in complex form

$$\psi = \psi(x,t) = \exp[i(kx - wt)],$$

we see that we can extract de Broglie's energy and momentum by differentiating:

$$i\hbar\, \partial\psi/\partial t = E_\psi \quad \text{and} \quad -i\hbar\, \partial\psi/\partial x = p_\psi.$$

This led Schrödinger to postulate *the identification of dynamical variables with operators* so that the first equation,

$$i\hbar\, \partial\psi/\partial t = E_\psi,$$

is promoted to the status of an equation of motion, while the second equation becomes the definition of momentum as an operator:

$$p = -i\hbar\, \partial/\partial x.$$

Once p is identified as an operator, the numerical value of momentum is associated with an eigenvalue of this operator, just as in the example above. In our example $p_\psi = \hbar k_\psi$.

In this formulation, the position operator is just multiplication by x itself. Once we have fixed specific operators for position and momentum, the operators for other

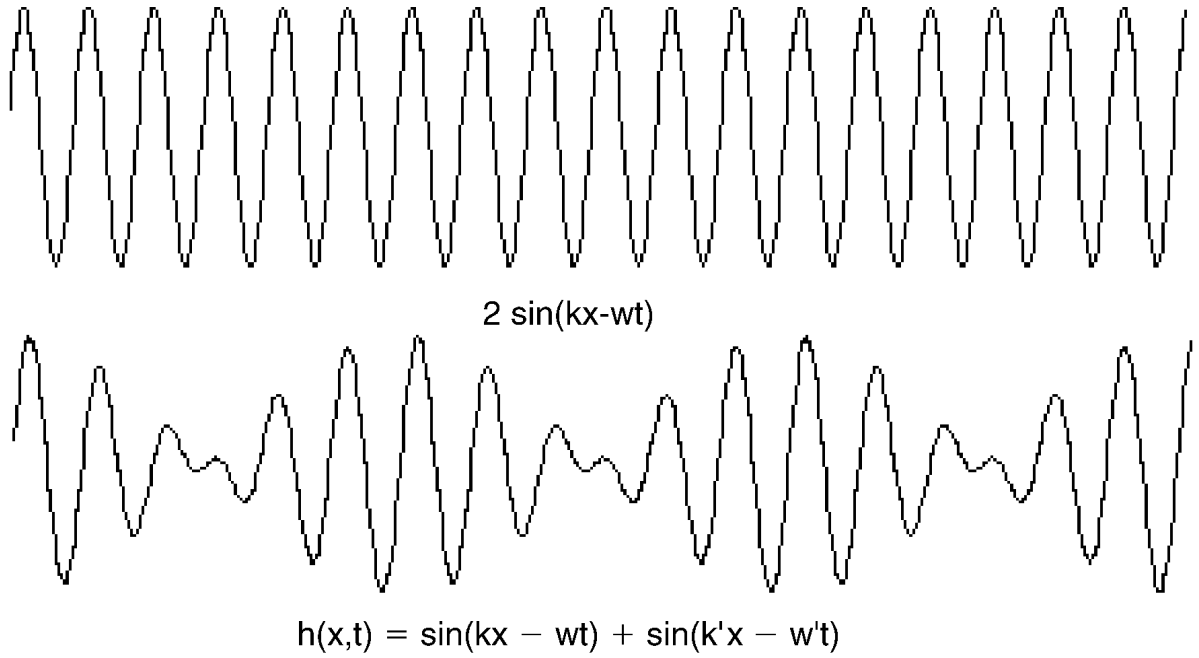


FIGURE 43 Waves and wave packets.

physical quantities can be expressed in terms of them. We obtain the energy operator by substitution of the momentum operator in the classical formula for the energy:

$$E = (1/2)mv^2 + V$$
$$E = p^2/2m + V$$
$$E = -(\hbar^2/2m)\,\partial^2/\partial x^2 + V.$$

Here V is the potential energy, and its corresponding operator depends upon the details of the application.

With this operator identification for E, Schrodinger's equation

$$i\hbar\,\partial\psi/\partial t = -(\hbar^2/2m)\,\partial^2\psi/\partial x^2 + V_\psi$$

is an equation in the first derivatives of time and in second derivatives of space. In this form of the theory one considers general solutions to the differential equation and this in turn leads to excellent results in a myriad of applications.

In quantum theory, observation is modeled by the concept of eigenvalues for corresponding operators. The quantum model of an observation is a projection of the wavefunction into an eigenstate.

An energy spectrum $\{E_k\}$ corresponds to wavefunctions ψ satisfying the Schrödinger equation such that there are constants E_k with $E_\psi = E_k\psi$. An *observable* (such as energy) E is a Hermitian operator on a Hilbert space of wavefunctions. Since Hermitian operators have real eigenvalues, this provides the link with measurement for the quantum theory.

It is important to notice that there is no mechanism postulated in this theory for how a wavefunction is "sent" into an eigenstate by an observable. Just as mathematical logic need not demand causality behind an implication between propositions, the logic of quantum mechanics does not demand a specified cause behind an observation. This absence of an assumption of causality in logic does not obviate the possibility of causality in the world. Similarly, the absence of causality in quantum observation does not obviate causality in the physical world. Nevertheless, the debate over the interpretation of quantum theory has often led its participants into asserting that causality has been demolished in physics.

Note that the operators for position and momentum satisfy the equation $xp - px = \hbar i$. This corresponds directly to the equation obtained by Heisenberg on other grounds, stating that dynamical variables can no longer necessarily commute with one another. In this way, the points of view of de Broglie, Schrödinger, and Heisenberg came together, and quantum mechanics was born. In the course of this development, interpretations varied widely. Eventually, physicists came to regard the wavefunction not as a generalized wave packet, but as a carrier of information about possible observations. In this way of thinking $\psi^*\psi$ (ψ^* denotes the complex conjugate of y) represents the probability of finding the "particle" (a particle is an observable with local spatial characteristics) at a given point in spacetime.

B. Dirac Brackets

We now discuss Dirac's notation $\langle a \mid b\rangle$ (Dirac, 1958). In this notation $\langle a|$ and $|b\rangle$ are vectors and covectors, respectively. $\langle a \mid b\rangle$ is the evaluation of $\langle a|$ by $|b\rangle$, hence it is a scalar, and in ordinary quantum mechanics it is a complex number. One can think of this as the amplitude for the state to begin in "a" and end in "b." That is, there is a process that can mediate a transition from state a to state b. Except for the fact that amplitudes are complex valued, they obey the usual laws of probability. This means that if the process can be factored into a set of all possible intermediate states $c_1, c_2, \ldots, c_n$, then the amplitude for $a \to b$ is the sum of the amplitudes for $a \to c_i \to b$. Meanwhile, the amplitude for $a \to c_i \to b$ is the product of the amplitudes of the two subconfigurations $a \to c_i$ and $c_i \to b$. Formally we have

$$\langle a \mid b\rangle = \sum \langle a \mid c_i\rangle\langle c_i \mid b\rangle,$$

where the summation is over all the intermediate states $i = 1, \ldots, n$.

In general, the amplitude for mutually disjoint processes is the *sum* of the amplitudes of the individual processes. The amplitude for a configuration of disjoint processes is the *product* of their individual amplitudes.

Dirac's division of the amplitudes into *bras* $\langle a|$ and *kets* $|b\rangle$ is done mathematically by taking a vector space V (a Hilbert space, but it can be finite dimensional) for the bras: $\langle a|$ belongs to V. The dual space V^* is the home of the kets. Thus $|b\rangle$ belongs to V^* so that $|b\rangle$ is a linear mapping $|b\rangle\colon V \to C$, where C denotes the complex numbers. We restore symmetry to the definition by realizing that an element of a vector space V can be regarded as a mapping from the complex numbers to V. Given $\langle a|\colon C \to V$, the corresponding element of V is the image of 1 (in C) under this mapping. In other words, $\langle a|(1)$ is a member of V. Now we have $\langle a|\colon C \to V$ and $|b\rangle\colon V \to C$. The composition $\langle a\|b\rangle = \langle a \mid b\rangle\colon C \to C$ is regarded as an element of C by taking the specific value $\langle a \mid b\rangle(1)$. The complex numbers are regarded as the "vacuum," and the entire amplitude $\langle a \mid b\rangle$ is a "vacuum-to-vacuum" amplitude for a process that includes the creation of the state a, its transition to b, and the annihilation of b to the vacuum once more.

Dirac notation has a life of its own. Let $P = |y\rangle\langle x|$ and $\langle x\|y\rangle = \langle x|y\rangle$. Then

$$PP = |y\rangle\langle x\|y\rangle\langle x| = |y\rangle\langle x \mid y\rangle\langle x| = \langle x \mid y\rangle P.$$

Up to a scalar multiple, P is a projection operator. That is, if we let $Q = P/\langle x \mid y\rangle$, then

$$QQ = PP/\langle x \mid y\rangle\langle x \mid y\rangle$$
$$= \langle x \mid y\rangle P/\langle x \mid y\rangle\langle x \mid y\rangle = P/\langle x \mid y\rangle = Q.$$

Thus, $QQ = Q$. In this language, the completeness of intermediate states becomes the statement that a certain sum of projections is equal to the identity: Suppose that $\sum_i |c_i\rangle\langle c_i| = 1$ (summing over i) with $\langle c_i \mid c_i\rangle = 1$ for each i. Then

$$\begin{aligned}\langle a \mid b\rangle &= \langle a \| b\rangle \\ &= \langle a| \sum |c_i\rangle\langle c_i \| b\rangle \\ &= \sum \langle a \| c_i\rangle\langle c_i \| b\rangle \\ &= \sum \langle a \mid c_i\rangle\langle c_i \mid b\rangle.\end{aligned}$$

Iterating this principle of expansion over a complete set of states leads to the most primitive form of the Feynman integral (Feynman and Hibbs, 1965). Imagine that the initial and final states a and b are points on the vertical lines $x = 0$ and $x = n + 1$, respectively, in the x–y plane, and that $(c(k)i(k), k)$ is a given point on the line $x = k$ for $0 < i(k) < m$. Suppose that the sum of projectors for each intermediate state is complete. That is, we assume that following sum is equal to one, for each k from 1 to $n - 1$:

$$|c(k)1\rangle\langle c(k)1| + \cdots + |c(k)m\rangle\langle c(k)m| = 1.$$

Applying the completeness iteratively, we obtain the following expression for the amplitude $\langle a \mid b\rangle$:

$$\langle a \mid b\rangle = \sum \langle a \mid c(1)i(1)\rangle\langle c(1)i(1) \mid c(2)i(2)\rangle \cdots$$
$$\times \langle c(n)i(n) \mid b\rangle,$$

where the sum is taken over all $i(k)$ ranging between 1 and m, and k ranging between 1 and n. Each term in this sum can be construed as a combinatorial path from a to b in the two-dimensional space of the x–y plane. Thus the amplitude for going from a to b is seen as a summation of contributions from all the "paths" connecting a to b. Feynman used this description to produce his famous path integral expression for amplitudes in quantum mechanics. His path integral takes the form

$$\int dP \exp(iS),$$

where i is the square root of minus one, the integral is taken over all paths from point a to point b, and S is the *action* for a particle to travel from a to b along a given path. For the quantum mechanics associated with a classical (Newtonian) particle the action S is given by the integral along the given path from a to b of the difference $T - V$, where T is the classical kinetic energy and V is the classical potential energy of the particle.

The beauty of Feynman's approach to quantum mechanics is that it shows the relationship between the classical and the quantum in a particularly transparent manner. Classical motion corresponds to those regions where all nearby paths contribute constructively to the summation. This classical path occurs when the variation of the action is null. To ask for those paths where the variation of the action is zero is a problem in the calculus of variations, and it leads directly to Newton's equations of motion. Thus, with the appropriate choice of action, classical and quantum points of view are unified.

The drawback of this approach lies in the unavailability at the present time of an appropriate measure theory to support all cases of the Feynman integral.

To summarize, Dirac's notation shows at once how the probabilistic interpretation for amplitudes is tied to the vector space structure of the space of states of the quantum mechanical system. Our strategy for bringing forth relations between quantum theory and topology is to pivot on the Dirac bracket. The Dirac bracket acts as intermediate between notation and linear algebra. In a very real sense, the connection of quantum mechanics with topology is an amplification of Dirac notation.

The next two sections discuss how topological invariants in low-dimensional topology are related to amplitudes in quantum mechanics. In these cases the relationship with quantum mechanics is primarily mathematical. Ideas and techniques are borrowed. It is not yet clear what the effect of this interaction will be on the physics itself.

VIII. KNOT AMPLITUDES

At the end of the last section we said that the connection of quantum mechanics with topology is an amplification of Dirac notation. Consider first a circle in a spacetime plane with time represented vertically and space horizontally (Fig. 44). The circle represents a vacuum-to-vacuum process that includes the creation of two "particles" (Fig. 45) and their subsequent annihilation (Fig. 46). In accord with our previous description, we could divide the circle into two parts, creation (a) and annihilation (b), and consider the amplitude $\langle a \mid b\rangle$. Since the diagram for the creation of the two particles ends in two separate points, it is natural to take a vector space of the form $V \otimes V$ (the tensor product of V with V) as the target for the bra and as the domain of the ket.

We imagine at least one particle property being catalogued by each dimension of V. For example, a basis of V could enumerate the spins of the created particles. If $\{e_a\}$ is a basis for V, then $\{e_a \otimes e_b\}$ forms a basis for

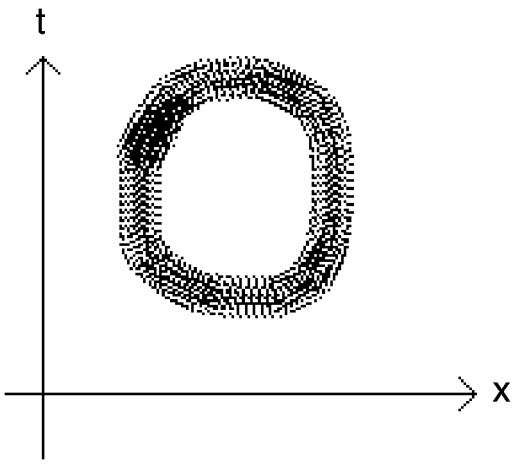


FIGURE 44 A Circle in Space Time.

$V \otimes V$. The elements of this new basis constitute all possible combinations of the particle properties. Since such combinations are multiplicative, the tensor product is the appropriate construction.

In this language the creation ket is a map CUP,

$$\text{CUP} = \langle a|: \quad C \to V \otimes V$$

and the annihilation bra is a mapping CAP,

$$\text{CAP} = |b\rangle: \quad V \otimes V \to C.$$

The first hint of topology comes when we realize that it is possible to draw a much more complicated simple closed curve in the plane that is nevertheless decomposed with respect to the vertical direction into many CUPs and CAPs. In fact, any non-self-intersecting differentiable curve can be rigidly rotated until it is in general position with respect to the vertical. It will then be seen to be decomposed into these minima and maxima. Our prescriptions for amplitudes suggest that we regard any such curve as an amplitude via its description as a mapping from C to C.

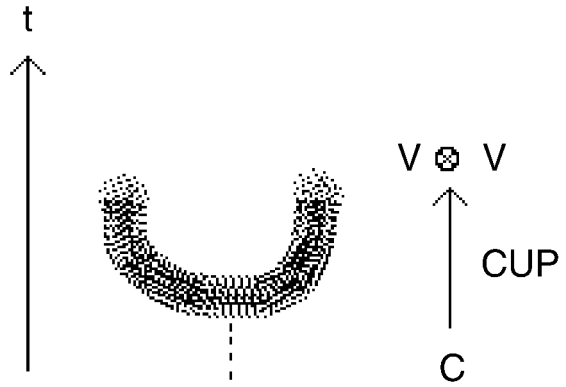


FIGURE 45 The Cup.



FIGURE 46 The Cap.

Each simple closed curve gives rise to an amplitude, but any simple closed curve in the plane is isotopic to a circle, by the Jordan curve theorem. If these are *topological amplitudes*, then they should all be equal to the original amplitude for the circle. Thus the question: What condition on creation and annihilation will insure topological amplitudes? The answer derives from the fact that all isotopies of the simple closed curves are generated by the cancellation of adjacent maxima and minima as illustrated in Fig. 47.

In composing mappings it is necessary to use the identifications

$$(V \otimes V) \otimes V = V \otimes (V \otimes V) \quad \text{and} \quad V \otimes k = k \otimes V = V.$$

Thus in Fig. 47, the composition on the left is given by

$$\begin{aligned} V &= V \otimes k \xrightarrow[1 \otimes \text{CUP}]{} V \otimes (V \otimes V) \\ &= (V \otimes V) \otimes V \xrightarrow[\text{CAP} \otimes 1]{} k \otimes V = V. \end{aligned}$$

This composition must equal the identity map on V (denoted 1 here) for the amplitudes to have a proper image of the topological cancellation.

This condition is said very simply by taking a matrix representation for the corresponding operators. Specifically, let $\{e_1, e_2, \ldots, e_n\}$ be a basis for V. Let $e_{ab} = e_a \otimes e_b$ denote the elements of the tensor basis for $V \otimes V$. Then there are matrices M_{ab} and M^{ab} such that $\text{cup}(1) = \sum M^{ab} e_{ab}$ with the summation taken over all values of a and b from 1 to n. Similarly, cap is described by $\text{cap}(e_{ab}) = Mab$. Thus, the amplitude for the circle is $\text{cap}[\text{cup}(1)] = \text{cap} \sum M^{ab} e_{ab} = \sum M^{ab} M_{ab}$.

In general, the value of the amplitude on a simple closed curve is obtained by translating it into an "abstract tensor expression" in the M_{ab} and M^{ab} and then summing over these products for all cases of repeated indices.

Returning to the topological conditions, we see that they are just that the matrices (M_{ab}) and (M^{ab}) are inverses in the sense that $\sum M^{ai} M_{ib} = \int_b^a$ and $\sum M_{ai} M^{ib} = \int_a^b$ are identity matrices.

Figure 48 shows the diagrammatic representative of the equation $\sum M^{ai} M_{ib} = \int_b^a$.

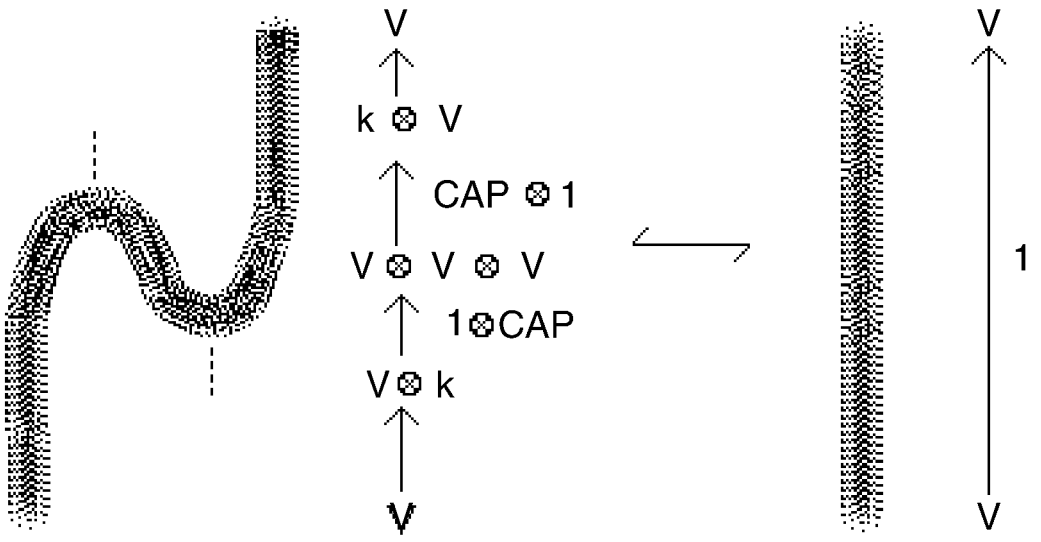


FIGURE 47 Cup–Cap Cancellation.

In the simplest case cup and cap are represented by 2×2 matrices. The topological condition implies that these matrices are inverses of each other. Thus the problem of the existence of topological amplitudes is very easily solved for simple closed curves in the plane.

Now we go to knots and links. Any knot or link can be represented by a picture that is configured with respect to a vertical direction in the plane. The picture will decompose into minima (creations) maxima (annihilations) and crossings of the two types shown below. (Here I consider knots and links that are unoriented. They do not have an intrinsic preferred direction of travel.) In Fig. 49, next to each of the crossings we have indicated mappings of $V \otimes V$ to itself, called R and $\bar{R}$, respectively. These mappings represent the transitions corresponding to these elementary configurations.

That R and $\bar{R}$ really must be inverses follows from the isotopy shown in Fig. 50 (this is the second Reidemeister move.)

We now have the vocabulary of cup, cap, R, and $\bar{R}$. Any knot or link can be written as a composition of these fragments, and consequently a choice of such mappings determines an amplitude for knots and links. In order for such an amplitude to be topological we want it to be

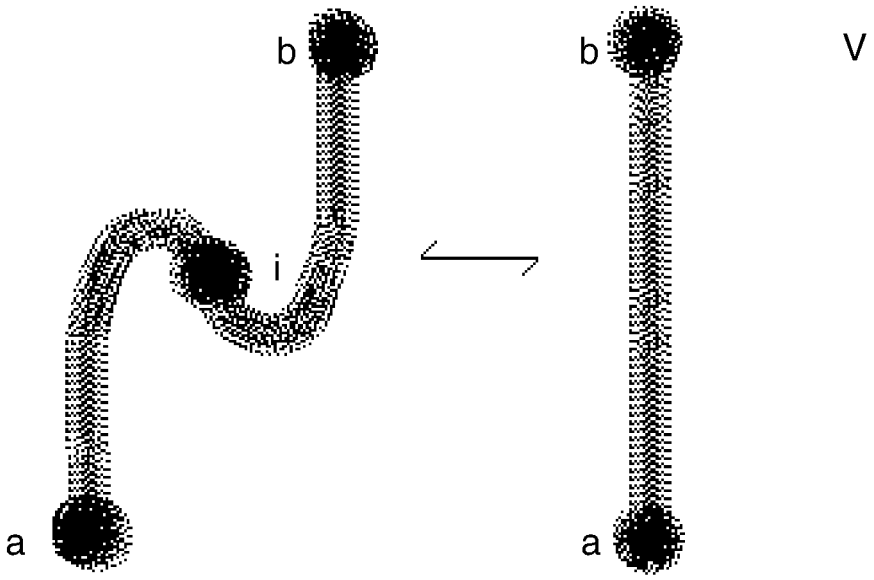


FIGURE 48 Cancellation.

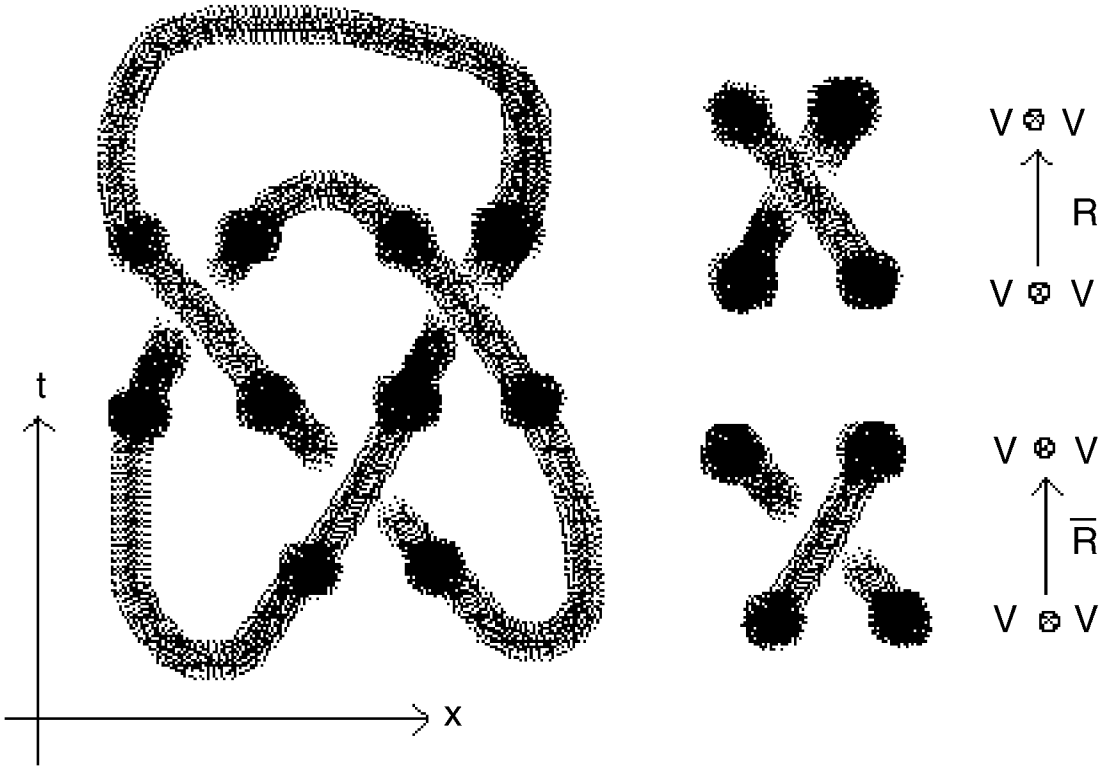


FIGURE 49

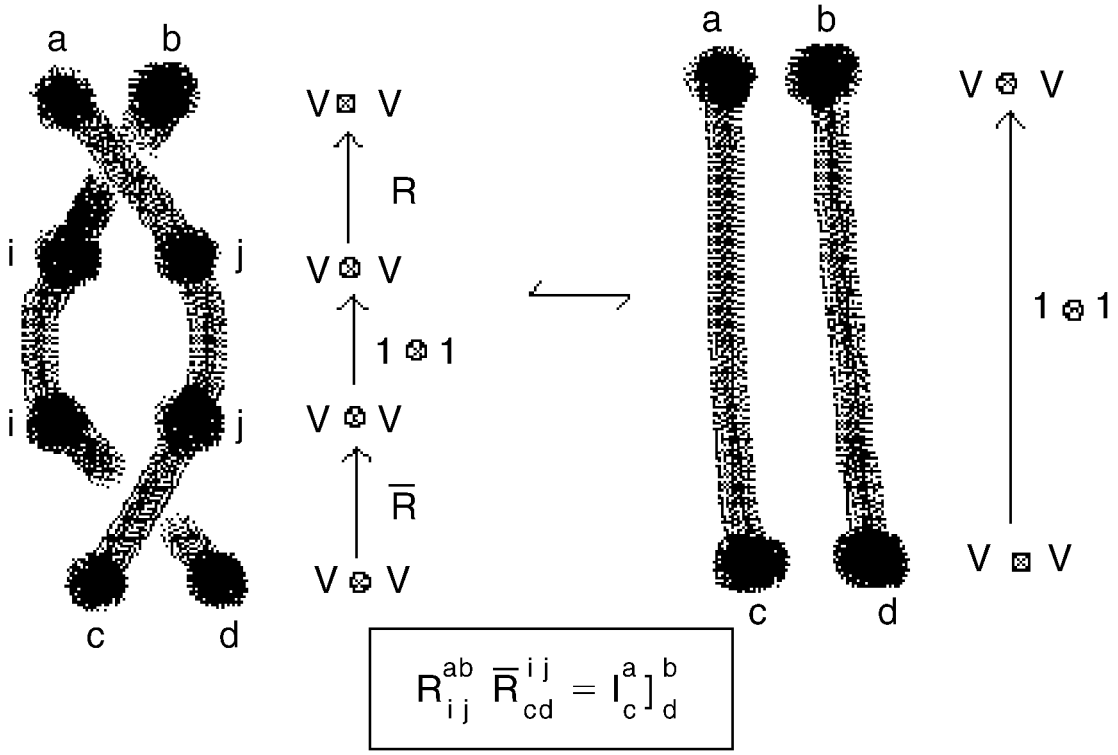


FIGURE 50 Knot as Vacuum–Vacuum Expectation.

invariant under the list of local moves on the diagrams shown in Fig. 51. These moves are an augmented list of the Reidemeister moves (see Fig. 4), adjusted to take care of the fact that the diagrams are arranged with respect to a given direction in the plane.

The equivalence relation generated by these moves is called *regular isotopy*. It is one move short of the relation known as *ambient isotopy*. The missing move is the first Reidemeister move shown in Fig. 4.

In the first Reidemeister move, a curl in the diagram is created or destroyed. Ambient isotopy (generated by all the Reidemeister moves) corresponds to the full topology of knots and links embedded in three-dimensional space. Two link diagrams are ambient isotopic via the Reidemeister moves if and only if there is a continuous family of embeddings in three dimensions leading from one link to the other. The moves give a combinatorial reformulation of the spatial topology of knots and links.

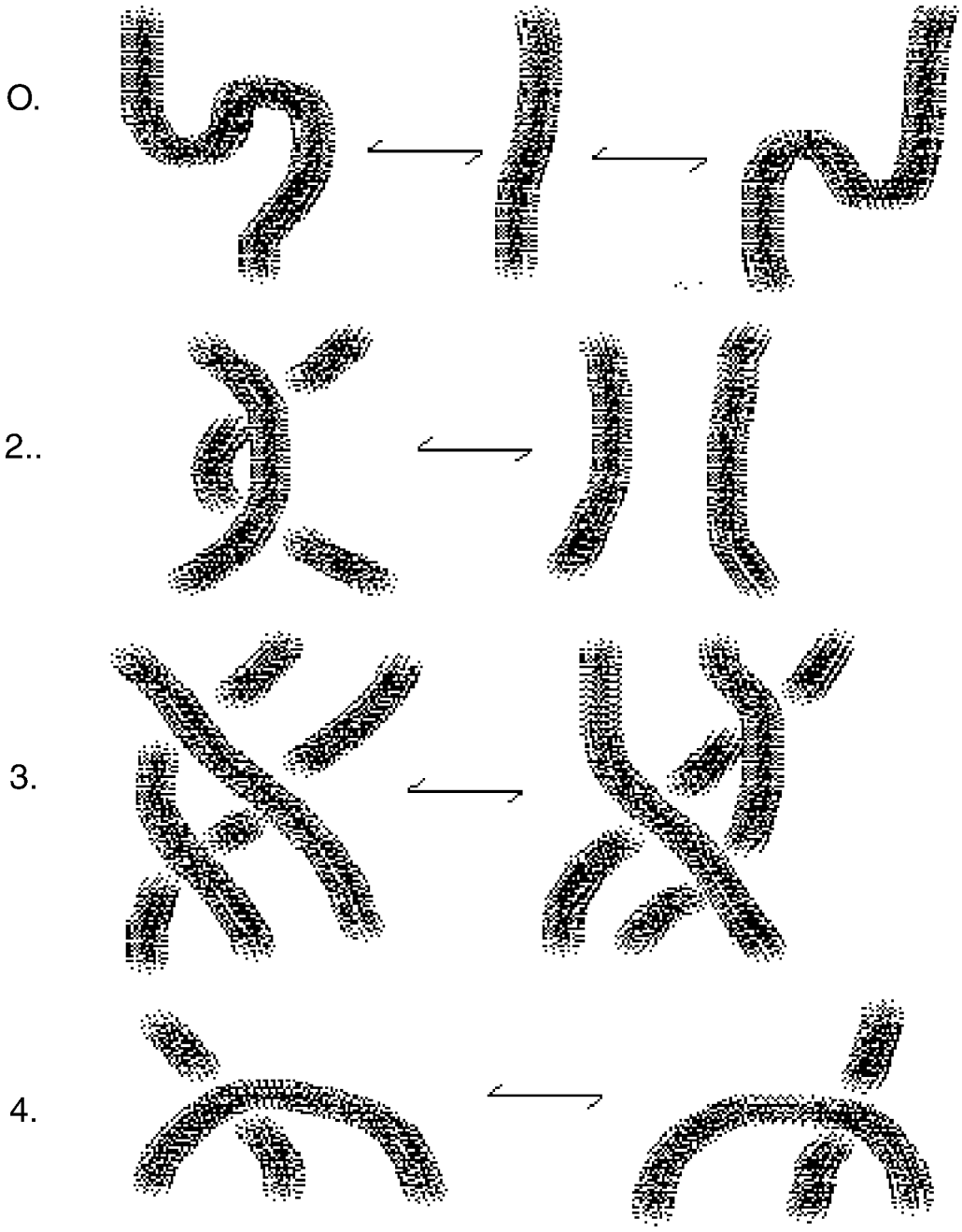


FIGURE 51 Augmented Reidemeister moves for regular isotopy.

By ignoring the first Reidemeister move, we allow the possibility that these diagrams can model framed links, that is, links with a normal vector field or, equivalently, embeddings of curves that are thickened into bands. It turns out to be fruitful to study invariants of regular isotopy. In fact, one can usually normalize an invariant of regular isotopy to obtain an invariant of ambient isotopy. We have already discussed this phenomenon with the bracket polynomial in Section 4.

As the reader can see, we have already discussed the algebraic meaning of moves 0 and 2. The other moves translate into very interesting algebra. Move 3, when translated into algebra, is the famous Yang–Baxter equation. The Yang–Baxter equation occurred for the first time in problems related to exactly solved models in statistical mechanics. All the moves taken together are directly related to the axioms for a quasi-triangular Hopf algebra ("quantum group"). We shall not go into this connection here.

There is an intimate connection between knot invariants and the structure of generalized amplitudes, as we have described them in terms of vector space mappings associated with link diagrams. This strategy for the construction of invariants is directly motivated by the concept of an amplitude in quantum mechanics. It turns out that the invariants that can actually be produced by this means (that is, by assigning finite dimensional matrices to the caps, cups and crossings) are incredibly rich. They encompass, at present, all of the known invariants of polynomial type (Alexander polynomial, Jones polynomial, and their generalizations).

It is now possible to indicate the construction of the Jones polynomial via the bracket polynomial as an amplitude, by specifying its matrices.

The cups and the caps are defined by $(M_{ab}) = (M^{ab}) = M$, where M is the 2×2 matrix (with $ii = -1$).

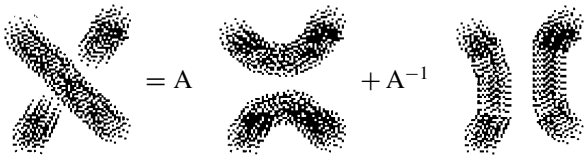
Note that $MM = I$, where I is the identity matrix. Note also that the amplitude for the circle is

$$\sum M_{ab}M^{ab} = \sum M_{ab}M_{ab} = \sum (M_{ab})^2$$
$$= (iA)^2 + (-iA^{-1})^2 = -A^2 - A^{-2}.$$

The matrix R is then defined by the equation

$$R^{ab}_{cd} = AM^{ab}M_{cd} + A^{-1}\int_c^a \int_d^b .$$

Since, diagrammatically, we identify R with a (right-handed) crossing, this equation can be written diagrammatically as



Taken together with the loop value of $A^2 - A^{-2}$,



these equations can be regarded as a recursive algorithm for computing the amplitude. This algorithm is the bracket state model for the (unnormalized) Jones polynomial. This model can be studied on its own grounds as we have already done in section 4.

IX. TOPOLOGICAL QUANTUM FIELD THEORY, FIRST STEPS

In order to further justify the idea of topology in relation to the amplification of Dirac notation, consider the following scenario. Let M be a three-dimensional manifold; that is, a space that is locally homeomorphic to Euclidean three-dimensional space. Suppose that F is a closed orientable surface inside M dividing M into two pieces M_1 and M_2. These pieces are 3-manifolds with boundary. They meet along the surface F. Now consider an amplitude $\langle M_1 \mid M_2 \rangle = Z(M)$. The form of this amplitude generalizes our previous considerations, with the surface F constituting the distinction between the "preparation" M_1 and the "detection" M_2. This generalization of the Dirac amplitude $\langle a \mid b \rangle$ amplifies the notational distinction consisting of the vertical line of the bracket to a topological distinction in a space M. The amplitude $Z(M)$ will be said to be a *topological amplitude for M* if it is a topological invariant of the 3-manifold M. Note that a topological amplitude does not depend upon the choice of surface F that divides M.

From a physical point of view the independence of the topological amplitude on the particular surface that divides the 3-manifold is the most important property. An amplitude arises in the condition of one part of the distinction carved in the 3-manifold acting as "the observed" and the other part of the distinction acting as "the observer." If the amplitude is to reflect physical (read topological) information about the underlying manifold, then it should not depend upon this particular decomposition into observer and observed. The same remarks apply to 4-manifolds and interface with ideas in relativity. We mention 3-manifolds because it is possible to describe many examples of topological amplitudes in three dimensions. The matter of four-dimensional amplitudes is a topic of current research. The notion that an amplitude be independent of the distinction producing it is prior to topology. Topological invariance

of the amplitude is a convenient and fundamental way to produce such independence.

This sudden jump to topological amplitudes has its counterpart in mathematical physics. Witten (1989) proposed a formulation of a class of 3-manifold invariants as generalized Feynman integrals taking the form $Z(M)$, where

$$Z(M) = \int dA e^{\frac{ik}{4\pi} S(M,A)}.$$

Here M denotes a 3-manifold without boundary and A is a gauge field (also called a gauge potential or gauge connection) defined on M. The gauge field is a one-form on M with values in a representation of a Lie algebra. The group corresponding to this Lie algebra is said to be the *gauge group* for this particular field. In this integral the "action" $S(M, A)$ is taken to be the integral over M of the trace of the Chern–Simons three-form $CS = A \wedge dA + (\frac{2}{3}) A \wedge A \wedge A$. (The product is the wedge product of differential forms.)

Instead of integrating over paths, the integral $Z(M)$ integrates over all gauge fields modulo gauge equivalence. This generalization from paths to fields is characteristic of quantum field theory. Quantum field theory was designed in order to accomplish the quantization of electromagnetism. In quantum electrodynamics the classical entity is the electromagnetic field. The question posed in this domain is to find the value of an amplitude for starting with one field configuration and ending with another. The analogue of all paths from point a to point b is "all fields from field A to field B."

Witten's integral $Z(M)$ is, in its form, a typical integral in quantum field theory. In its content $Z(M)$ is highly unusual. The formalism of the integral and its internal logic supports the existence of a large class of topological invariants of 3-manifolds and associated invariants of knots and links in these manifolds.

Invariants of 3-manifolds were initiated by Witten as functional integrals and at the same time defined in a combinatorial way by Reshetikhin and Turaev (1991). The Reshetikhin–Turaev definition proceeds in a way that is quite similar to the definition that we gave for the bracket model for the Jones polynomial in Section 1. It is an amazing fact that Witten's definition seems to give the very same invariants. We are not in a position to go into the details of this correspondence here. However, one theme is worth mentioning: For k large, the Witten integral is approximated by those gauge connections A for which $S(M, A)$ has zero variation with respect to change in A. These are the so-called *flat connections*. It is possible in many examples to calculate this contribution via both the functional integral and by the combinatorial definition of Reshetikhin and Turaev. In all cases, the two methods agree (see, e.g., Freed and Gompf, 1991; Lawrence and Rozansky, 1997). This is one of the pieces of evidence in a puzzle that everyone expects will eventually justify the formalism of the functional integral.

In order to obtain invariants of knots and links from Witten's integral, one adds an extra bit of machinery to the brew. The new machinery is the *Wilson loop*. The Wilson loop is an exponentiated version of integrating the gauge field along a loop K. We take this-loop K in three-space to be an embedding (a knot) or a curve with transversal self-intersections. It is usually indicated by the symbolism $\mathrm{tr}(\mathbb{P}e^{\oint_K A})$, where P denotes *path-ordered integration*, that is, we are integrating and exponentiating matrix-valued functions, and one must keep track of the order of the operations. The symbol tr denotes the trace of the resulting matrix.

With the help of the Wilson loop function on knots and links, Witten writes down a functional integral for link invariants in a 3-manifold M:

$$Z(M, K) = \int dA e^{\frac{ik}{4\pi} \int (M,A)} \mathrm{tr}(\mathbb{P}e^{\oint_K A}).$$

Here $S(M, A)$ is the Chern–Simons Lagrangian, as in the previous discussion.

If one takes the standard representation of the Lie algebra of $SU(2)$ as 2×2 complex matrices then it is a fascinating exercise to see that the formalism of $Z(S^3, K)$ ($S3$ denotes the three-dimensional sphere) yields the original Jones polynomial with the basic properties as discussed in Section 1. See Witten (1989) or Kauffman (1995) for discussions of this part of the heuristics.

This approach to link invariants crosses boundaries between different methods. There are close relations between $Z(S^3, K)$ and the invariants defined by Vassiliev (Kauffman, 1995), to name one facet of this complex crystal.

A. Links and the Wilson Loop

We shall now indicate an analysis of the formalism of this functional integral that reveals quite a bit about its role in knot theory. This analysis depends upon some key facts relating the curvature of the gauge field to both the Wilson loop and the Chern–Simons Lagrangian. To this end, let us recall the local coordinate structure of the gauge field $A(x)$, where x is a point in three-space. We can write $A(x) = A_h^a T_a \, dx^h$, where the index a ranges from 1 to m with the Lie algebra basis $\{T_1, T_2, T_3, \ldots, T_m\}$ and the index k goes from 1 to 3. For each choice of a and k, $A_k^a(x)$ is a smooth function defined on three-space. In $A(x)$ we sum over the values of repeated indices. The Lie algebra generators T_a are actually matrices

corresponding to a given representation of an abstract Lie algebra.

B. Difference Formula

One can deduce a difference formula for the Witten invariants from the formal properties of the functional integral. Let K_+ and K_- denote knots that differ at a single crossing with $+$ and $-$ signs, respectively, and $K_{\#\#}$ the result of replacing the crossing by a transverse singularity (i.e., with distinct tangent directions for the two local curve segments). We take $K_\#$ to denote the insertion of a graphical node at the transverse crossing, as we have done in our discussion of the Vassiliev invariant. The notation $K_{\#\#}$ indicates that the curve intersects itself in space at one point. Let $K_{\#\#}T_aT_a$ denote the result of placing the matrices of the Lie algebra basis into the Wilson line at the singular crossing as shown in Fig. 52.

These matrices become part of the big matrix product that generates the Wilson line. Then, up to order $(1/k)$ one has the difference relation

$$Z(K^+) - Z(K^-) = (4pi/k)Z\big(K_{\#\#}T_aT_a\big).$$

This formula is the key to unwrapping many properties of the knot invariants.

C. Graph Invariants and Vassiliev Invariants

Recall, from Section 5, that $V(G)$ is a Vassiliev invariant if

$$V_{K+} - V_{K-} = V_{K\#}.$$

$V(G)$ is said to be of *finite type* k if $V(G) = 0$ whenever $\#(G) > k$, where $\#(G)$ denotes the number of 4-valent nodes in the graph G. See Section 5.

With this definition in hand, let us return to the invariants derived from the functional integral $Z(K)$. We have that

$$Z(K+) - Z(K-) = \left(\frac{4\pi i}{k}\right) Z\big(K_{\#\#}T^aT^a\big).$$

This formula tells us that for the Vassiliev invariant associated with Z we have

$$Z(K\#) = \left(\frac{4\pi i}{k}\right) Z\big(K_{\#\#}T^aT^a\big).$$

Furthermore, if $V_j(K)$ denotes the coefficient of $\frac{4\pi i}{k}$ in the expansion of $Z(K)$ in powers of $(1/k)$, then the ambient difference formula implies that $(1/k)^j$ divides $Z(G)$ when G has j or more nodes. Hence $V_j(G) = 0$ if G has more than j nodes. Therefore $V_j(K)$ is a Vassiliev invariant of finite type. [This result was proved by Birman and Lin (1993) by different methods and by Bar-Natan (1995) by methods equivalent to ours.]

The fascinating thing is that the ambient difference formula, appropriately interpreted, actually tells us how to compute $V_k(G)$ when G has k nodes. This result is equivalent to the description of weight systems derived from Lie algebras that we described in Section 6. Thus the approach to link invariants via the functional integral motivates and explains the fundamental structure of Vassiliev invariants.

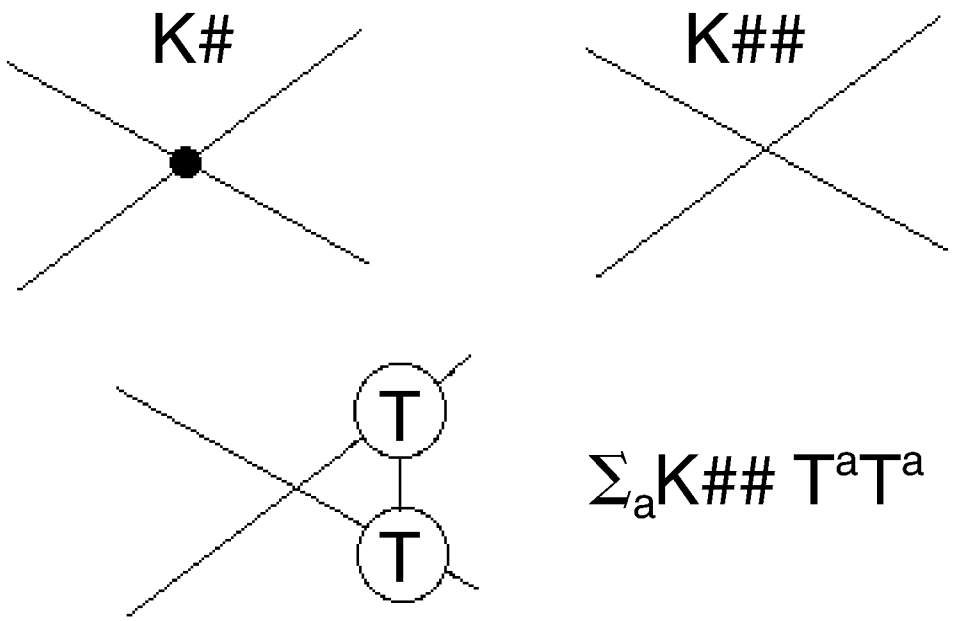


FIGURE 52 Lic Algebra Insertion.

This deep relationship between topological invariants in low-dimensional topology and quantum field theory in the sense of Witten's functional integral is still in its infancy. There will be many surprises in the future as we discover that what has so far been uncovered is only the tip of an iceberg.

ACKNOWLEDGMENT

It gives me great pleasure to thank Vaughan jones, Ed Witten, Nicolai Reshetikhin, Mario Rasetti, Sostenes Lins, Massimo Ferri, Lee Smolin, Louis Crane, David Yetter, Ray Lickorish, DeWitt Sumners, Hugh Morton, Joan Birman, John Conway, John Simon and Dennis Roseman for many conversations related to the topics of this paper. This research was partially supported by the National Science Foundation Grant DMS-2528707.

SEE ALSO THE FOLLOWING ARTICLES

QUANTUM MECHANICS • QUANTUM THEORY • STATISTICAL MECHANICS • TOPOLOGY, GENERAL

BIBLIOGRAPHY

Alexander, J. W. (1923). "Topological invariants of knots and links," *Trans. Amr. Math. Soc.* **20,** 275–306.

Atiyah, M. F. (1990). "The Geometry and Physics of Knots," Cambridge University Press, Cambridge.

Bar-Natan, D. (1995). "On the Vassiliev knot invariants," *Topology* **34,** 423–472.

Birman, J., and Lin, X. S. (1993). "Knot polynomials and Vassiliev's invariants," *Invent. Math.* **111,** 225–270.

Conway, J. H. (1970). "An enumeration of knots and links and some of their algebraic properties," *In* "Computational Problems in Abstract Algebra," Pergamon Press, New York, pp. 329–358.

Crowell, R. H., and Fox, R. H. (1963). "Introduction to Knot Theory," Ginn, Boston.

Dirac, P. A. M. (1958). "Principles of Quantum Mechanics," Oxford University Press, Oxford.

Feynman, R., and Hibbs, A. R. (1965). "Quantum Mechanics and Path Integrals," McGraw-Hill, New York.

Freed, D., and Gompf, R. (1991). "Computer calculations of Witten's 3-manifold invariants," *Comm. Math. Phys.* **41,** 79–117.

Jones, V. F. R. (1985). "A polynomial invariant for links via von Neumann algebrea," *Bull. Amr. Math. Soc.* **129,** 103–112.

Kauffman, L. H. (1980). "The Conway polynomial," *Topology* **20,** 101–108.

Kauffman, L. H. (1983). "Formal Knot Theory," Princeton University Press, Princeton, NJ.

Kauffman, L. H. (1987a). "State models and the Jones polynomial," *Topology* **26,** 395–407.

Kauffman, L. H. (1987b). "On Knots," Princeton University Press, Princeton, NJ.

Kauffman, L. H. (1989). "Statistical mechanics and the Jones polynomial." *In* "Proceedings of the 1986 Santa Cruz Conference on Artin's Braid Group," pp. 263–298, AMS, Providence, RI. [Reprinted in M. Rasetti, ed. (1990). "New Problems, Methods and Techniques in Quantum Field Theory and Statistical Mechanics," pp. 175–222, World Scientific, Singapore.]

Kauffman, L. H. (1993). "Knots and Physics," 2nd ed. World Scientific, Singapore.

Kauffman, L. H. (1995). "Functional integration and the theory of knots," *J. Math. Phys.* **36,** 2402–2429.

Kontsevich, M. (1994). "Feynman diagrams and low-dimensional topology." First European Congress of Mathematics, Vol. II (Paris, 1992), 97–121, Progro Math., 120, Birkhauser, Basel.

Lawrence, R., and Rozansky, L. (1997). "Witten–Reshetikhin–Turaev invariants of Seifert manifolds," Preprint.

Murasugi, K. (1987a). "The Jones polynomial and classical conjectures in knot theory," *Topology* **26,** 187–194.

Murasugi, K. (1987b). "Jones polynomials and classical conjectures in knot theory II," *Math. Proc. Camb. Phil. Soc.* **102,** 317–318.

Reidemeister, K. (1948, 1932). "Knotentheorie," Chelsea, New York and Julius Springer, Berlin.

Reshetikhin, N. Y., and Turaev, V. (1990). "Ribbon graphs and their invariants derived from quantum groups," *Comm. Math. Phys.* **127,** 1–26.

Reshetikhin, N. Y., and Turaev, V. (1991). "Invariants of three-manifolds via link polynomials and quantum groups," *Invent. Math.* **103,** 547–597.

Stanford, T. (1996). "Finite-type invariants of knots, links and graphs," Topology, **35,** 1027–1050.

Tnistlethwaite, M. (1987). " A spanning tree expansion of the Jones polynomial," Topology, **26,** pp 297–309.

Vassiliev, V. (1990). "Cohomology of knot spaces." *In* "Theory of Singularities and Its Applications" (V. I. Arnold, ed.), pp. 23–69, AMS, Providence, RI.

Witten, E. (1989). "Quantum field theory and the Jones polynomial," *Commun. Math. Phys.* **121,** 351–399.



Linear Optimization

C. Roos
Delft University of Technology/University of Leiden

I. Historical Background
II. The Simplex Method
III. Interior-Point Methods
IV. Related Topics
V. Further Extensions

GLOSSARY

Bounded problem An LO-problem whose objective function is bounded from above if the problem is a maximization problem, and bounded from below if it is a minimization problem.
Canonical LO-model All constraints are inequality constraints and all variables are nonnegative.
Feasible problem An LO-problem whose feasible region is nonempty.
Feasible region The set of points satisfying all the constraints.
Infeasible problem An LO-problem whose feasible region is empty.
Integer LO-problem LO-problem whose variables are required to be integral.
Interior-point condition (IPC) There exists a feasible point for which all inequality constraints in the LO-problem are satisfied strictly.
LO-relaxation The LO-problem that arises when the integrality condition on the variables in an integer LO-problem is removed.
Polynomial method A solution method for LO-problems whose running time depends polynomially on the size of the problem.
Size of a problem Length of a binary string needed to encode an LO-problem.
Standard LO-model An LO-problem in which all constraints are equality constraints and all variables are nonnegative.
Unbounded problem An LO-problem that is not bounded.

LINEAR OPTIMIZATION (LO) is the branch of mathematics that deals with minimizing (or maximizing) a linear function whose variables are subject to linear equality or inequality constraints. The *standard form* of the LO-problem has the form

$$\begin{aligned} \text{minimize} \quad & p = \sum_{j=1}^{n} c_j x_j \\ \text{subject to} \quad & \sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, \ldots, m \\ & x_j \geq 0, \quad j = 1, \ldots, n, \end{aligned}$$

or, in matrix form,

$$\begin{aligned} \text{minimize} \quad & p = c^T x \\ \text{subject to} \quad & Ax = b \\ & x \geq 0, \end{aligned} \tag{1}$$

where A is an $m \times n$ matrix and x, b, and c are vectors of dimension n, m, and m respectively. Here, $x \geq 0$ means that all entries of x are nonnegative and the superscript ^{T} refers to taking the transpose.

The relevance of the subject is due to the fact that real-life problems in many branches of applied sciences like economy, logistics, and engineering, can be modeled as LO-problems. As a result, LO is probably one of the most applied mathematical tools. The beauty of the subject is due to its rich mathematical theory, part of which has developed over the last twenty years.

I. HISTORICAL BACKGROUND

The field of LO[1] arose in the 1940s, due to important work of Dantzig, Kantorovich, Koopmans, Von Neumann, and Morgenstern. Dantzig was involved in the research project Scientific Computation of the Optimum Programs (SCOOP) at the U.S. Air Force. He visited Koopmans in June 1947 and told him of his work on linear models for the optimization of military operations. Koopmans immediately recognized the relevance of this approach for his economic models and made clear to Dantzig that the economists did not have an algorithm for solving such models. In the summer of 1947 Dantzig invented the Simplex method. Koopmans was the leader of a group of economists who developed the theory of optimal assignment of resources by using LO-models. For this work he received the Nobel prize, in 1975, together with Kantorovich.

In the autumn of 1947 another important meeting took place, when Dantzig visited Von Neumann. This meeting clarified the relation between Dantzig's work and the game theory as developed by Von Neumann and Morgenstern. Dantzig also heard for the first time of Farkas' lemma and the notion of duality in LO.

From a computational point of view the Simplex method is very simple. Only the elementary arithmetic operations addition, subtraction, division, and multiplication are performed on a rectangular table that initially contains the data of the LO-model. The number of operations, however, is so large that only relatively small models can be solved manually. Fortunately, the invention of the Simplex method coincided in time more or less with the invention of the electronic computer. The calculations could be automatized, thus paving the way for large-scale applications. Oil companies belonged to the early users of the computer code MPS/360 for LO released by IBM in 1966 and that could run on the IBM 360 computers that were introduced around that time. At the end of the 1960s and in the early 1970s, other software packages for LO became available: MPSX, and later MPSX/370 of IBM, MPS III of Exxon, the UMPIRE system for the UNIVAX 1108, APX III for the CDC computers, LAMPS (of John Forrest), and LINDO. For the history of the implementations of the Simplex method the reader is referred to W. Orchard-Hays (1990).

From a theoretical point of view the Simplex method has some properties that deserve mentioning here, because they were important in the history of LO. First, the method may *cycle*, i.e., it may happen that after hours of calculations a tableau is generated that already occurred before during the calculations. This phenomenon is called *cycling*. For a mathematical algorithm this is disastrous. It inplies that a computer program may run forever without terminating with a solution. Dantzig already recognized this danger and found a so-called *cycle-breaking* rule; as a consequence, the Simplex method always solves any LO-problem correctly.

A second property concerns the computational time. It may be expected that this time will grow with the size of the problem, i.e., with the number n of variables as well as with the number m of constraints. In practice the behavior is such that, as a rule of thumb, the computational time depends linearly on n and m. Many researchers tried to find a theoretical estimate for the computational time in terms of n and m. In all cases they ended up with a formula that is exponential in either n or m. In 1972 it became clear that this behavior is inherent to the Simplex method. In that year Klee and Minty gave an elegant example for which the computational time is exponential in n [Klee and Minty (1972)].

The result of Klee and Minty initiated a period of search for new, more efficient methods in the LO world. In 1979 the front page of the New York Times announced an important result: the Russian mathematician Khachiyan had found a new method with the desired property, the Ellipsoid method [Khachiyan (1979)]. This method is polynomial, i.e., the computational time depends polynomially on the size of the problem. Theoretically, this was an enormous breakthrough, from the practical point of view the result was disappointing. Computer programs based on the Ellipsoid method proved to be much less

[1]Historically, the field is named Linear Programming, or LP. This name was proposed shortly after World War II [Dantzig (1963)]. Since then the modern computer came to life and the word "programming" usually refers to the activity of writing computer programs. As a consequence, its use instead of the more natural word "optimization," gives rise to confusion [Williams (1990)].

efficient and less robust than those using the Simplex method. The result was an unpleasant tension between theory and practice: the theoretically efficient Ellipsoid method could not match the theoretically inefficient Simplex method.

A new breakthrough came in 1984. In that year the Indian mathematician Karmarkar published a method that reconciled theory and practice [Karmarkar (1984)]. It was again front page news. Karmarkar worked at the well-known American telephone company AT&T, and had found a completely new approach to the LO-problem. His so-called Projective method was polynomial and, as claimed by Karmarkar, was in practice at least 100 times faster than the Simplex method, especially for large-scale problems. The last claim gave rise to much commotion.

It is hard to find another mathematical result that caused so much excitement and dispute. The disputes were due to the fact that the published version of Karmarkar's method differed from the version implemented at AT&T. This was not made public. Later on the reason became clear. AT&T had made a big investment to design a computer program for LO that should be about 200 times faster than the commercial codes for solving LO-problems available at that time. A project group was formed including many prominent researchers, with the task to devise a new LO-program on the basis of Karmarkar's method. During its existence the size of the group grew from 20 to 200 researchers (R. E. Bixby and R. Vanderbei, Private communication). The system would be a hardware/software "complete" solution. For the hardware a $1 million vector/parallel machine of Alliant Computers was chosen.

About 5 years later, in 1989, the LO-package was launched under the name KORBX. In 1991, after selling one system to the military airlift command for about $4 million and one system to Delta airlines for $12 million, the venture was essentially abandoned. The LO-package was not portable, it ran only on the Alliant computer. In retrospect, this may have been the main reason for the failure of the project.

Outside AT&T much research activity took place. Initial implementations based on Karmarkar's paper proved to be about 100 times slower than the Simplex-based codes. It looked as if the Simplex method would survive this attack from the Projective method, just as it had survived the Ellipsoid method. Karmarkar, however, persisted with his claims. This inspired further worldwide research. Over a period of 10 years more than 2000 scientific publications appeared on the subject. This led to many new theoretical concepts and algorithmic ideas. These insights were immediately incorporated into computer programs. The work of Lustig, Marsten, and Shanno resulted in a computer code, called OB1, that could compete with the Simplex method. At the time that KORBX entered the market, OB1 was faster, suitable for all current computer platforms, and freely available for research purposes.

The computer code OB1 was based on a method proposed already in the 1950s by Frisch (1956), the *logarithmic barrier method.* By adjusting this method the so-called *path-following methods or interior-point methods* arose. The first term refers to the *central path* of an LO-problem; this is a curve in the interior of the feasible region that converges to an optimal solution. After its rediscovery, independently by several researchers, it became the basis of all modern interior-point methods. These methods are nothing else than numerical recipes that generate a sequence of points, on or close to the central path, that converge to an optimal solution. In this way, several efficient (i.e., polynomial) methods came to life whose practical performance justify the original claim of Karmarkar [Roos *et al.* (1997)].

In the meantime it turned out that the implementations of the Simplex method could be accelerated dramatically by implementing techniques already available in the literature. Especially Bixby (Rice University, Houston, TX) did important work in this respect. As a result an exciting competition arose between him and the makers of OB1. In some large-scale applications in the airline industry this has led to a nice synthesis of both approaches, where a close-to-optimal solution is generated by an interior-point method which is then used as input for the Simplex method to generate an exact solution.

A result of the sketched developments is that nowadays LO-problems can be solved about 1,000,000 times faster than in 1984. A factor of 1000 is due to the better algorithmic methods and the other factor of 1000 to the improvements in computer technology (Moore's law). The consequences for the applications are far-reaching. An LO-model that required 1 year of computational time 16 years ago can now be solved in about 30 seconds. It is clear that problems that require a computational time of 1 year are unsolvable from a practical point of view. Therefore, the improved performance, although of a quantitative nature, has qualitative consequences: problems that could not be solved 15 years ago can now be solved in a few seconds. This explains why the use of commercial LO-packages has shown an explosive growth during the last years.

Modern LO-packages contain both a Simplex-based solver and an interior-point solver. Some of the most well-known packages are OSL of IBM, CPLEX (based on the Simplex code of Bixby and on OB1), XPRESS-MP of Dash Associates, and MOSEK of the brothers E. D. and K. D. Andersen. The environments in which these packages are used nowadays is quite diverse: aviation industry, oil industry, engineering design, finance, water management, etc.

II. THE SIMPLEX METHOD

The Simplex method can be used to solve any LO-problem in the standard form. The method constructs a sequence of basic solutions of the problem until it either finds an optimal basic solution or detects that such a solution does not exist.

A. Reduction to Standard Form

If an LO-problem does not have the standard form it can easily be put into this form as follows.

If a constraint has the form $a_i^T x \geq b_i$ we replace it by $-a_i^T x \leq -b_i$. Thus, we may assume that the inequalities have the form $a_i^T x \leq b_i$. For each such inequality constraint we introduce a *slack variable* s_i according to $a_i^T x + s_i = b_i$. We are then left with a situation where all constraints are equality constraints. Let us write them as $Ax = b$. If b does not belong to the column space of the matrix A, then this system is inconsistent and the problem infeasible. Otherwise, we remove redundant constraints until the matrix A has full row rank. If for each variable a nonnegativity constraint exists, the problem already has the standard form. Otherwise, let F denote the index set of the "free" variables, i.e., the variables without nonnegativity constraints. The classical way to "remove" the free variables is to substitute $x_i = x_i^{(1)} - x_i^{(2)}$, with $x_i^{(1)} \geq 0$ and $x_i^{(2)} \geq 0$, for each free variable x_i. This approach, although theoretically sound, has some serious drawbacks: it increases the number of variables, and when solving the problem with an interior-point method, the new model will not satisfy the interior point condition (IPC). A better approach is as follows. Let r denote the rank of A_F (the submatrix of A formed by the columns indexed by F). As long as free variables occur in the problem formulation we choose a free variable and a constraint in which it occurs. Then, using this (equality) constraint, we express the free variable in terms of the other variables and by substitution we eliminate it from the other constraints and from the objective function. Since F has rank r, we can do this r times, and then the remaining constraints no longer contain free variables. We are left with m equality conditions, r of which express free variables in the remaining variables, while the remaining $m - r$ equalities contain no free variables. Observe that the first r equalities do not impose a condition on the feasibility of the vector x; they simply tell us how the values of r of the free variables in x can be calculated from the remaining variables. Hence, these equalities can be neglected from now on, and we are left with $m - r$ equality constraints in nonnegative variables. If the objective function still contains free variables, the problem is unbounded from below, and we have solved the problem. Otherwise, the remaining problem has the standard form.

B. Basic Solution

Let x denote any solution of $Ax = b$ and let $\sigma(x)$ denote its support:

$$\sigma(x) := \{i : x_i \neq 0\}.$$

We call x a *basic solution* if the columns in $A_{\sigma(x)}$ are linearly independent. If x is not a basic solution we can easily derive a basic solution from it, as follows. The columns in $A_{\sigma(x)}$ being linearly dependent, there exists a nonzero vector λ such that $A\lambda = 0$, with $\lambda_i = 0$ for $i \notin \sigma(x)$. Now let $\alpha \geq 0$ grow until one of coordinates in $\sigma(x)$ of $x \pm \alpha\lambda$ becomes zero and let $\bar{x}$ denote $x \pm \alpha\lambda$ for this value of α. Then $\sigma(\bar{x})$ is a proper subset of $\sigma(x)$. If $\bar{x}$ is not a basic solution the above procedure is repeated. Since $\sigma(x)$ is finite this yields a basic solution after a finite number of steps.

C. Basic Feasible Solution

Let x denote any nonnegative solution of $Ax = b$ and let $\sigma(x)$ be its support. We call x a *basic feasible solution* if x is a basic solution. If x is not a basic solution, the procedure described in Section II.B can be used to produce a nonnegative basic solution $\bar{x} = x + \alpha\lambda$ (if necessary, use $-\lambda$ instead of λ). Even more, if $p = c^T x$, we can find a basic feasible solution x such that $c^T \bar{x} \leq p$, or establish that the problem is unbounded. To this end we modify the procedure of Section II.B by taking λ such that $c^T \lambda \leq 0$. If this works it gives a basic feasible solution with the desired property. If it does not work and $c^T \lambda < 0$, it follows that $\bar{x} = x + \alpha\lambda$ remains nonnegative, however large the value of α. For large α, $c^T \bar{x}$ goes to minus infinity, proving that the problem is unbounded. If $c^T \lambda = 0$, replacing λ by $-\lambda$ provides a basic feasible solution with the same objective value.

Let us mention that, in general, interior-point methods generate a nonbasic optimal solution. Note that the procedure described above enables us to construct a basic optimal solution from any given optimal solution.

D. A Simplex Iteration

Suppose we are given a basic feasible solution $\bar{x}$. There are two main questions:

1. How can we determine whether the given solution is optimal?
2. If the solution is not optimal, how can we find a better solution?

In this section we deal with these two questions.

We assume (without loss of generality) that the problem matrix A has rank m. Since $\bar{x}$ is a basic solution, its support contains at most m indices and the corresponding columns of A are linearly independent. If necessary, we extend the support of $\bar{x}$ to obtain a set B of m indices such that the columns A_k of A with $k \in B$ are linearly independent. Any such set B is called a *basic index set* (or *basis*) for $\bar{x}$. The remaining indices are called the *nonbasic indices;* their set is denoted as N.

By construction, the submatrix A_B of A, consisting of the columns indexed by elements of B, is nonsingular. Hence, the current basic feasible solution $\bar{x}$ is given by

$$\bar{x}_B = A_B^{-1} b, \quad \bar{x}_N = 0, \tag{2}$$

and, defining $\bar{y} \in \mathbb{R}^m$ by

$$A_B^T \bar{y} = c_B, \tag{3}$$

the current objective value satisfies

$$c^T \bar{x} = c_B^T \bar{x}_B + c_N^T \bar{x}_N = c_B^T A_B^{-1} b = b^T \bar{y}. \tag{4}$$

Letting

$$\bar{s} := c - A^T \bar{y}, \tag{5}$$

we claim that $\bar{x}$ is an optimal solution if $\bar{s} \geq 0$, thus answering question (1).

Here is the proof of this claim. We rewrite the constraints as

$$A_B x_B + A_N x_N = b, \quad x_B \geq 0, \quad x_N \geq 0 \tag{6}$$

and multiply from the left by the inverse of A_B, leaving us with the equivalent system

$$x_B = \bar{x}_B - A_B^{-1} A_N x_N, \quad x_B \geq 0, \quad x_N \geq 0. \tag{7}$$

The objective function can now be written as

$$\begin{aligned} c^T x &= c_B^T x_B + c_N^T x_N \\ &= c_B^T \bar{x}_B + \left(c_N^T - c_B^T A_B^{-1} A_N\right) x_N \\ &= c_B^T \bar{x}_B + \left(c_N - A_N^T \bar{y}\right)^T x_N \\ &= c^T \bar{x} + \bar{s}_N^T x_N. \end{aligned} \tag{8}$$

Now suppose that $\bar{s} \geq 0$, and let x' be any feasible solution. Then, by Eq. (8), we have

$$c^T x' = c^T \bar{x} + \bar{s}_N^T x'_N \geq c^T \bar{x},$$

since $\bar{s}_N \geq 0$ and $x'_N \geq 0$. This proves the claim that $\bar{x}$ is optimal if $\bar{s} \geq 0$.

Next, we deal with question (2) and consider the case where $\bar{s}$ is not nonnegative. So $\bar{s}$ has at least one negative entry. Let $\bar{s}_k < 0$. Then $k \in N$, because $\bar{s}_B = 0$. Expression (8) makes clear how $c^T x$ will change if we increase x_k while leaving the other nonbasic variables zero (their current value). We then have

$$c^T x = c^T \bar{x} + \bar{s}_k x_k. \tag{9}$$

This makes clear that the objective value will decrease if x_k increases. Of course, the larger x_k, the larger the decrease of the objective value. However, when increasing x_k we have to take care of the feasibility. In this respect Eq. (7) is useful. Taking for x_N the vector with x_k in the k-position, and zeros elsewhere, we obtain x_B as a function of x_k:

$$x_B = \bar{x}_B - x_k A_B^{-1} A_k.$$

It is convenient to introduce the matrix

$$T^B := A_B^{-1} A. \tag{10}$$

Note that this is an $m \times n$ matrix whose rows are naturally indexed by the elements of B. We can now rewrite the above expression for x_B as follows:

$$x_B = \bar{x}_B - x_k T_k^B. \tag{11}$$

Hence, if $T_k^B \leq 0$ then $x_B \geq 0$ for all nonnegative values of x_k. Letting x_k go to infinity, the objective value goes to minus infinity. We conclude that the problem is unbounded if $T_k^B \leq 0$.

In the other case, when T_k^B has one or more positive entries, there exists a maximal value of x_k for which $x_B \geq 0$. In fact this value of x_k is given by

$$\tilde{x}_k = \min\left\{ \frac{\bar{x}_i}{T_{ik}^B} : T_{ik}^B > 0 \right\}. \tag{12}$$

For this value of x_k at least one (but possibly more) entries of x_B vanish. Let x_ℓ be one of these entries. Then the new solution is a basic solution with index set

$$B' := (B \cup \{k\}) \backslash \{\ell\}. \tag{13}$$

The new objective value follows from Eq. (9). If $\tilde{x}_k > 0$ the new value is smaller than $c_B^T A_B^{-1} b$, the current value.

Thus, we have answered question (2). In doing so, we have just described a typical iteration of the Simplex method. Starting with a basic feasible solution, with index set B for the basic variables, we moved to another basic feasible solution whose basic index set is given by Eq. (13). We call k the *entering index* and ℓ the *leaving index*. The entering index is chosen first: for this one may take *any* index k such that $\bar{s}_k < 0$. Given k we choose the leaving index ℓ according to the above *quotient rule* (Eq. 12). For the leaving index we may take *any* index ℓ yielding the minimal quotient in Eq. (12). The pair (k, ℓ) is called the *pivoting pair* of the Simplex iteration.

E. Simplex Tableaus

When the given LO-problem is small enough, the Simplex iteration can be performed by hand. In such cases it is usual, and useful, to make use of a so-called Simplex tableau. The Simplex tableau belonging to the basis B looks as follows.

$c^T\bar{x}$	$-\bar{s}^T$
$\bar{x}_B$	T^B

(14)

If the pivoting pair is (k, ℓ), then the Simplex tableau belonging to the new basis B' simply arises by pivoting on the element $T^B_{\ell k}$.

To illustrate the use of the Simplex tableau we consider the following LO-problem.

$$\begin{aligned} \min \quad & 3x_1 + 2x_2 \\ \text{s.t.} \quad & x_1 + x_2 \geq -1 \\ & x_1 + 2x_2 \geq 1 \\ & x_1 \geq 0, \quad x_2 \geq 0. \end{aligned}$$

By adding surplus variables x_3 and x_4 the problem assumes the standard form:

$$\begin{aligned} \min \quad & 3x_1 + 2x_2 \\ \text{s.t.} \quad & x_1 + x_2 - x_3 = -1 \\ & x_1 + 2x_2 - x_4 = 1 \\ & x_1 \geq 0, \quad x_2 \geq 0, \quad x_3 \geq 0, \quad x_4 \geq 0. \end{aligned}$$

The vector $\bar{x} = (1, 0, 2, 0)$ is a basic solution. The set B of basic indices is $\{1, 3\}$. We thus have

$$A_B = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \quad A_B^{-1} = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}.$$

Furthermore,

$$\bar{y} = A_B^{-T} c_B = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \end{pmatrix},$$

and, using Eq. (5),

$$\bar{s} = \begin{pmatrix} 3 \\ 2 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ -4 \\ 0 \\ 3 \end{pmatrix}.$$

Using this, the Simplex tableau for B becomes

		x_1	x_2	x_3	x_4
	3	0	4	0	−3
x_1	1	1	**2**	0	−1
x_3	2	0	1	1	−1

The variable names outside the borders serve to indicate to which variables the rows and columns belong. The tableau shows that $\bar{s}_2 = -4 < 0$. In its column we find the pivot (set in boldface) $(k, \ell) = (2, 1)$ by using the quotient rule. After the pivot we get the tableau

		x_1	x_2	x_3	x_4
	1	−2	0	0	−1
x_2	$\frac{1}{2}$	$\frac{1}{2}$	1	0	$-\frac{1}{2}$
x_3	$\frac{3}{2}$	$-\frac{1}{2}$	0	1	$-\frac{1}{2}$

Note that new basic variables are still nonnegative, as it should. But we now have $\bar{s} \geq 0$, which means that the optimal solution has been found. We read it from the last tableau: the nonbasic variables are equal to zero and the basic variables can be read from its first column: $\bar{x} = (0, \frac{1}{2}, \frac{3}{2}, 0)$. Hence, the optimal solution of the original problem is $x_1 = 0$, $x_2 = \frac{1}{2}$, and the optimal value is 1.

F. Finiteness of the Simplex Method

Unfortunately, it may happen that the new solution is not really "better" than the old one. The new objective value may be equal to the old one, and this happens exactly if $\tilde{x}_k = 0$, which means that $\bar{x}_\ell = 0$. In that case we call the iteration *degenerate*. Then no progress is made in terms of the objective value. Even worse, examples exist in which, after a sequence of degenerate iterations, the new basic index set is equal to a basic index set occurring earlier in the sequence [Avis and Chvátal (1978); Beale (1955); Hoffman (1953); Lee (1997)]. This is the so-called phenomenon of *cycling*.

It is most important to avoid cycling. Although it is impossible to avoid degenerate iterations, it is possible to avoid the occurrence of cycles. This can be achieved by using a *cycle-breaking* rule. Such a rule imposes some conditions on the choice of the entering and the leaving variable and its use guarantees that the Simplex method will never produce the same basic index set twice, see Section II.G. What does this mean? Since the number of basic index sets is finite it means that after a finite number of iterations the Simplex method terminates. There are two possible reasons for the method to terminate. The first reason is that there is no candidate for k. In that case we have

$\bar{s} \geq 0$, and as we saw before this implies the optimality of the current solution $\bar{x}$. The second possible reason for termination of the method is that after having found the entering variable k there is no candidate for ℓ; as we established earlier this implies that the problem is unbounded.

G. Cycle-Breaking Rules

One elegant way to avoid cycles is to use a so-called *least index rule*. Then, whenever there is any ambiguity in the choice of k or ℓ we choose the smallest possible index among the candidate indices. This rule was first proposed by Bland (1977).

A second cycle-breaking rule is the *lexico-graphic rule* [Danzig *et al*. (1955)]. This rule uses the lexico-graphic ordering "$\prec$" of vectors: we say that $v \prec w$ ($v, w \in \mathbb{R}^p$) if the vector $u = w - v$ is *lexicographically positive*, i.e. $u \neq 0$ and the first nonzero coordinate of u is positive. The lexicographic rule gives no condition on the choice of the entering index k, but it requires the leaving index ℓ to be taken such that the pivot element $T^B_{\ell k}$ is positive and the vector

$$\frac{\left(\bar{x}_B T^B\right)_{\ell,:}}{T^B_{\ell k}}$$

is lexicographically minimal. Here, $(\bar{x}_B T^B)$ is the matrix obtained by extending T^B to the left with the column $\bar{x}_B$. So this matrix is nothing else than the part of the Simplex tableau below its first row; $(\bar{x}_B T^B)_{\ell:}$ denotes the row of the tableau indexed by the basic variable ℓ.

If initially all rows in the tableau below the first row are lexicographically positive, which can be realized easily, then the lexicographic rule guarantees that in all subsequent tableaus this property is maintained, and that the first row is lexicographically decreasing. The last property prevents the occurrence of cycles.

H. Initialization: Two-Phase Method

In Section II.D we assumed that we were given a basic feasible solution $\bar{x}$. We show in this section how to obtain such a solution, if it exists.

We introduce *artificial variables* t_i ($1 \leq i \leq n$) and consider the problem

$$\begin{aligned} \text{minimize} \quad & e^T t \\ \text{subject to} \quad & Ax + t = b \\ & x \geq 0, \quad t \geq 0, \end{aligned} \tag{15}$$

where e denotes the all-one vector. Without loss of generality we may assume that $b \geq 0$ in Eq. (1), because otherwise we multiply constraints for which the corresponding entry in b is negative, by -1. Then $x = 0, t = b$ is a feasible solution for Eq. (15), and, obviously, this solution is basic. Hence, we can solve Eq. (15) with the Simplex method. Since Eq. (15) is a bounded problem, this yields a basic feasible solution $(\bar{x}, \bar{t})$ of Eq. (15). Now two cases can occur: $e^T \bar{t} > 0$ or $e^T \bar{t} = 0$. In the first case the problem must be infeasible, because any feasible solution of Eq. (1) yields a feasible solution to Eq. (15) with $t = 0$. In the second case $\bar{x}$ is a basic feasible solution of Eq. (1).

It is now clear that we can solve Eq. (1) with the Simplex method in two phases. In phase I we solve Eq. (15); if the problem is infeasible it is detected in this phase, otherwise we obtain a basic feasible solution of Eq. (1) that can be used in phase II to solve the problem. Phase II either yields an optimal basic solution, or detects that the problem is (feasible and) unbounded.

I. Duality

Suppose that Eq. (1) has an optimal solution. Then, by applying the Simplex method, we can obtain an optimal basic index set B and the corresponding basic feasible solution $\bar{x}$. Below, we use again the vectors $\bar{y} \in \mathbb{R}^m$ and $\bar{s} \in \mathbb{R}^n$ as introduced in Eqs. (3) and (5), respectively.

1. Dual of the Standard Problem

Recall from Section II.D that $\bar{s} \geq 0$. Hence, $(\bar{y}, \bar{s})$ is feasible for the following maximization problem:

$$\begin{aligned} \text{maximize} \quad & d = b^T y \\ \text{subject to} \quad & A^T y + s = c \\ & s \geq 0. \end{aligned} \tag{16}$$

2. Duality Results

We claim that $(\bar{y}, \bar{s})$ is an optimal solution of Eq. (16). This is a consequence of the following result.

Theorem 1 (Weak duality) Suppose that x and (y, s) are feasible for Eqs. (1) and (16), respectively. Then

$$c^T x - b^T y \geq 0.$$

Proof: We have

$$\begin{aligned} x^T s &= x^T (c - A^T y) \\ &= c^T x - (Ax)^T y = c^T x - b^T y. \end{aligned}$$

Since $x \geq 0$ and $s \geq 0$, $x^T s \geq 0$. □

Theorem 1 reveals a close relation between problems (1) and (16). From now we call these problems the *primal problem* and *dual problem*, respectively. The theorem implies that if x is feasible for the primal problem (or shortly,

primal feasible) then $c^T x$ is an upper bound for the optimal value d^* of the dual problem, and, vice versa, if (y, s) is *dual feasible* then $b^T y$ is a lower bound for the optimal value p^* of the primal problem. As a consequence, if $c^T x = b^T y$, then x is an optimal solution of the primal problem and (y, s) is an optimal solution of the dual problem. Hence, by taking $x = \bar{x}$, $y = \bar{y}$ and $s = \bar{s}$, the above claim now follows from Eq. (4).

A direct consequence of Theorem 1 is that if either of problems (1) or (16) is unbounded, then the other problem is infeasible.

In summary, if the primal problem has an optimal solution then so has the dual problem and the optimal values coincide. If the primal problem is unbounded then the dual problem is infeasible. In Section II.I.3 we will see that problem (16) also has a dual problem, which is exactly (1). Hence, we may interchange the words "primal" and "dual" in the first sentence of this paragraph. Thus, we have the following result, due to Von Neumann (1947).

Theorem 2 (Strong duality) If the primal and dual problem are both feasible then both problems have optimal solutions and their optimal values are equal. Otherwise, neither of the two problems has optimal solutions.

If neither of the two problems has optimal solutions then both are infeasible, or one is infeasible while the other is unbounded.

A second duality result is due to Goldman and Tucker (1956).

Theorem 3 (Goldman-Tucker) If both the primal and dual problem are feasible then there exists a strictly complementary pair of optimal solutions, i.e., optimal solutions x and (y, s) such that $x + s > 0$.

This duality result is less well-known. Its interest is that interior-point methods produce strictly complementary solutions (cf. Section III.J). It has recently become clear that such solutions play an important role when dealing with sensitivity analysis (cf. Section IV.B).

Let us also mention that the primal problem is infeasible if and only if there exists a vector y such that $A^T y \leq 0$ and $b^T y > 0$, and the dual problem is infeasible if and only if there exists a vector $x \geq 0$ such that $Ax = 0$ and $c^T x < 0$. These statements are examples of *theorems of the alternatives* and are equivalent to the well-known Farkas' lemma. See, e.g., Schrijver (1986).

3. Dual of a General LO-Problem

The dual of the standard problem, as given in Section II.I.3, can be reformulated in the following way:

$$\begin{aligned} &\text{maximize} && b^T y \\ &\text{subject to} && A^T y \leq c. \end{aligned}$$

TABLE I Scheme for Dualizing

$\min c^T x$	$\max b^T y$
'=' Constraint	Free variable
'≥' Constraint	Variable ≥ 0
'≤' Constraint	Variable ≤ 0
Free Variable	'=' Constraint
Variable ≥ 0	'≤' Constraint
Variable ≤ 0	'≥' Constraint

In this way we get a 1–1 correspondence between the variables in the primal problem and the constraints in the dual problem, and vice versa. Note that the primal constraints are equality constraints and the dual variables are free (i.e., without sign constraints). On the other hand, the primal variables are nonnegative and the dual constraints are inequality constraints of '≤' type. Also note that the roles of b and c are exchanged when taking the dual, and the problem matrix A is transposed.

We can associate with each LO-problem (not necessarily in the standard from) a dual problem. To this end we first put the problem in the standard from (cf. Section II.A), and then take the dual of this problem. The dual problem obtained in this way, however, in general does not have such a nice 1–1 correspondence between variables on one side and constraints on the other side. With little extra effort it is possible to reformulate the dual such that we retain a 1–1 correspondence. In this way we get a simple and natural relation between a general LO-problem and its dual, at the same time making it quite easy to write down the dual problem of an arbitrary LO-problem. The relation is shown in Table I.

The scheme can be used both from the left to the right and from the right to the left. Thus, e.g., if the primal problem is a maiximization problem, the dual problem is a minimization problem and, e.g., a '≥'-constraint in the primal problem gives rise to a nonpositive variable in the dual problem. An obvious consequence is that we may say, in short, that the dual of the dual problem is the primal problem.

J. The Dual Simplex Method

Given any basic index set B, following the method of Section II.I we can associate with B a basic solution $\bar{x}$ of the primal problem and a basic solution $(\bar{y}, \bar{s})$ of the dual problem as done in Eqs. (2), (3), and (5). Note that if $\bar{x} \geq 0$ and $\bar{s} \geq 0$ then $\bar{x}$ and $(\bar{y}, \bar{s})$ are primal and dual feasible respectively, and due to Eq. (4), these solutions are optimal. In this section we do not assume that $\bar{x}$ is primal feasible but we assume that $(\bar{y}, \bar{s})$ is dual feasible.

A typical iteration in the dual Simplex method then goes as follows. Since $\bar{x}$ is not primal feasible we may choose

an index $\ell \in B$ such that $\bar{x}_\ell < 0$. Our aim is to replace the index $\ell \in B$ by some index $k \in B$, thus getting the new basic index set [see Eq. (13)]

$$B' := (B \cup \{k\}) \backslash \{\ell\}.$$

The new Simplex tableau will be obtained by using $T^B_{\ell k}$ as pivot. Denoting the new basic solutions as $\tilde{x}$ and $(\tilde{y}, \tilde{s})$, we then have

$$\tilde{x}_\ell = \tilde{x}_\ell - \tilde{x}_k T^B_{\ell k} = 0 \tag{17}$$

and

$$\tilde{s}_i = \bar{s}_i - \frac{T^B_{\ell i}}{T^B_{\ell k}} \bar{s}_k, \quad 1 \le i \le n. \tag{18}$$

We want to have $\tilde{x}_k > 0$. Therefore, since $\bar{x}_\ell < 0$, Eq. (17) makes it clear that we need

$$T^B_{\ell k} < 0. \tag{19}$$

We further want to maintain dual feasibility, i.e., $\tilde{s} \ge 0$. Assuming Eq. (19), it is obvious that $\tilde{s}_i \ge 0$ whenever $T^B_{\ell i} \ge 0$, because $\bar{s}_i \ge 0$ and $\bar{s}_k \ge 0$. Thus, dual feasibility is maintained if and only if k is such that

$$\bar{s}_i - \frac{T^B_{\ell i}}{T^B_{\ell k}} \bar{s}_k \ge 0$$

whenever $T^B_{\ell i} < 0$. As a consequence, we obtain

$$k = \operatorname{argmin}_i \left\{ \frac{\bar{s}_i}{-T^B_{\ell i}} : T^B_{\ell i} < 0 \right\}.$$

This is the *quotient rule* for the dual Simplex method. With the given leaving index ℓ this rule finds an entering index k such that dual feasibility of the new basic solution is guaranteed.

The above method works only if there are suitable candidates for the entering index k. What if there are no such candidates? Then we have

$$\bar{x}_\ell < 0 \text{ and } T^B_{\ell k} \ge 0, \quad \text{for all } k. \tag{20}$$

Now recall from Eq. (7) that

$$x_B = \bar{x}_B - T^B_N x_N.$$

Hence, if Eq. (20) holds then we have for the ℓ-entry of any primal feasible vector x:

$$x_\ell = \bar{x}_\ell - \sum_{k \in N} T^B_{\ell k} x_k \le \bar{x}_\ell < 0,$$

showing that the problem cannot be feasible. We conclude that if no candidate for the entering variable exists then the problem is infeasible. Note that this argument does not use the dual feasibility of the current basic solution.

Let us conclude this section by mentioning that in some natural way the dual Simplex method is completely equivalent with the (primal) Simplex method as treated earlier. The natural correspondence between the two methods follows by noting that when applying the Simplex method to solve problem (1) then this also yields the solution of the dual problem (16). Therefore, since the dual of (16) is exactly (1), when solving (16) with the primal Simplex method we also obtain the solution of (1). In fact, it can be seen that applying the dual Simplex method to (1) is essentially the same as applying the primal Simplex method to (16).

With the above observation in mind it will be no surprise that cycling of the dual Simplex method can be prevented by using the *least index rule*: whenever there is any ambiguity in the choice of ℓ or k, choose the smallest possible index among the candidate indices.

K. The Criss-Cross Method

So far, we have dealt with two variants of the Simplex method: the primal Simplex method and the dual Simplex method. The primal Simplex method generates a sequence of primal feasible basic solutions, whereas the dual Simplex method uses only dual feasible basic solutions. Feasiblility is maintained by choosing the pivoting pair (k, ℓ) in each iteration according to the primal and dual quotient rule, respectively. To start such a method, one first needs to generate a feasible basic solution, thus separating the work into two phases. In Phase I a feasible basic solution is generated; this phase requires the introduction of the artificial variables. Phase II is used to generate an optimal solution.

We discuss in this section how we can avoid artificial variables, and solve the problem in one phase. The underlying idea, due to Zionts (1969), is to neglect the feasibility issue and to aim for feasibility at both sides simultaneously.

Let us call index i *primal infeasible* if $\bar{x}_i < 0$ and otherwise *primal feasible*. Similarly, index i is *dual infeasible* if $\bar{s}_i < 0$ and otherwise *dual feasible*. Recall that any basic solution satisfies $\bar{x}_N = 0$ and $\bar{s}_B = 0$. Hence, for each index either $\bar{x}_i = 0$ or $\bar{s}_i = 0$. Therefore, it is impossible for an index i to be primal infeasible and dual infeasible at the same time. If all indices are feasible then the basic solutions are optimal, and we are done. Otherwise, we take an infeasible index i. If i is dual infeasible then, putting $k := i$ we look for an index ℓ such that $T^B_{\ell k} > 0$; if such an ℓ does not exist the problem is unbounded, or infeasible. If i is primal infeasible then, putting $\ell := i$ we look for an index k such that $T^B_{\ell k} > 0$; if such a k does not exist the problem is infeasible. If a pair (k, ℓ) has been found then we perform an iteration with this pair as pivoting pair. Hoping for the best, this process is repeated.

L. Example of Cycling

Note that this so-called *criss-cross* method differs from the primal and the dual Simplex method in many ways. Since no quotient rule is used there is much more freedom in the choice of the pivoting pair, and primal and dual iterations are used in a more or less arbitrary order. It may be clear that there is no guarantee that the process will stop. In fact, it is quite easy to find examples of cycling with this method. Below is an example of a cycle that starts with the first tableau from the example in Section II.E.

		x_1	x_2	x_3	x_4
	3	0	4	0	−3
x_1	1	1	2	0	−1
x_3	2	0	**1**	1	−1

		x_1	x_2	x_3	x_4
	−5	0	0	−4	1
x_1	−3	1	0	−2	**1**
x_2	2	0	1	1	−1

		x_1	x_2	x_3	x_4
	−2	−1	0	−2	0
x_4	−3	1	0	−2	1
x_2	−1	1	1	**−1**	0

		x_1	x_2	x_3	x_4
	0	−3	−2	0	0
x_4	−1	**−1**	−2	0	1
x_3	1	−1	−1	1	0

		x_1	x_2	x_3	x_4
	3	0	4	0	−3
x_1	1	1	2	0	−1
x_3	2	0	1	1	−1

The fifth tableau is exactly the same as the first tableau.

Cycling of the criss-cross method can be prevented by the least index rule: finiteness of the criss-cross method is then guaranteed [Terlaky (1985)].

M. The Klee-Minty Example

For $m \geq 1$ we define the polytope $\mathcal{P}_m$ in $\mathbb{R}^m$ as follows:

$$\{y: \varepsilon y_{j-1} \leq y_j \leq 1 - \varepsilon y_{j-1},\ 1 \leq j \leq m\},$$

where $y_0 = 0$ and $0 \leq \varepsilon < 0.5$. We consider the LO-problem

$$\max\{y_m: y \in \mathcal{P}_m\}.$$

Note that $y = 0$ is feasible for this problem, and one easily verifies that it is a basic solution.

If $\varepsilon = 0$ then $\mathcal{P}_m$ is the m-dimensional unit cube:

$$\{y \in \mathbb{R}^m: 0 \leq y_j \leq 1,\ 1 \leq j \leq m\},$$

which has 2^m vertices. Hence, for (small) positive values of ε we can consider $\mathcal{P}^m$ as a perturbation of the unit cube. In fact, because of $0 \leq \varepsilon < 0.5$, the polytope $\mathcal{P}_m$ has also 2^m vertices. Moreover, both the primal Simplex method and the criss-cross method, when started at $y = 0$, and with the least-index rule, pass through all these vertices. Hence, the number of iterations will be $2^m - 1$, in both cases [Roos (1990) and Zionts (1969)]. See Fig. 1, which depicts the Klee-Minty path for $m = 3$. Being optimistic, let us assume that one needs 10^{-9} seconds per iteration. Then the computational time becomes $2^m \cdot 10^{-9}$ sec. For $m = 50$ this is about 13 days, and for $m = 60$ even 36.5 years!

N. Computational Issues

Despite the exponential example of Klee and Minty, in practice the Simplex method is a very powerful method. This is due to the fact that the average-case behavior is not exponential but polynomial. On the average, taken over a wide class of representative LO-problems, the number of iterations is bounded above by a polynomial in m and n. This has been shown by Smale (1983) and Borgwardt (1982). As a rule of thumb one may state that on the average the number of iterations grows linearly with the number n of variables and the number m of constraints. Actually, this number highly depends on the choice of the pivots. Quite efficient pivot rules are used nowadays. The search for a theoretically efficient pivot rule goes on. So far, the best thing a pivot rule can have is a finiteness proof. Unfortunately, for all rules with such a proof an exponential example like that of Klee and Minty exists. The history of implementations of the Simplex method is well documented in Orchard-Hays (1990).

III. INTERIOR-POINT METHODS

For a long time the Simplex algorithm was the only practical algorithm for linear optimization. Many people tried to justify the remarkable efficiency of the method by providing a theoretical bound on the number of Simplex iterations. To date, no one has yet given a polynomial bound for general problems.

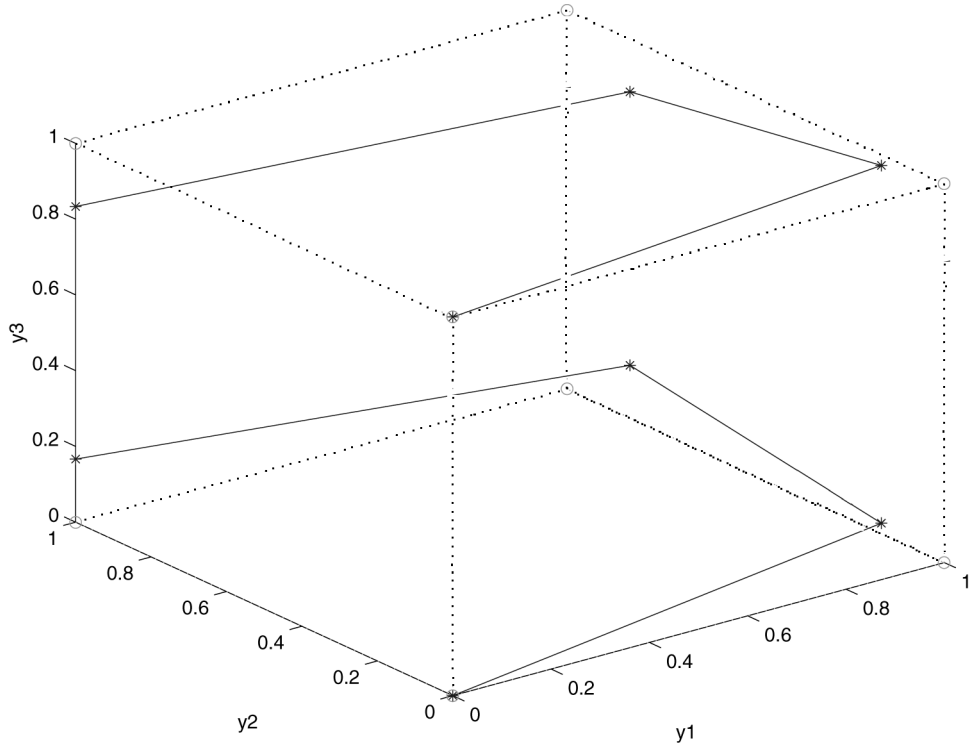


FIGURE 1 Klee-Minty path for $m=3$.

The interest in interior-point methods for linear programming emerged from Karmarkar's contribution in 1984. He proposed a totally new method that enjoyed both polynomial complexity and practical efficiency. The field became one of the most active in the area of mathematical optimization. It introduced new ideas and techniques that now have received their own place among the basic tools in optimization. In this section we survey the theory of interior-point methods, including the analysis of their complexity.

The heart of this theory is the concept of *central path*, a continuous curve in the interior of the feasible set that converges to an optimal point. Most interior-point methods follow the central path, and can be characterized as *path-following methods*. The concept captures the basic ideas that underlie some of the most prominent interior-point methods. It provides a unified framework for a variety of existing efficient methods.

The problem we consider in this section is

$$\begin{aligned} \text{minimize} \quad & p = c^T z \\ \text{subject to} \quad & Az \geq b \\ & z \geq 0. \end{aligned} \tag{21}$$

This is the LO-problem in canonical from. The dual problem is

$$\begin{aligned} \text{minimize} \quad & d = b^T y \\ \text{subject to} \quad & A^T y \leq b \\ & y \geq 0. \end{aligned} \tag{22}$$

A. Reduction to Canonical Form

If a given LO-problem does not have the canonical form, using the method of the Section II.A the problem is first brought into the standard form (1). Choosing an arbitrary basic index set B, using the notation introduced in Section II.D, we rewrite the constraints as in Eq. (7), using also Eq. (10),

$$z_B = \bar{z}_B - T_N^B z_N, \quad z_B \geq 0, \quad z_N \geq 0,$$

and the objective function as

$$c^T z = c^T \bar{z} + \bar{s}_N^T z_N.$$

Omitting the constant term $c^T \bar{z}$ the problem gets the canonical form

$$\begin{aligned} \text{minimize} \quad & \bar{s}_N^T z_N \\ \text{subject to} \quad & -T_N^B z_N \geq -\bar{z}_B \\ & z_N \geq 0. \end{aligned}$$

B. Reduction to Feasibility Problem

By Theorem 2, if one of the two problems (21) and (22) has an optimal solution then so has the other, and then their optimal values coincide. Using also Theorem 1, this is the case if and only if the system

$$\begin{aligned} Az &\geq b, \qquad z \geq 0, \\ -A^T y &\geq -c, \quad y \geq 0, \\ b^T y - c^T z &\geq 0 \end{aligned}$$

has a solution, and any such solution provides optimal solutions for (21) and (22). With $\kappa = 1$, the above system can be rewritten as

$$\begin{pmatrix} 0 & A & -b \\ -A^T & 0 & c \\ b^T & -c^T & 0 \end{pmatrix} \begin{pmatrix} y \\ z \\ \kappa \end{pmatrix} \geq 0, \quad \begin{pmatrix} y \\ z \\ \kappa \end{pmatrix} \geq 0. \tag{23}$$

Here the zeros represent matrices (or vectors) of appropriate sizes. Note that by introducing the variable κ, the system became homogeneous: if (y, z, κ) is a solution then $\lambda(y, z, \kappa)$ is also a solution, for any positive λ. Hence, problem (23) has a solution with $\kappa = 1$ if and only if it has a solution with $\kappa > 0$. Given any solution (y, z, κ) of (23) with $\kappa > 0$, then

$$z^* = \frac{z}{\kappa}, \quad y^* = \frac{y}{\kappa}$$

are optimal solutions of (21) and (22).

Thus we may conclude that (21) and (22) have optimal solutions if and only if (23) has a solution with $\kappa > 0$. The extra variable κ is called the *homogenizing variable*. Note that problem (23) always admits the zero solution $z = 0$, $y = 0$, and $\kappa = 0$. If $\kappa = 0$ for every solution, then we may conclude that (21) and (22) have no optimal solutions, i.e., these problems are infeasible or unbounded.

C. Embedding into Self-Dual Model

To simplify notations, we use the matrix $\bar{M}$ and the vector $\bar{x}$ defined by

$$\bar{M} = \begin{pmatrix} 0 & A & -b \\ -A^T & 0 & c \\ b^T & -c^T & 0 \end{pmatrix}, \quad \bar{x} = \begin{pmatrix} y \\ z \\ \kappa \end{pmatrix}.$$

Then Eq. (23) can be written as

$$\bar{M}\bar{x} \geq 0, \quad \bar{x} \geq 0. \tag{24}$$

We need to find out whether or not this inequality system has a solution with $\kappa > 0$.

When using interior-point methods it is necessary that the IPC is satisfied. In other words, there should exist a vector $x^0 > 0$ such that $\bar{M}x^0 > 0$. The system (23) certainly does not satisfy this condition. Because if y, z, and κ solve (23) then $c^T z - b^T y = 0$, and hence the last coordinate of $M\bar{x}$ vanishes.

To circumvent this difficulty the problem is embedded into a slightly larger problem that satisfies the IPC. This goes as follows. Letting $n - 1$ denote the size of $\bar{M}$, and with e denoting the all-one vector, as usual of appropriate size, we introduce

$$r = e - \bar{M}e, \quad q = \begin{pmatrix} 0 \\ n \end{pmatrix}.$$

Now consider the system

$$\begin{pmatrix} \bar{M} & r \\ -r^T & 0 \end{pmatrix} \begin{pmatrix} \bar{x} \\ \vartheta \end{pmatrix} \geq -q, \quad \begin{pmatrix} \bar{x} \\ \vartheta \end{pmatrix} \geq 0. \tag{25}$$

Note that if $(\bar{x}, \vartheta)$ satisfies (25) and if $\vartheta = 0$, then $\bar{x}$ is a solution of (24). On the other hand, a solution $\bar{x}$ of (24) gives rise to a solution of (25) with $\vartheta = 0$ if and only if

$$r^T \bar{x} \leq n.$$

This certainly holds if $r^T \bar{x} \leq 0$; otherwise a positive multiple of $\bar{x}$ yields a solution of problem (25) with $\vartheta = 0$. We conclude that the set of solutions of (25) with $\vartheta = 0$ contains all solutions of (24), possibly up to a positive factor. The new variable ϑ is called the *lifting variable.*

We now show that the new system satisfies the IPC. In fact the all-one vector does the work. Taking $\bar{x} = e$ and $\vartheta = 1$ we get

$$\bar{M}\bar{x} + \vartheta r = \bar{M}e + r = e$$

and

$$-r^T \bar{x} = -(e - \bar{M}e)^T e = -e^T e = -n + 1,$$

where we used that $e^T \bar{M} e = 0$, since $\bar{M}$ is skew-symmetric. Hence, we obtain

$$\begin{pmatrix} \bar{M} & r \\ -r^T & 0 \end{pmatrix} \begin{pmatrix} e \\ 1 \end{pmatrix} + q = \begin{pmatrix} e \\ 1 \end{pmatrix},$$

proving the claim.

We already observed that a solution of (25) can be useful for us only if $\vartheta = 0$. How do we get such a solution? Note that, since $q \geq 0$, $\bar{x} = 0$ and $\vartheta = 0$ are feasible for problem (25). Therefore, when minimizing ϑ subject to (25) the optimal value will be equal to zero. Defining

$$M = \begin{pmatrix} \bar{M} & r \\ -r^T & 0 \end{pmatrix}, \quad x = \begin{pmatrix} \bar{x} \\ \vartheta \end{pmatrix},$$

$q^T x = n\vartheta$ vanishes if and only if $\vartheta = 0$, and thus we are interested in the optimal solutions of the problem

$$\begin{aligned} \text{minimize} \quad & q^T x \\ \text{subject to} \quad & Mx \geq -q \\ & x \geq 0. \end{aligned} \tag{26}$$

When taking the dual of this problem we get

$$\begin{aligned} \text{maximize} \quad & -q^T u \\ \text{subject to} \quad & -Mu \le q \\ & u \ge 0, \end{aligned}$$

where we used that $M^T = -M$. Since maximizing $-q^T u$ is equivalent to minimizing $q^T u$ this is exactly the same problem as (26). This is expressed by saying that problem (26) is *self-dual.*

We finally point out that the optimal set of (26) is bounded. For any feasible x let $s(x) = Mx + q$. Then

$$\begin{aligned} e^T Mx &= e^T(\bar{M}\bar{x} + \vartheta r) - r^T \bar{x} \\ &= e^T(\bar{M}\bar{x} + \vartheta r) - (e - \bar{M}e)^T \bar{x} \\ &= e^T(\bar{M}\bar{x} + \vartheta r) - e^T \bar{x} + e^T \bar{M}^T \bar{x} \\ &= \vartheta e^T r - e^T \bar{x}, \end{aligned}$$

where we used once more that $\bar{M}^T = -\bar{M}$. For the same reason we have $e^T \bar{M} e = 0$, whence $e^T r = e^T e = n - 1$. Hence,

$$e^T s(x) = \vartheta(n-1) - e^T \bar{x} + n$$

and, finally,

$$\begin{aligned} e^T x + e^T s(x) &= e^T \bar{x} + \vartheta + e^T s(x) \\ &= n(1 + \vartheta). \end{aligned}$$

Taking $\vartheta = 0$ we get

$$e^T x + e^T s(x) = n.$$

Since $x \ge 0$ and $s(x) \ge 0$, this implies that the optimal set is bounded.

D. Central Path

In the previous section we reduced the general LO-problem to the problem of solving the self-dual problem (26). This is not exactly true, however. Remember that the real issue is to find an optimal solution with a positive homogenizing variable $\kappa = x_{n-1}$, or to establish that such a solution does not exist.

Recall that the all-one vector is feasible and

$$s(e) = e, \tag{27}$$

so the IPC is satisfied. Although the M and q introduced in the previous section have a very special structure, in the analysis below we allow M to be any skew-symmetric matrix and q any nonnegative vector.

Denoting the optimal set of (26) as $\mathcal{S}$, and defining the index set B by

$$\mathcal{B} := \{i : x_i > 0 \quad \text{for some } x \in \mathcal{S}\}, \tag{28}$$

the question is whether $n - 1 \in \mathcal{B}$ or not. If the answer is yes, then any solution with positive homogenizing variable will yield optimal solutions for our original problems (21) and (22).

Since M is skew-symmetric we have for any vector u

$$u^T Mu = 0. \tag{29}$$

Hence, if x is feasible,

$$q^T x = (Mx + s(x))^T x = x^T s(x), \tag{30}$$

and x is optimal if and only if $x_i s_i(x) = 0$ for each i.

We will use a shorthand notation for the vector with coordinates $x_i s_i(x)$, namely $xs(x)$. Then the optimality conditions for problem (26) are given by

$$x \ge 0, \quad s(x) = Mx + q \ge 0, \quad xs(x) = 0. \tag{31}$$

The last condition is the *complementarity condition*.

For more appreciation of the next theorem, let us indicate that problem (31) may have multiple solutions. For example, if

$$M = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad q = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

then

$$s(x) = M \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 - x_2 \\ x_1 \end{pmatrix},$$

and the solution to (31) is $x = (0, x_2)$ with $0 \le x_2 \le 1$.

Now consider the following system.

$$x \ge 0, \quad s(x) = Mx + q \ge 0, \quad xs(x) = \mu e, \tag{32}$$

where μ is any positive number.

Theorem 4 For any $\mu > 0$ the system (32) has a unique solution.

The solution of (32) is denoted as $x(\mu)$; it is called the μ-center of (26). When μ runs through all positive reals then $x(\mu)$ follows a curve in the interior of the feasible region of (26). Note that $z(\mu)/\kappa(\mu)$ and $x(\mu)/\kappa(\mu)$ are feasible solutions for the problems (21) and (22); the corresponding curves are the respective central paths of these two problems.

The central path is quite relevant for two reasons. First, observe that for any $\mu > 0$ we have

$$q^T x(\mu) = n\mu. \tag{33}$$

This is an obvious consequence of Eq. (30) and $x_i s_i(x) = \mu$ for each i. Hence, if μ approaches 0 then $q^T x(\mu)$ goes to zero, which means that $x(\mu)$ approaches the optimal set. The second reason is that not only the limit

$$x(0) := \lim_{\mu \downarrow 0} x(\mu)$$

exists but, even more importantly, that the support of $x(0)$ is the set $\mathcal{B}$. As a consequence, if we know $x(0)$, we have enough information to decide whether the homogeneous variable can be positive in the optimal set, and if this is the case we can derive from $x(0)$ optimal solutions for (21) and (22).

Path-following methods use the central path as a guide to the optimal set. Such a method starts at the point on the central path corresponding to $\mu = 1$, because the 1-center is known: due to Eq. (27) we have $x(1) = e$.

E. Conceptual Method

The parameter is μ is called *barrier parameter*. The question we have to deal with is how to obtain the μ-centers for small values of the barrier parameter.

Suppose that we know $x(\mu)$ for some $\mu > 0$, and let μ' be obtained from μ by

$$\mu' := (1 - \theta)\mu,$$

where θ is a positive constant smaller than 1. We may expect that if θ is not too large, the μ'-center will be close to the given μ-center. For the moment, let us assume that we are able to calculate the μ'-center, provided θ is not too large. Then the following conceptual algorithm can be used to find an ε-optimal solution of problem (26).

Conceptual Algorithm

Input:
An accuracy parameter $\varepsilon > 0$;
a barrier update parameter θ, $0 < \theta < 1$.
begin
$x = x(1)$; $\mu := 1$;
while $n\mu \geq \varepsilon$ **do**
begin
$\mu := (1 - \theta)\mu$;
$x := x(\mu)$;
end
end

The output of this algorithm is a feasible solution for problem (26) such that the objective value does not exceed ε. How many iterations are needed by the algorithm? The answer is provided by the following lemma.

Lemma 1 After at most

$$\left\lceil \frac{1}{\theta} \log \frac{n}{\varepsilon} \right\rceil$$

iterations we have $n\mu \leq \varepsilon$.

Proof: Initially, the objective value is n and in each iteration it is reduced by the factor $1 - \theta$. Hence, after k iterations the objective value, given by $q^T x(\mu) = n\mu$, is smaller than, or equal to ε if

$$(1 - \theta)^k n \leq \varepsilon.$$

Taking logarithms, this becomes

$$k \log(1 - \theta) + \log n \leq \log \varepsilon.$$

Since $-\log(1 - \theta) \geq \theta$, this certainly holds if

$$k\theta \geq \log n - \log \varepsilon = \log \frac{n}{\varepsilon}.$$

This implies the lemma. □

The above algorithm uses exact μ-centers. These can be obtained only by solving the nonlinear system (32). To make the algorithm more practical, we have to avoid this. This is the subject of the following sections.

F. Using Approximate Centers

Let us now assume that we have an *approximate* μ-center, i.e., a positive feasible solution x that is close to $x(\mu)$. The meaning of "close" will be made more precise later on.

We want to find a displacements Δx such that

$$x' = x + \Delta x \tag{34}$$

is the μ-center. Denoting $s' = s(x')$, and

$$s' = s + \Delta s, \tag{35}$$

neglecting the inequality constraints for the moment, this means that Δx and Δs should satisfy

$$M(x + \Delta x) + q = s + \Delta s$$

$$(x + \Delta x)(s + \Delta s) = \mu e.$$

This system can be rewritten as

$$M\Delta x = \Delta s, \tag{36}$$

$$s\Delta x + x\Delta s + \Delta x \Delta s = \mu e - xs. \tag{37}$$

The above system is nonlinear and hard to solve. Following Newton's method, we linearize the *centering condition* (37) by neglecting the quadratic term $\Delta x \Delta s$, thus obtaining the *Newton equation*

$$s\Delta x + x\Delta s = \mu e - xs. \tag{38}$$

Substitution of Eq. (36) leads to the linear equation

$$(S + XM)\,\Delta x = \mu e - xs.$$

Here $X = \mathrm{diag}(x)$ and $S = \mathrm{diag}(s)$. The coefficient matrix being nonsingular, this equation uniquely defines the *Newton direction* Δx at x to the *target* $x(\mu)$.

G. Analysis of the Newton Step

Using the notation of Eqs. (34) and (35) we may write

$$\begin{aligned} x's' &= (x + \Delta x)(s + \Delta s) \\ &= xs + (s\Delta x + x\Delta s) + \Delta x \Delta s. \end{aligned}$$

Due to the Newton equation (38) this implies

$$x's' = \mu e + \Delta x \Delta s. \tag{39}$$

Since M is skew-symmetric,

$$(\Delta x)^T \Delta s = (\Delta x)^T M \Delta x = 0,$$

proving that Δx and Δs are orthogonal. Hence, using Eq. (39)

$$q^T x' = e^T(x's') = e^T(\mu e + \Delta x \Delta s) = n\mu.$$

Thus we have shown that after the Newton step, the objective value has the value at the target $x(\mu)$.

But what about x'? We may not expect that x' coincides with $x(\mu)$, due to the "error term" $\Delta x \Delta s$ in Eq. (39). But hopefully, x' is closer to $x(\mu)$ then x. For dealing with this we need a quantity to measure proximity to $x(\mu)$. For this purpose we introduce the quantity $\delta(x, \mu)$ defined by

$$\delta(x, \mu) = \frac{1}{2} \left\| \sqrt{\frac{xs}{\mu}} - \sqrt{\frac{\mu e}{xs}} \right\|.$$

The notation, although not quite common, explains itself: all operations are meant to be coordinatewise. Note that

$$x = x(\mu) \Leftrightarrow \frac{xs}{\mu} = e \Leftrightarrow \frac{\mu e}{xs} = e.$$

Hence, if $x = x(\mu)$ then $\delta(x, \mu) = 0$ and $\delta(x, \mu) > 0$ otherwise. One may prove the following.

Theorem 5 If $\delta := \delta(x, \mu) \leq 1$, then the Newton step is feasible, i.e., x' and s' are nonnegative. Moreover, if $\delta < 1$, then x' and s' are positive and

$$\delta(x', \mu) \leq \frac{\delta^2}{\sqrt{2(1 - \delta^2)}}.$$

This remarkable result shows not only that the Newton step is feasible if x is close enough (i.e., $\delta < 1$) to the μ-center, but also that the Newton process is quadratically convergent. Especially,

$$\delta(x, \mu) \leq \frac{1}{\sqrt{2}} \Rightarrow \delta(x', \mu) \leq \delta(x, \mu)^2.$$

It is now clear that with Newton's method we can obtain good approximations of the μ-center in a very efficient way.

H. Approximate Center Method

We modify the conceptual algorithm in Section III.E slightly, by replacing the statement $x := x(\mu)$ with $x := x + \Delta x$.

Approximate Center Algorithm

Input:
An accuracy parameter $\varepsilon > 0$;
a barrier update parameter θ, $0 < \theta < 1$.
begin
$x = x(1)$; $\mu := 1$;
while $n\mu \geq \varepsilon$ **do**
begin
$\mu := (1 - \theta)\mu$;
$x := x + \Delta s$;
end
end

In the next section we discuss how an appropriate value of the parameter θ can be obtained, so that during the course of the algorithm the iterates are always close enough to the current μ-center to guarantee that Newton's method is quadratically convergent.

I. Complexity Analysis

At the start of the algorithm we have $\mu = 1$ and $x = x(1)$, whence $q^T x = n$ and $\delta(x, \mu) = 0$. In each iteration μ is first reduced with the factor $1 - \theta$ and then the Newton step is made to the new μ-center. It will be clear that the reduction of μ has effect on the value of the proximity measure. This effect is fully described by the following lemma.

Lemma 2 Let $x > 0$ and $\mu > 0$ be such that $s = s(x) > 0$ and $q^T x = n\mu$. Moreover, let $\delta := \delta(x, \mu)$ and $\mu' = (1 - \theta)\mu$. Then

$$\delta(x, \mu')^2 = (1 - \theta)\delta^2 + \frac{\theta^2 n}{4(1 - \theta)}.$$

Now let us have

$$q^T x = n\mu \quad \text{and} \quad \delta(x, \mu) \leq \frac{1}{2} \tag{40}$$

at the start of some iteration. It certainly holds in the first iteration. We claim that these properties are maintained during the course of the algorithm if $\theta = 1/(2\sqrt{n})$.

When the barrier parameter is updated to $\mu' = (1 - \theta)\mu$, the lemma gives

$$\delta(x, \mu')^2 \leq \frac{1 - \theta}{4} + \frac{\theta^2 n}{4(1 - \theta)}.$$

Due to the choice of θ, we get

$$\delta(x,\mu')^2 < \frac{1}{4} + \frac{1}{16(1-\theta)} \leq \frac{1}{4} + \frac{1}{8} < \frac{1}{2}.$$

Therefore, after the Newton step $\delta(x',\mu') \leq 1/2$. Also, $q^T x' = n\mu'$. This proves the claim.

Thus we have shown that the following theorem holds.

Theorem 6 If $\theta = 1/(2\sqrt{n})$ then the algorithm with full Newton steps requires at most

$$\left\lceil 2\sqrt{n}\, \log \frac{n}{\varepsilon} \right\rceil$$

iterations. The output is a feasible $x > 0$ such that $q^T x = n\mu \leq \varepsilon$ and $\delta(x,\mu) \leq \frac{1}{2}$.

This theorem shows that we can get an ε-solution x of our self-dual model with ε as small as desirable. But note that x will always be an interior solution, so $x > 0$ and $s = s(x) > 0$.

A crucial question is whether the variable $\kappa = x_{n-1}$ is positive or zero in the limit, when μ goes to zero. In practice, for small enough ε it is usually no serious problem to decide which of the two cases occurs. But the theory can help us in an elegant way, as we explain in the next section. This requires some further analysis of the central path.

Before proceeding, we want to agree upon the following. This section has made clear that the conceptual algorithm of Section III.E can be turned into a practical algorithm; the iterates are no longer on the central path but move in a narrow neighborhood around the central path.

In the next sections we will assume that the iterates are on the central path, because this simplifies the analysis significantly. At the cost of some additional technicalities in the analysis we can obtain similar results for the iterates of the practical algorithm. For the technical details the reader is referred to Roos *et al.* (1997).

J. Finding the Optimal Partition

If $x \in \mathcal{S}$, $\mathcal{S}$ being the set of optimal solutions, then

$$xs(x) = 0, \quad x + s(x) \geq 0.$$

The equality represents the *complementarity property* of optimal solutions. If the inequality is strict, we say that x is a *strictly complementary solution*. Recall from Section III.D that the support of the limit point $x(0)$ of the central path is the set $\mathcal{B}$, defined by Eq. (28). Let $\mathcal{N}$ denote the complementary set. Then another property of the central path is that the support of $s(0) := s(x(0))$ is $\mathcal{N}$. As a consequence, $x(0)$ is a strictly complementary solution:

$$x(0)s(0) = 0, \quad x(0) + s(0) > 0.$$

In fact, this property can be used to prove Theorem 3 in Section II.I.2. Another consequence is that

$$\mathcal{N} := \{i : s_i > 0 \quad \text{for some } x \in \mathcal{S}\}. \tag{41}$$

The partition of the index set $\{1,\ldots,n\}$ into the classes $\mathcal{B}$ and $\mathcal{N}$ is called the *optimal partition* of (26).

We define the *condition number* of (26) by

$$\sigma := \min_i \max_{x \in \mathcal{S}} (x_i + s_i(x)).$$

The calculation of this (positive!) number is even more cumbersome than solving Eq. (26). For our purpose, however, it is sufficient to know the following lower bound

$$\sigma \geq \frac{1}{\prod_{j=1}^n \|M_j\|}, \tag{42}$$

where M_j denotes the jth column of M. For this bound we need to make the assumption that the entries of M and q are integral. In the sequel, in all the bounds where the condition number occurs, it can be safely replaced by this easily computable lower bound.

Lemma 3 For any positive μ one has

$$x_i(\mu) \geq \frac{\sigma}{n}, \quad i \in \mathcal{B},$$

$$x_i(\mu) \leq \frac{n\mu}{\sigma}, \quad i \in \mathcal{N}.$$

Proof: Let $i \in \mathcal{N}$ and $\tilde{x} \in \mathcal{S}$ be such that $\tilde{s}_i := s_i(\tilde{x})$ is maximal. Then the definition of the condition number σ implies that $\tilde{s}_i \geq \sigma$. Using the skew-symmetry of M one easily sees that

$$(x(\mu) - \tilde{x})^T (s(\mu) - \tilde{s}) = 0.$$

Since $q^T \tilde{x} = \tilde{x}^T \tilde{s} = 0$, it follows that

$$x(\mu)^T \tilde{s} + s(\mu)^T \tilde{x} = n\mu.$$

This implies

$$x_i(\mu)\tilde{s}_i \leq x(\mu)^T \tilde{s} \leq n\mu.$$

Dividing by $\tilde{s}_i$ and using that $\tilde{s}_i \geq \sigma$ we obtain the second inequality of the lemma:

$$x_i(\mu) \leq \frac{n\mu}{\tilde{s}_i} \leq \frac{n\mu}{\sigma}.$$

The first inequality can be derived in a similar way. □

Lemma 3 gives a complete separation between variables in $\mathcal{B}$ and variables in $\mathcal{N}$, provided that μ is so small that

$$\frac{\sigma}{n} > \frac{n\mu}{\sigma}.$$

or, equivalently,

$$n\mu < \frac{\sigma^2}{n}.$$

Thus, applying the algorithm with

$$\varepsilon = \frac{\sigma^2}{n},$$

we obtain a solution x which reveals the optimal partition according to

$$\mathcal{B} = \{i : x_i > s_i(x)\}$$
$$\mathcal{N} = \{i : x_i < s_i(x)\}.$$

By substitution of the above value of ε in Theorem 6 the required number of iterations follows:

$$\left\lceil 2\sqrt{n}\,\log\frac{n^2}{\sigma^2}\right\rceil.$$

After this number of iterations we know if κ belongs to $\mathcal{B}$ or not. In other words, we then know whether the original problem has an optimal solution or not. In the last case we are done, otherwise we are usually interested in finding a solution. If we are interested in an ε-solution we can proceed with the algorithm until such a solution is obtained. An alternative approach may be to use the rounding procedure described in the next section, which yields an exact solution.

K. Rounding to an Exact Solution

Knowing the optimal partition, we aim for finding a strictly complementary solution of Eq. (26).

In one case this is very easy. If it happens that the set $\mathcal{B}$ in the optimal partition is empty then $x(0) = 0$ is a strictly complementary solution and we are done. Note that in that case $s(x(0))$ must be positive. Since $s(x(0)) = q$, this case can easily be seen to occur if and only if $q > 0$.

Therefore, we assume from now on that the class $\mathcal{B}$ is not empty. Assuming this we describe a rounding procedure that can be applied to any x generated by the algorithm to yield a vector $\bar{x}$ such that $\bar{x}$ and its surplus vector $\bar{s} = s(\bar{x})$ are complementary (in the sense that $\bar{x}_N = \bar{s}_B = 0$). In general, however, these $\bar{x}$ and $\bar{s}$ are not necessarily nonnegative. But, as we will see, after sufficient additional iterations of the algorithm the separation between the "small" and the "large" variables is strong enough to get a strictly complementary solution. All of this can be done in polynomial time.

Partitioning the matrix M and the vectors x and s according to the optimal partition, the relation $s = Mx + q$ can be rewritten as

$$\begin{pmatrix} s_{\mathcal{B}} \\ s_{\mathcal{N}} \end{pmatrix} = \begin{pmatrix} M_{\mathcal{BB}} & M_{\mathcal{BN}} \\ M_{\mathcal{NB}} & M_{\mathcal{NN}} \end{pmatrix} \begin{pmatrix} x_{\mathcal{B}} \\ x_{\mathcal{N}} \end{pmatrix} + \begin{pmatrix} q_{\mathcal{B}} \\ q_{\mathcal{N}} \end{pmatrix}.$$

Since $q_{\mathcal{B}}^T x_{\mathcal{B}}(0) = 0$ and $x_B(0) > 0$, it follows that $q_B = 0$. Hence, we have

$$s_{\mathcal{B}} = M_{\mathcal{BB}} x_{\mathcal{B}} + M_{\mathcal{BN}} x_{\mathcal{N}}.$$

Consider the system of equations in the unknown vector ξ given by

$$M_{BB}\xi = s_B - M_{BN} x_N. \tag{43}$$

Note that $\xi = x_{\mathcal{B}}$ is a "large" solution of (43), because the entries of x_B are "large" variables. We can easily see that Eq. (43) has more solutions. This follows by using any optimal solution $\tilde{x}$ with $\tilde{x}_{\mathcal{B}} \neq 0$. One has $\tilde{x}_{\mathcal{N}} = 0$ and $s_{\mathcal{B}}(\tilde{x}) = 0$, whence $M_{\mathcal{BB}}\tilde{x}_{\mathcal{B}} = 0$. Since $x_{\mathcal{B}}(0) \neq 0$, it follows that the matrix $M_{\mathcal{BB}}$ must be singular, and hence Eq. (43) has multiple solutions.

Now let ξ be any solution of Eq. (43) and consider the vector $\bar{x}$ defined by

$$\bar{x}_{\mathcal{B}} = x_B - \xi, \quad \bar{x}_{\mathcal{N}} = 0.$$

Define $\bar{s} = s(\bar{x})$. Since $\bar{x}_{\mathcal{N}} = 0$, we have

$$\bar{s}_{\mathcal{B}} = M_{\mathcal{BB}}\bar{x}_{\mathcal{B}} = M_{\mathcal{BB}}(x_{\mathcal{B}} - \xi) = 0.$$

Therefore, $\bar{x}_N = \bar{s}_B = 0$, showing that the vectors $\bar{x}$ and $\bar{s}$ are complementary. It will be clear, however, that the vectors $\bar{x}$ and $\bar{s}$ are not necessarily nonnegative, let alone strictly complementary. This only holds if

$$\bar{x}_B = x_B - \xi > 0, \tag{44}$$

and

$$\begin{aligned} \bar{s}_N &= M_{NB}\bar{x}_B + M_{NN}\bar{x}_N + q_N \\ &= M_{NB}(x_B - \xi) + q_N \\ &= s_N - M_{NN}x_N - M_{NB}\xi > 0. \end{aligned} \tag{45}$$

Note that if we run the algorithm long enough, $x_{\mathcal{B}}$ and $s_{\mathcal{N}}$ converge to positive vectors, whereas $x_{\mathcal{N}}$ and $s_{\mathcal{B}}$ (and hence also ξ) converge to zero. This makes it plausible that if we take ε small enough in the algorithm, then we will get a solution ξ of (43) that satisfies (44) and (45), hence giving rise to a strictly complementary solution of Eq. (43).

Omitting further details, we conclude this section by stating the main complexity result for interior-point results: when solving Eq. (43) by Gaussian elimination, an exact solution of Eq. (26) can be found after at most $7\sqrt{nL}$ iterations, where L denotes the binary size of the problem.

L. Other Interior-Point Methods

In the preceding sections we have shown that the LO-problem can be solved in polynomial time. If the problem is infeasible or unbounded then this is detected by the method, otherwise it generates an exact solution. When executed, the number of iterations will be about the same as predicted by the theory. In practice this means that the method is much slower than the Simplex method. There are several ways to make the method more efficient and

competitive to the Simplex method. We briefly discuss some of them.

1. Adaptive-Update Methods

The method we considered generates iterations in a narrow neighborhood of the central path, since at the start of each iterations we have $\delta(x,\mu) \le \frac{1}{2}$. In the theoretical analysis this property is maintained by taking small updates of the barrier barrier parameter ($\theta = 1/(2\sqrt{n})$. In practice, Newton's method behaves much better than predicted by Theorem 5. As a consequence, $\delta(x,\mu)$ will usually be much smaller than $\frac{1}{2}$. It means that we can take larger values of θ without loosing the above property. The adaptive-update method seeks to take θ as large as possible such that at the start of the next iteration we still have $\delta(x,\mu) \le \frac{1}{2}$. This strategy accelerates the method significantly and does not deteriorate the theoretical iteration bound.

2. Large-Update Methods

A popular strategy is to use larger values of θ, say $\theta = 0.5$, or even $\theta = 0.99$. In this case, after the barrier update, $\delta(x,\mu)$ will be much larger than 1, and Theorem 5 can be no longer used to measure progress. Even worse, the (full) Newton step will be no longer feasible. The remedy is to take a *damped* Newton step $x' = x + \alpha\Delta x$, with damping factor α; this factor can be chosen such that x' is feasible and at the same time the proximity decreases sufficiently to get a provable polynomial method. In practice this approach yields very efficient methods.

3. Predictor-Corrector Method

This is the most popular method. We describe a simple variant. It is based on a very greedy strategy that uses the Newton step targeting at the zero vector. So, in the definition of the Newton step one takes $\mu = 0$. The resulting direction is called the *affine scaling* direction. To stay feasible, this step is damped again, but with a greedy damping factor. As a result, after such a step we may end up far from the central path. To restore the proximity, a so-called *centering step* is taken. This is a (damped) Newton step with $\mu = q^T x/n$, where x is the current iterate. This method has not only turned out to be extremely efficient, but also has the nice property that asymptotically the objective value converges quadratically to zero.

4. Infeasible-Start Methods

To start an interior-point method one needs an interior solution of the problem at hand. Usually such a solution is not available, and then the method cannot even be started. The embedding technique, as described in Section III.C, elegantly resolves this initialization problem, at the cost of two additional variables, the homogenizing variable κ and the lifting variable ϑ. There exist other solutions to the initialization problem, leading to the so-called *infeasible-start methods*. We briefly discuss such a method for the standard form (1) and its dual (16). The centering condition for this pair of problems is simply $xs = \mu e$.

Infeasible-start methods take any two positive n-vectors x and s and an arbitrary m-vector y. These vectors may be not feasible. Defining the primal and dual residuals by

$$r_p(x) = Ax - b, \quad r_d(y,s) = A^T y + s - c,$$

respectively, the Newton-like search directions are defined by the system

$$\begin{aligned} A\Delta x &= -r_p(x) \\ A^T\Delta y + \Delta s &= -r_d(y,s) \\ x\Delta s + s\Delta x &= \mu e - xs. \end{aligned}$$

The steps are damped to keep the iterates sufficiently bounded away from zero. When starting with $y = 0$ and $x = s = \zeta e$ for some suitable ζ, under a mild condition on ζ it can be shown that within $O(n^2|\log\varepsilon|)$ iterations an ε-solution can be found [Wright (1996)]. The generated solutions are not feasible, in general, but the norms of their residuals are bounded in terms of ε.

IV. RELATED TOPICS

A. Integer Programming

Suppose we have a linear maximization problem with feasible region $\mathcal{P}$:

$$\max\{c^T x : x \in \mathcal{P}\}, \tag{46}$$

and one or more of the variables are required to be integral. In many situations it is natural to impose such a condition on the variable, for example if a variable represents a number of vehicles in the model. If all variables need to be integral the problem is called a *pure integer problem*, otherwise a *mixed integer problem*. In the next sections we restrict the discussion to pure integer problems, the generalization to mixed integer problems is straightforward.

1. Branch-and-Bound Methods

This method is an intelligent way of searching for integral solutions in $\mathcal{P}$ with an objective value that is better than the current lower bound, which is initially put to $-\infty$.

Let $\bar{x}$ be an optimal solution of Eq. (46). Suppose that $\bar{x}_i$ is fractional. Let $\mathcal{P}_1$ be the region obtained by adding the

constraint $x_i \leq \lfloor \bar{x}_i \rfloor$ to $\mathcal{P}$ and let $\mathcal{P}_2$ be the region obtained by adding the constraint $x_i \geq \lfloor \bar{x}_i \rfloor + 1$ to $\mathcal{P}$. Here, $\lfloor \bar{x}_i \rfloor$ denotes the integral part of $\bar{x}_i$. Then it will be clear that the integral solution to Eq. (46) is the best of the integral solutions of the two *subproblems*

$$\max\{c^T x : x \in \mathcal{P}_i\}, \quad i = 1, 2.$$

These subproblems can be solved in a similar way: solve the LO-relaxation and if necessary *branch* to two new subproblems.

In this way a binary search tree of subproblems is generated. In principle all subproblems in the generated tree need to be solved. When a subproblem happens to have an integral solution its objective value is a lower bound for the objective value of the original problem, and such a subproblem needs no further branching; if its objective value is higher than the current lower bound we refresh the lower bound. Also, if a subproblem is infeasible no further branching is needed. All subproblems for which the optimal value of the LO-relaxation does not exceed the current lower bound can be considered as *fathomed*. If all subproblems in the tree are solved or fathomed, then there are two possibilities: the lower bound is $-\infty$ or it is a finite value. In the first case no integer solution has been found and hence the original problem has no integer solution. In the second case the subproblem that gave rise to the current lower bound provides the best integer solution.

2. Cutting-Plane Methods

If the solution $\bar{x}$ of the problem (46) is not integral, we call an inequality $a^T x \leq \beta$ a *cutting plane* of (46) if $a^T \bar{x} > \beta$ whereas all integer vectors in $\mathcal{P}$ satisfy $a^T x \leq \beta$. A cutting-plane method adds one or more cutting planes to the problem and then solves the resulting LO-problem. The process is repeated until an integer solution is found or a problem arises that is infeasible.

When applying the Simplex method, there is a nice systematic way to generate cutting planes from optimal tableaus, originally proposed by Gomory. These cutting planes are added to the tableau, and then the tableau is re-optimized. If necessary, the new optimal tableau is used to generate new Gomory-cuts, etc. The process converges to an integer solution, if it exists. Unfortunately, for interior-point methods the situation is not as good. No satisfactory systematic methods exist at present that produce an exact integral solution, although some promising work has been done by Mitchell and Borchers for some specific problems. Part of the problem is due to the fact that at present interior-point methods do not allow fast re-optimizing.

3. Branch-and-Cut Methods

Successful solution of integer LO-problems requires a clever combination of many different ideas. Branch-and-cut methods combine branch-and-bound and cutting-plane methods. The cutting-planes are generated throughout the branch-and-bound tree. The underlying idea is to work on getting as tight as possible bounds in each node of the tree and thus reducing the number of nodes in the search tree. Of course, there is an obvious trade-off. If many cuts are added at a node, re-optimization may slow down. In addition, keeping all the information in the tree is more difficult. On the other hand, the size of the search tree may be reduced significantly. For more details the reader is referred to Wolsey (1998).

B. Sensitivity Analysis

In practice it is often quite important to know how sensitive the optimal value is to perturbations of the data in the problem. We restrict ourselves to perturbations in the objective vector c. For perturbations in the right-hand side vector b one may consider the dual problem, which has b as objective vector, and similar results will follow.

Assuming that the given LO-problem is a minimization problem, let us consider the case where one of the coefficients of c, c_j say, is varied. It can easily be seen that the objective value is a convex and piecewise linear function of c_j. The derivative of the optimal-value function at c_j is called its *shadow price* and the linearity interval to which c_j belongs, its *range*. If c_j is a break point of the optimal value function, we need to distinguish between a left- and a right shadow price, and then the range of c_j is just a singleton.

A crucial result in this respect is that the break points are precisely the points where the optimal partition of the problem changes. If the current optimal partition is known, the range and the shadow price(s) of c_j can be obtained as follows. Assuming that the LO-problem is in standard form, the range of c_j is obtained by minimizing and maximizing c_j over the set

$$\{c_j : A^T y + s = c,\ s_{\mathcal{B}} = 0,\ s_{\mathcal{N}} \geq 0\},$$

and the shadow price(s) follow(s) by minimizing and maximizing x_j over the set

$$\{x_j : Ax = b,\ x_{\mathcal{B}} \geq 0,\ x_{\mathcal{N}} = 0\}.$$

The classical approach, as treated in all textbooks, instead of using the optimal partition in the above formulas, uses the partition formed by the classes B of basic and N of nonbasic variables for some optimal basic solution; this usually gives rise to ambiguous information [Jansen *et al.* (1997) and Roos *et al.* (1997)].

V. FURTHER EXTENSIONS

It should be noted that many phenomena in this world cannot be described adequately by a linear model, but require a nonlinear model. The solution of such models goes beyond the scope of this chapter. Some remarks are in order, however.

Inspired by the success of the interior-point approach to LO, Nesterov and Nemirovski recognized that the underlying ideas can be extended to a wide class of nonlinear optimization problems, the so-called *convex cone optimization problems*. Such problems have the form

$$\min\{c^T x : Ax = b,\ x \in \mathcal{K}\}, \tag{47}$$

where $\mathcal{K}$ denotes a *convex cone*. If $\mathcal{K} = \mathbb{R}^n_t$, the nonnegative orthant, this is exactly the standard LO-problem. The above authors showed that the interior-point approach also applies to problems of this type for different cones. Note that Eq. (47) looks like a linear problem, but one should realize that the cone $\mathcal{K}$ can hide a lot of nonlinearity. One striking example occurs when $\mathcal{K}$ is the cone of positive semidefinite matrices. For example, the nonlinear constraint $uv \geq 1$, with $u \geq 0$ and $v > 0$, can be modeled as

$$\begin{pmatrix} u & 1 \\ 1 & v \end{pmatrix} \text{ is positive semidefinite.}$$

As a consequence, the field of *semidefinite optimization* at present receives much attention. It has important applications in system theory [Boyd *et al*. (1994)] and in combinatorial optimization. Many combinatorial optimization problems admit a natural relaxation to a semidefinite optimization problem. Using this, Goemans and Williamson could generate approximate solutions of a famous and hard combinatorial problem (finding a maximal cut in a graph), not more than 13% from optimal.

It is generally believed that Karmarkar revealed only the tip of an iceberg; there is still a big mass to be explored.

SEE ALSO THE FOLLOWING ARTICLES

COMPUTER ALGORITHMS • DYNAMIC PROGRAMMING • GAME THEORY • LINEAR SYSTEMS OF EQUATIONS • NONLINEAR PROGRAMMING • OPERATIONS RESEARCH

BIBLIOGRAPHY

Avis, D., and Chvátal, V. (1978). Notes on Bland's pivoting rule. *Math. Program. Study* **8,** 24–34.

Beale, E. M. L. (1955). Cycling in the dual simplex algorithm. *Nav. Res. Logist. Q.* **2,** 269–276.

Bland, R. (1977). New finite pivoting rules for the simplex method. *Math. Oper. Res.* **2,** 103–107.

Borgwardt, K.-H. (1982). The average number of pivot steps required by the simplex method is polynomial. *Z. Oper. Res.* **26,** 157–177.

Boyd, S. E., El Ghaoui, L., Feron, E., and Balakrishnan, V. (1994). Linear matrix inequalities in system and control theory. SIAM Studies in Applied Mathematics, Vol. 15. SIAM, Philadelphia.

Dantzig, G. B. (1963). "Linear Programming and Extensions," Princeton Univ. Press, Princeton, NJ.

Dantzig, G. B., Orden A., and Wolfe, Ph. (1955). Notes on linear programming: Part I—the generalized simplex method for minimizing a linear form under linear inequality restrictions. *Pac. J. Math.* **5**(2), 183–195.

Frisch, K. R. (1956). La resolution des problemes de programme lineaire par la methode du potential logarithmique. *Cah. Semin. D'Econ.* **4,** 7–20.

Goemans, M. X., and Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. Assoc. Comput. Mach.* **42**(6), 1115–1145.

Goldman, A. J., and Tucker, A. W. (1956). Theory of linear programming. *In* "Linear Inequalities and Related Systems" (H.W. Kuhn and A.W. Tucker, eds.) Annals of Mathematical Studies, No. 38, pp. 53–97, Princeton Univ. Press, Princeton, NJ.

Hoffman, A. J. (1953). Cycling in the simplex algorithm. Technical Report 2974, National Bureau of Standards.

Jansen, B., de Jong, J. J., Roos, C., and Terlaky, T. (1997). Sensitivity analysis in linear programming: just be careful! *Eur. J. Oper. Res.* **101,** 15–28.

Karmarkar, N. K. (1984). A new polynomial-time algorithm for linear programming. *Combinatorica* **4,** 373–395.

Khachiyan, L. G. (1979). A polynomial algorithm in linear programming. *Dok. Akad. Nauk SSSR* **244,** 1093–1096. Translated into English in *Sov. Math. Dok.* **20,** 191–194.

Klee, V., and Minty, G. J. (1972). How good is the simplex algorithm? *In* "Inequalities III" (O. Shisha, ed.), Academic Press, New York.

Lee, J. (1997). Hoffman's circle untangled. *SIAM Rev*. **39,** 98–105.

Mitchell, J. E., and Borchers, B. (1996). Solving real-world linear ordering problems using a primal-dual interior point cutting plane method. *Ann. Oper. Res.* **62,** 253–276.

Nesterov, Y., and Nemirovskii, A. S. (1994). Interior point polynomial algorithms in convex programming. SIAM Studies in Applied Mathematics, Vol. 13. SIAM, Philadelphia.

von Neumann, J. (1947). On a maximization problem. Manuscript, Institute for Advanced Studies, Princeton University, Princeton, NJ.

Orchard-Hays, W. (1990). History of the development of LP solvers. *Interfaces* **20**(4), 61–73.

Roos, C. (1990). An exponential example for Terlaky's pivoting rule for the criss-cross method. *Math. Program.* **46,** 79–84.

Roos, C., Terlaky, T., and Vial, J.-Ph. (1997). "Theory and Algorithms for Linear Optimization. An Interior Approach," John Wiley & Sons, Chichester, UK.

Schrijver, A. (1986). "Theory of Linear and Integer Programming," John Wiley & Sons, New York.

Smale, S. (1983). On the average number of steps of the simplex method of linear programming. *Math. Program.* **27,** 241–262.

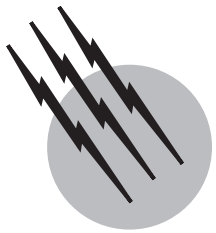
Terlaky, T. (1985). A convergent criss–cross method. *Math. Oper. Stat. ser. Optimization* **16,** 683–690.

Williams, H. P. (1990). "Model Building in Mathematical Programming," 3rd edition, John Wiley & Sons, New York. USA.

Wolsey, L. A. (1998). "Integer Programming," John Wiley & Sons, New York.

Wright, S. J. (1996). Primal-Dual Interior-Point Methods. SIAM, Philadelphia.

Zionts, S. (1969). The criss-cross method for solving linear programming problems. *Manag. Sci.* **15,** 426–445.



Loop Groups

Andrew Pressley
King's College, London

I. Basic Properties of Loop Groups
II. The Fundamental Homogeneous Space
III. Representation Theory of Loop Groups
IV. Relations with Other Parts of Mathematics

GLOSSARY

Central extension A central extension of a group G by an Abelian group A is a group $\tilde{G}$ that has A as a central subgroup such that the quotient group $\tilde{G}/A$ is isomorphic to G.
Deformation retract A space X is a deformation retract of a space Y if the identity map $Y \to Y$ can be deformed through continuous maps to a map $Y \to X$.
Lie group A group that is also a smooth manifold such that the group operations are smooth.
Maximal torus A maximal connected Abelian subgroup of a compact Lie group.
Symplectic manifold A smooth manifold equipped with a closed 2-form that is nondegenerate at each point.

LOOP GROUPS are groups of maps from the circle into a finite-dimensional Lie group, usually assumed to be compact. They arise in many areas of mathematics and physics, such as the theory of integrable systems, singularity theory, and two-dimensional quantum field theory. The geometry and representation theory of loop groups is closely analogous to that of compact Lie groups. The central extensions of the Lie algebras of loop groups are the simplest infinite-dimensional examples of Kac–Moody algebras. (Throughout this article, G will denote a compact Lie group.)

I. BASIC PROPERTIES OF LOOP GROUPS

A. Definitions

A loop group LG is the group of maps from the circle S^1 into a Lie group G. Except where stated otherwise, the maps will be assumed to be smooth, that is, infinitely differentiable, and the group G to be compact. The group operation in LG is given by pointwise multiplication. When provided with the C^∞-topology, LG can be given the structure of an infinite-dimensional Lie group, modeled on the space $L\mathfrak{g}$ of smooth maps from S^1 into the Lie algebra $\mathfrak{g}$ of G. The model space $L\mathfrak{g}$ is a Lie algebra, again under pointwise operations, and is the Lie algebra of LG. It is called a loop algebra. There is an exponential map $L\mathfrak{g} \to LG$ induced from that of G; unlike the exponential map of G itself, that of LG is not surjective, although its image is dense in LG.

One of the many properties of loop groups that does not hold for infinite-dimensional Lie groups in general is the existence of a complexification. This is simply the group $LG_{\mathbb{C}}$ of smooth loops in the complexification $G_{\mathbb{C}}$ of G itself. Its Lie algebra is the complexification of $L\mathfrak{g}$.

B. Twisted Loop Groups

Geometrically, it LG can be thought of as the space of smooth sections of the trivial principal G-bundle on S^1. More generally, if α is any automorphism of G, one can form a G-bundle on S^1 by taking the quotient of $G \times \mathbb{R}$ by the equivalence relation that identifies (x, t) with $(\alpha(x), t + 2\pi)$ for all $x \in G$. The cross sections of this bundle form a group $L_\alpha G$ called a twisted loop group. It is clear that $L_\alpha G$ depends only on the class of α modulo inner automorphisms of G, and hence may be assumed to be of finite order if G is semi-simple. Since all G-bundles on S^1 are trivial if G is connected, any twisted loop group is isomorphic to an untwisted loop group as an abstract group. However, there are good reasons for treating them separately.

Twisted loop groups arise naturally in the study of maximal Abelian subgroups of LG. If T is a maximal torus in G, then LT is obviously a maximal Abelian subgroup of LG. More generally, if τ is any smooth loop in the space of maximal tori of G there is a maximal Abelian subgroup consisting of the loops f such that $f(\theta) \in \tau(\theta)$ for all θ. Up to isomorphism, it depends only on the free homotopy class of τ. The space of maximal tori is $G/N(T)$, where $N(T)$ is the normalizer of T in G. Its fundamental group is the Weyl group $W = N(T)/T$ of G. Hence, there is a maximal Abelian subgroup associated to every conjugacy class in W. The group associated to $w \in W$ is the twisted loop group of T associated to the automorphism of T induced by conjugation with a representative of w in $N(T)$.

C. Automorphisms

One of the most important properties of loop groups is the existence of the one-parameter group of automorphisms R_ϕ of LG that rotates a loop f rigidly through an angle ϕ:

$$R_\phi(f)(\theta) = f(\theta + \phi)$$

More generally, the group $\mathrm{Diff}(S^1)$ of all smooth differomorphisms of the circle acts as a group of automorphisms of LG. There are also the loops in the group of automorphisms of G itself.

Theorem 1. If G is simple, every automorphism of LG, even as an abstract group, is a product of these two types.

D. Polynomial Loops

Apart from the smooth loops, it is often convenient to consider other classes of loops. One of the most important is the Laurent polynomial loops; these have finite expansions

$$f(z) = \sum_{k=N}^{N} A_k z^k$$

The polynomial loops form a subgroup $L_{\mathrm{pol}}G$ of LG, but $L_{\mathrm{pol}}G$ is not a Lie group in any reasonable sense. However, it is a good approximation to LG in at least two ways:

Theorem 2. If G is any compact group, $L_{\mathrm{pol}}G$ is a deformation retract of LG. If G is also semi-simple, $L_{\mathrm{pol}}G$ is dense in LG in the C'-topology.

E. Central Extensions

Throughout this section G will be assumed to be simply-connected. A central extension of the loop algebra $L\mathfrak{g}$ by the one-dimensional Lie algebra $\mathbb{R}$ is a Lie algebra $\tilde{L}\mathfrak{g}$ that fits into a short exact sequence of the form

$$0 \to \mathbb{R} \to \tilde{L}\mathfrak{g} \to L\mathfrak{g} \to 0$$

with $\mathbb{R}$ lying in the center of $\tilde{L}\mathfrak{g}$. Every such extension is isomorphic to one corresponding to the cocycle

$$\omega(\xi,\ \eta) = \frac{1}{2\pi} \int_{S^1} \langle \xi, d\eta \rangle,$$

where $\langle\ ,\ \rangle$ is an invariant inner product on $\mathfrak{g}$ and $\xi,\ \eta \in \mathfrak{g}$. This extension is the vector space $L\mathfrak{g} \oplus \mathbb{C}$ with the bracket

$$[(\xi,\ \lambda), (\eta, \mu)] = ([\xi, \mu], \omega(\xi, \mu))$$

Note that this cocycle is obviously invariant under the action of the group $\mathrm{Diff}^+(S^1)$ of orientation-preserving diffeomorphisms of S^1. The cocycle ω extends to a left invariant closed 2-form on LG and thus defines a cohomology class in $H^2(LG, \mathbb{R})$.

A central extension of LG by the circle group S^1 is defined in an analogous way. Such an extension gives rise by differentiation to a central extension of $L\mathfrak{g}$. The converse question is answered by the following result.

Theorem 3.

(a) There is a central extension of LG by S^1 with Lie algebra cocycle ω if and only if $\omega/2\pi$ is an integral cohomology class. In this case, the central extension $\tilde{L}\mathfrak{g}$ is unique up to isomorphism, and $\mathit{Diff}^+(S^1)$ acts on it as a group of automorphisms.
(b) If no nonzero multiple of ω is integral, $\tilde{L}\mathfrak{g}$ is not the Lie algebra of any Lie group.
(c) The class of $\omega/2\pi$ is integral if and only if $\langle \check{\alpha},\ \check{\alpha} \rangle \in \mathbb{Z}$ for all coroots $\check{\alpha}$ of $\mathfrak{g}$.

(d) If G is simple, the extension of $L\mathfrak{g}$ defined by any nonzero cocycle ω of the above form is universal, in the sense that every central extension of $L\mathfrak{g}$ by any Abelian Lie algebra A arises from it via a homomorphism $A \to \mathbb{R}$.

For the definition of coroots, see the next subsection.

F. Affine Lie Algebras

The complexified central extensions $\tilde{L}\mathfrak{g}_{\mathbb{C}}$ become, after taking the semi-direct product with the one-dimensional algebra generated by the derivation $d = zd/dz$, examples of Kac–Moody algebras, and in this context they are referred to as affine Lie algebras. One can describe analogs of all the usual notions associated to finite-dimensional semi-simple Lie algebras.

Recall the decomposition

$$\mathfrak{g}_{\mathbb{C}} = \mathfrak{t}_{\mathbb{C}} \oplus \bigoplus_{\alpha} \mathfrak{g}_{\alpha}$$

where $\mathfrak{g}_{\alpha}$ is the subspace of $\mathfrak{g}$ on which the maximal torus T acts (by conjugation) via the homomorphism $\alpha : T \to S^1$. The homomorphisms that occur are called the roots of $\mathfrak{g}$. Roots are often identified with their derivatives at the identity, which are elements of the dual space $\mathfrak{t}^*$ of $\mathfrak{t}$. The element $\check{\alpha} \in [\mathfrak{g}_{\alpha}, \mathfrak{g}_{-\alpha}]$ such that $\alpha(\check{\alpha}) = 2$ is called the coroot associated to the root α.

To formulate the corresponding definitions for loop groups, consider the semi-direct product $LG \tilde{\times} S^1$, where S^1 acts on LG by rotating loops. Then $T \times S^1$ is a maximal Abelian subgroup of $LG \tilde{\times} S^1$ (T is identified with the constant loop whose image lies in T). The Weyl group of LG is thus $W_{\text{aff}} = N(T \times S^1)/(T \times S^1)$. It is called the affine Weyl group, and is isomorphic to the semi-direct product $\check{T} \tilde{\times} \mathbb{Z}$, where $\check{T}$ is the lattice of homomorphisms $S^1 \to T$.

One has the following decomposition of the complexified Lie algebra of $LG \tilde{\times} S^1$:

$$L\mathfrak{g}_{\mathbb{C}} \oplus \mathbb{C} = (\mathfrak{t}_{\mathbb{C}} \oplus \mathbb{C}) \oplus \bigoplus_{k \neq 0} \mathfrak{t}_{\mathbb{C}} z^k \oplus \bigoplus_{(\alpha,k)} \mathfrak{g}_{\alpha} z^k$$

The pairs (α, k) that occur, including those corresponding to the second summand where $\alpha = 0$, are called the roots of LG. The pair (α, k) is regarded as the homomorphism $T \times S^1 \to S^1$ given by $(t, z) \to \alpha(t) z^k$. As in the finite-dimensional case, a subset of the roots is called a simple system if every root can be written as a linear combination of the simple roots, the coefficients being integers all of which have the same sign. A root is called positive if all the coefficients are positive.

The affine Weyl group acts on the Lie algebra of $T \tilde{\times} S^1$ and is generated by the reflections in the hyperplanes $\gamma(x) = 0$, where γ runs through the simple roots of LG. The length $l(w)$ of an element $w \in W_{\text{aff}}$ is the number of reflections in a minimal expression for w.

Choose elements e_{γ} in the root space $\mathfrak{g}_{\alpha} z^k$ corresponding to each root $\gamma = (\alpha, k)$ with the property

$$[e_{\gamma}, e_{-\gamma}] = \check{\gamma}$$

for all simple roots γ of LG. Let $\gamma_1, \ldots, \gamma_l$ be the simple roots of LG and write $e_i = e_{\gamma_i}$, $f_i = e_{-\gamma_i}$. Then, the elements of e_i, f_i, $\check{\gamma}_i$, $i = 1, \ldots, l$ satisfy the following commutation relations:

$$\begin{aligned}
[\gamma_i,\ \gamma_j] &= 0 \\
[e_i,\ f_j] &= \delta_{ij} \check{\gamma}_i \\
[\check{\gamma}_i,\ e_j] &= a_{ij} e_j \\
[\check{\gamma}_i,\ f_j] &= -a_{ij} f_j \\
(\mathrm{ad} e_i)^{-a_{ij}+1} e_j &= 0 \quad \text{if} \quad i \neq j \\
(\mathrm{ad}\, f_i)^{-a_{ij}+1} f_j &= 0 \quad \text{if} \quad i \neq j
\end{aligned}$$

Here, a_{ij} is a square matrix of integers, called the Cartan matrix of LG, satisfying

$$\begin{aligned}
a_{ii} &= 2 \\
a_{ij} &\leq 0 \quad \text{if} \quad i \neq j \\
a_{ij} &= 0 \quad \text{if and only if} \quad a_{ij} = 0
\end{aligned}$$

Theorem 4. These relations define the universal central extension of $L\mathfrak{g}$ defined in the previous section.

If (a_{ij}) is any square matrix of integers satisfying the above conditions, then the preceding relations define a so-called Kac–Moody Lie algebra. If (a_{ij}) is positive definite, the Lie algebra is finite-dimensional and semi-simple. In the case of affine Lie algebras, $\det(a_{ij}) = 0$, but every principal minor of (a_{ij}) is positive definite. In fact, this condition characterizes the affine Lie algebras, and their twisted analogs, among Kac–Moody algebras.

The information contained in the Cartan matrix is often expressed graphically in the form of the Dynkin diagram of the Lie algebra in question. This is a graph that has one node for each generator e_i, the ith and jth nodes being joined by $a_{ij}a_{ji}$ bonds. If $|a_{ij}| > |a_{ji}|$, the bonds carry an arrow pointing toward the ith node.

II. THE FUNDAMENTAL HOMOGENEOUS SPACE

A. Differential-Geometric Properties

The homogeneous space $X = LG/G$ plays a fundamental role in the study of loop groups. It is obviously diffeomorphic to the subgroup ΩG of LG consisting of the based

loops, that is, those satisfying $f(1) = 1$; however, it is better to regard it as a homogeneous space of LG, partly because its properties are closely analogous to those of the homogeneous space G/T of G, where T is a maximal torus of G; of course, G/T is not a group.

The first example of this phenomenon is that X, like G/T, is a complex manifold. This follows from the following factorization theorem, in which $L^+G_{\mathbb{C}}$ denotes the subgroup of $LG_{\mathbb{C}}$ consisting of the smooth maps $f : S^1 \to G_{\mathbb{C}}$ that extend smoothly to holomorphic maps of the disc $\{z \in \mathbb{C} : |z| < 1\}$ in to $G_{\mathbb{C}}$.

Theorem 5. $LG_{\mathbb{C}} = LG, L^+G_{\mathbb{C}}$.

This result shows that X is a homogeneous space of the complex Lie group $LG_{\mathbb{C}}$:

$$X \cong LG_{\mathbb{C}}/L^+G_{\mathbb{C}}$$

One of the most striking facts about X is that it behaves like a compact complex manifold.

Theorem 6.

(a) Every holomorphic map $X \to \mathbb{C}$ is constant on each connected component of X.
(b) If M is any compact complex manifold, the space of based holomorphic maps $M \to X$ lying in a fixed homotopy class is finite-dimensional.

The second important aspect of the geometry of X is the existence of an invariant symplectic structure. The tangent space to ΩG at the identity element is $L\mathfrak{g}/\mathfrak{g}$, where $\mathfrak{g}$ is the Lie algebra of G. The formula

$$\omega(\xi, \eta) = \frac{1}{2\pi} \int_0^{2\pi} \langle \xi(\theta), \eta'(\theta) \rangle \, d\theta$$

where $\langle , \rangle$ is an invariant inner product on $\mathfrak{g}$, defines a skew form on $L\mathfrak{g}/\mathfrak{g}$; extending by left translation gives a nondegenerate closed 2-form on X (the nondegeneracy is in the weak sense, that $\omega(\xi, \eta) = 0$ for all η implies $\xi = 0$).

Moreover, the complex structure and the symplectic structure are compatible, in the sense that they fit together to give a Kähler structure on X; this means that $\omega(J\xi, J\eta) = \omega(\xi, \eta)$ and that $\omega(\xi, J\eta)$ is a Riemannian metric on X, where $J : L\mathfrak{g}/\mathfrak{g} \to L\mathfrak{g}/\mathfrak{g}$ is the infinitesimal complex structure. The Ricmannian metric in question is the Sobolev $\frac{1}{2}$-metric, given by the L^2-norm of the $\frac{1}{2}$-th derivative.

B. Stratifications

There is a canonical real-valued function on X, the energy function, given by

$$E(f) = \frac{1}{4\pi} \int_0^{2\pi} \langle f^1 f'(\theta), f^1 f'(\theta) \rangle \, d\theta$$

The Hamiltonian vector field associated to E by the symplectic structure is the derivative of the circle action on X that rigidly rotates loops. The critical points of E are thus the homomorphisms $S^1 \to G$; they fall into conjugacy classes under the action of G and each conjugacy class is a connected, compact complex manifold; the conjugacy classes are in one-to-one correspondence with the orbits of the action of the Weyl group W of G on the lattice of homomorphisms $S^1 \to T$.

The Kähler metric on X defined above allows one to define the gradient vector field of E. Let $\{f_t\}$ be the downward gradient flow of E passing through a loop f at time $t = 0$; in other words, the solution of the ordinary differential equation

$$\frac{\partial f}{\partial t} = -\text{grad } E$$

Since X is infinite-dimensional, it is not clear *a priori* that this exists.

Theorem 7.

(a) The integral curve f_t of the downward gradient flow exists for all $t > 0$ for any initial loop f. The integral curve exists for all $t < 0$ if and only if the initial loop f is polynomial.
(b) For all integral curves f_t, $\lim_{t \to \infty} f_t$ exists and is a critical point of E. If f_0 is a polynomial loop, $\lim_{t \to -\infty} f_t$ exists and is a critical point of E.

For any conjugacy class C of homomorphisms $S^1 \to G$, let X_C and X^C denote the parts of X that tend to a point of C as $t \to \infty$ and as $t \to -\infty$, respectively. Note that X is the disjoint union of the X_C and $X_{\text{pol}} = L_{\text{pol}}G/G$ is the disjoint union of the X^C.

Theorem 8.

(a) For any conjugacy class C, X_C and X^C are locally closed complex submanifolds of X of finite codimension and finite dimension, respectively.
(b) The intersection of X_C and X^C is transverse and consists of the conjugacy class C.
(c) The stratum X_1 corresponding to the identity homomorphism is open and dense in the identity component of X.
(d) If λ is any homomorphism in C,

$$X_C = LG_{\mathbb{C}} \cdot \lambda$$
$$X_C = L^+_{\text{pol}}G_{\mathbb{C}} \cdot \lambda$$

Here, $LG_{\mathbb{C}}$ is the subgroup of $LG_{\mathbb{C}}$ consisting of the loops that are boundary values of holomorphic maps from the disc $|z| > 1$ in the Riemann sphere into $G_{\mathbb{C}}$.

The first half of part (c) is equivalent to the Birkhoff factorization theorem, which asserts that every loop f in $G_{\mathbb{C}}$ can be written as

$$f = f \cdot \lambda \cdot f^{+}$$

where $f^{+} \in L^{+}G_{\mathbb{C}}$.

C. Grassmannian Embedding

Let H be a separable infinite-dimensional Hilbert space equipped with a polarization, that is, an orthogonal splitting $H = H_{+} \oplus H$ into a pair of closed infinite-dimensional subspaces. The restricted general linear group $GL_{\text{res}}(H)$ is defined as the group of invertible operators on H whose block decomposition

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with respect to the polarization has the property that b and c are Hilbert–Schmidt operators. Then a and d are necessarily Fredholm operators, and index $(a) = -\text{index}(d)$. The group $GL_{\text{res}}(H)$ is a Hilbert Lie group and has $\mathbb{Z}$ connected components determined by the index of a. Also of importance is the subgroup $U_{\text{res}}(H)$ of unitary operators in $GL_{\text{res}}(H)$.

The restricted Grassmannian $\text{Gr}(H)$ of H is defined to be the set of closed subspaces W of H such that the orthogonal projections $W \to H_{+}$ and $W \to H_{-}$ are Fredholm and Hilbert–Schmidt, respectively. The group $GL_{\text{res}}(H)$ acts transitively on $\text{Gr}(H)$. The Grassmannian $\text{Gr}(H)$ is a Hilbert manifold and is homotopy equivalent to $\mathbb{Z} \times BU$, the universal classifying space for vector bundles.

To apply this theory to loop groups, one chooses a finite-dimensional unitary representation V of G and takes H to be the space of L^2 functions on the circle with values in V. The subspaces H_{+} and H_{-} consist of the functions whose Fourier expansions are of the form $\sum_{n \geq 0} v_n z^n$ and $\sum_{n<0} v_n z^n$, respectively. The loop group $LG_{\mathbb{C}}$ acts on H, and LG acts by unitary operators.

Theorem 9. The action of $LG_{\mathbb{C}}$ on H induces a homomorphism $LG_{\mathbb{C}} \to GL_{\text{res}}(H)$ and a smooth map $X \to \text{Gr}(H)$. Both maps are embeddings if V is a faithful representation of G.

If $G = U_n$ and V is the natural representation on $\mathbb{C}^n$, the image of X_{pol} in $\text{Gr}(H)$ can be characterized as follows:

$$X_{\text{pol}} = \{W \in \text{Gr}(H) : zW \subset W \quad \text{and} \quad z^N H_{+} \subset W \subset z^{N} H_{+}\}.$$

Similar characterizations are possible for classes of loops other than polynomial and groups other than U_n.

D. Determinant Bundle and Central Extensions

If H were a finite-dimensional space, there would be a holomorphic line bundle Det on $\text{Gr}(H)$ whose fiber over $W \in \text{Gr}(H)$ is the top exterior power of W. In the infinite-dimensional case, one can still define the determinant bundle, using the fact that operators on H of the form $1 + (\textit{traceclass})$ have well-defined determinants. But whereas, in the finite-dimensional case, Det is homogeneous under the action of $GL(H)$, in the infinite-dimensional case only a central extension $\tilde{G}L_{\text{res}}(H)$ of $GL_{\text{res}}(H)$ by C^{+} acts on Det.

By combining this with the embedding in Proposition 9, one obtains a central extension $\tilde{L}G_{\mathbb{C}}$ of $LG_{\mathbb{C}}$; up to finite coverings, every extension of $LG_{\mathbb{C}}$ by C^{+} arises in this way. In particular, this shows that all the central extensions of LG described in Section I.E have complexifications.

III. REPRESENTATION THEORY OF LOOP GROUPS

A. Positive-Energy Representations

A representation of LG is said to be of positive energy if it extends to a representation of the semidirect product $S^1 \tilde{\times} LG$, with S^1 acting on LG by rotation, and if $e^{i\theta} \in S^1$ acts by $e^{iR\theta}$, where the spectrum of the self-adjoint operator R is bounded below. The most interesting representations of LG are those of positive energy; it turns out that these representations are projective, that is, representations of a central extension of LG. Apart from that, their properties and classification are closely analogous to those of the finite-dimensional representations of G. The irreducible positive-energy representations of LG correspond to the so-called integrable highest weight representations of affine Lie algebras.

There are also representations that are not of positive energy. For any $a \in S^1$, there is a homomorphism $LG \to G$ given by evaluating the loop at a; pulling back an irreducible representation V of G by this homomorphism gives an irreducible representation $V(a)$ of LG; every finite-dimensional irreducible representation of LG is a tensor product of representations of this form. A closely related construction is to let $f \in LG$ act on LV by $(f, v)(z) = f(az)v(z)$; tensor products of such representations are generically irreducible. Neither of these types of representations are of positive energy.

B. The Fundamental Representation and The Spin Representation

The space Γ of holomorphic sections of Det is a completion of the exterior algebra $\Lambda(H_{+} \oplus \bar{H}_{-})$, where $\bar{H}$ is the orthogonal complement of H_{+}, but with the

complex conjugate complex structure. The induced action of the central extension $\tilde{G}L_{\text{res}}(H)$ on Γ is irreducible; it is called the fundamental representation of $\tilde{G}L_{\text{res}}(H)$. "Irreducible" here means that Γ contains no proper invariant subspace that is closed in the compact-open topology. Moreover, Γ is a unitary representation of $\tilde{U}_{\text{res}}(H)$, the central extension of $U_{\text{res}}(H)$ by S^1 obtained by restricting the extension $\tilde{G}L_{\text{res}}(H)$. This means that Γ contains as a dense subspace a Hilbert space on which $\tilde{U}_{\text{res}}(H)$ acts by unitary operators.

One can define in a similar way the restricted orthogonal group $O_{\text{res}}(H_{\mathbb{R}})$ of the real Hilbert space $H_{\mathbb{R}}$ underlying H. The fundamental representation can then be realized as the spin representation of $O_{\text{res}}(H_{\mathbb{R}})$. The space of this realization is a direct sum of symmetric algebras indexed by the integers.

The isomorphism between these two realizations is called the boson-fermion correspondence, by analogy with quantum field theory. The space $\Lambda(H_+ \oplus \bar{H}_-)$ of the first realization is the Hilbert space of a system of free fermions, with H_+ and H_- being the states of a single particle of positive and negative energy, respectively. Similarly, the symmetric algebra is the Fock space of a system of free bosons.

By restricting the fundamental representation, one obtains an irreducible unitary representation of $L\tilde{U}_n$, called the basic representation of LU_n.

C. Borel–Weil Theory

To construct the positive-energy representations of LG for a general compact Lie group G, one must consider the homogeneous space $Y = LG/T$, where T is a maximal torus of G. Dividing by the action of G exhibits Y as a bundle over X with fiber G/T. For simplicity, we shall only consider the case where G is simply-connected and simple. Then Y is connected and simply connected so the complex line bundles over Y are classified by their first Chern class, which is an element of

$$H^2(Y, \mathbb{Z}) \cong \mathbb{Z} \oplus H^2(G/T) \cong \mathbb{Z} \oplus \hat{T}$$

where $\hat{T}$ is the lattice of characters of T, that is, the homomorphisms $T \to S^1$. Let $L_{n,\lambda}$ be the line bundle associated to $(n, \lambda) \in \mathbb{Z} \oplus \hat{T}$. Then we have the following classification of the positive-energy representations of LG, which is a precise analog of the Borel–Weil theorem, which describes the finitedimensional representations of G as sections of line bundles over G/T.

Theorem 10. Let G be a simply-connected, simple compact Lie group.

(a) Each complex line bundle $L_{n,\lambda}$ has a unique holomorphic structure.

(b) $L_{n,\lambda}$ has nonzero holomorphic sections if and only if (n, λ) is dominant, in the sense that

$$0 \le \lambda(\check{\alpha}) \le n\langle\check{\alpha}, \check{\alpha}\rangle$$

for every simple coroot $\check{\alpha}$ of G. Here, $\langle\ ,\ \rangle$ is the unique inner product on $\mathfrak{g}$ such that $\langle\check{\alpha}, \check{\alpha}\rangle = 2$ for every simple coroot $\check{\alpha}$.

(c) If (n, λ) is dominant, the space $\Gamma(L_{n,\lambda})$ of holomorphic sections of $L_{n,\lambda}$ is an irreducible positive-energy representation of a central extension of $LG_{\mathbb{C}}$.

(d) Every irreducible positive energy representation of LG arises in this way.

The positive integer n is called the level of the representation $\Gamma(L_{n,\lambda})$. The pair (n, λ) is usually regarded as an element of the dual of the Lie algebra of $\tilde{T}$, the part of the central extension $\tilde{L}G$ lying over T.

One can show also that every positive-energy representation of LG is unitary, and hence breaks up into a direct sum of representations of the above type. In particular, this implies that every positive-energy representation of LG extends to the complexification $LG_{\mathbb{C}}$. No such property holds for infinite-dimensional groups in general.

D. Kac Character Formula

V. Kac proved an analog for affine Lie algebras (in fact, for all symmetrizable Kac–Moody algebras) of the Weyl formula for the characters of the irreducible representations of G. The character of a representation V of $LG \tilde{\times} S^1$ is the formal power series

$$\chi_V = \sum_\gamma (\dim V_\gamma)\gamma(t, z)$$

where V_γ is the part of V on which the maximal Abelian subgroup $T \times S^1$ acts by the homomorphism γ. In some cases this can be interpreted as some kind of generalized function of $(t, z) \in T \times S^1$. The formula for the character is often written

$$\chi_V = \sum_\gamma (\dim V_\gamma)e^\gamma$$

identifying the homomorphism γ with an element of the dual of the Lie algebra of $T \times S^1$.

Kac Character Formula

Let L_V be the holomorphic line bundle on LG/T associated to a dominant element Λ of the dual of the Lie algebra of $\tilde{T}$, as in theorem 10. Let Γ_V be the space of holomorphic sections of L_V. Then the character of Γ_V, is

$$\chi_V = \frac{\sum_{w \in w_{\text{aft}}} (-1)^{l(w)} e^{w(V,\rho)^+ \rho}}{\prod_{\gamma,0}(1 - e^\gamma)}$$

Here, ρ is characterized by $\rho(\check{\gamma}) = 1$ for all simple roots γ of LG.

Regarded as functions of $z \in S^1$, these characters are boundary values of holomorphic functions in the disc $|z| < 1$. Since the group $SL_2(\mathbb{Z})$ acts on the disc by appropriate linear fractional transformations, it acts also on the space of such functions. One of the remarkable facts about the characters is that the finite-dimensional vector space spanned by the characters of the irreducible positive-energy representations of a given level is preserved by the action of $SL_2(\mathbb{Z})$. The proper explanation of this is to be found in conformal field theory.

E. Vertex Operators

Historically, the first realization of the basic representation was given in terms of so-called vertex operators, the construction and properties of which were known to physicists in dual resonance theory. The restriction of the central extension of LG to the Abelian group LT is an infinite-dimensional Heisenberg group, and thus has a canonical level 1 irreducible unitary representation, say H. One attempts to extend this representation, initially to the Lie algebra $L\mathfrak{g}$, and then to LG, by defining, for each coroot $\check{\alpha}$ of G, and each complex-valued function f on S^1, operators $V_\alpha(f)$ that obey the correct commutation relations with each other and with the generators of the action of LT on H. Since $\mathfrak{g}_{\mathbb{C}}$ is spanned by $\mathfrak{t}_{\mathbb{C}}$ and the coroots $\check{\alpha}$, this will accomplish the task of extending the representation to $L\mathfrak{g}$.

The construction works, at least in its simplest form, only when G is simply-laced, which means that all the simple coroots have the same length; the coroots can then be identified with the homomorphisms $S^1 \to T$ of minimal length. One defines

$$V_\alpha(f) = \int_0^{2\pi} f(\theta) V_\alpha(\theta)\, d\theta$$

for some "operator-valued distribution" $V_\alpha(\theta)$ on the circle. To define $V_\alpha(\theta)$, think of α as a homomorphism $S^1 \to T$. For each $\varepsilon > 0$, let $V_{\alpha,\theta,\varepsilon}$ be the element of LT such that

$$V_{\alpha,\theta,\varepsilon} = 1 \quad \text{if} \quad |\theta' - \theta| > \varepsilon$$

and such that $V_{\alpha,\theta,\varepsilon}(\theta')$ traces out the loop α when θ' goes from $\theta - \varepsilon$ to $\theta + \varepsilon$. Then the limit

$$\lim_{\varepsilon \to 0} \varepsilon^{-1} V_{\alpha,\theta,\varepsilon}$$

exists in the distributional sense and defines the required $V_\alpha(\theta)$.

The vertex operator construction is equivariant with respect to the action of diffeomorphisms of the circle. Recall from Section I that $\mathrm{Diff}(S^1)$ acts as a group of automorphisms of LG. Moreover, the action of the orientation-preserving diffeomorphisms $\mathrm{Diff}^+(S^1)$ preserves the cocycle defining the central extension LG, and $\mathrm{Diff}^+(S^1)$ acts on $\tilde{L}G$.

Theorem 11. There is a (projective) action of $\mathrm{Diff}^+(S^1)$ on all positive-energy representations of LG that intertwines with the action of LG.

More recently, the algebra of vertex operators has been formalized and generalized and is capable of describing representations of LG other than the basic one. Vertex algebras have also found other applications, notably to the construction of the "moonshine" representation of the Monster simple group.

IV. RELATIONS WITH OTHER PARTS OF MATHEMATICS

Due to limitations of space, we shall restrict discussion of applications of loop groups to the following two combinatorial topics.

A. Macdonald Identities

I. G. Macdonald discovered a remarkable series of multivariable power series identities, one for each simple compact Lie group G. When $G = SU_2$, this gives Jacobi's formula

$$\eta(q)^3 = \sum_{j \in \mathbb{Z}} (-1)^j q^{(1/24)(6j+1)^2}$$

for the cube of the Dedekind eta-function

$$\eta(q) = q^{1/24} \prod_{j=1} (1 - q^j).$$

For a general G, Macdonald's identities give an analogous formula for $\eta(q)^{\dim G}$. Kac and R. V. Moody independently observed that Macdonald's identities are precisely the statement of the Kac character formula for the triyial representation (the socalled denominator formula).

More recently, J. Lepowsky has shown that many other well-known power series identities have a Lie algebraic proof.

B. Quivers

A quiver is a finite set of points together with directed arrows joining certain pairs of points. A representation of a quiver is an assignment to each point i of a finite-dimensional vector space V_i and to each arrow going from point i to point j of a linear map from V_i to V_j. Quivers having only finitely many indecomposable representations correspond exactly to the Dynkin diagrams of simply-laced finite-dimensional complex simple Lie algebras. Quivers whose representations can be classified

correspond exactly the Dynkin diagrams of affine Lie algebras that contain no double bonds.

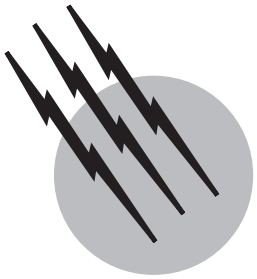
SEE ALSO THE FOLLOWING ARTICLE

MANIFOLD GEOMETRY

BIBLIOGRAPHY

Kac, V. G. (1985). "Infinite Dimensional Lie Algebras," 2nd ed., Cambridge Univ. Press, Cambridge.

Pressley, A. N., and Segal, G. B. (1986). "Loop Groups," Oxford Univ. Press, Oxford.



Manifold Geometry

C. T. J. Dodson
University of Manchester Institute of Science and Technology

I. Preliminary Notions
II. Geometrical Spaces
III. Manifolds and Bundles
IV. Calculus of Sections
V. Metric Geometry
VI. Connection Geometry
VII. Singular Geometry
VIII. Topology, Geometry, and Physics

GLOSSARY

Bundle Superstructure over a base manifold used in the definition and analysis of geometrical entities.
Cohomology Algebraic procedure for studying global properties of spaces on manifolds, via differential forms.
Completeness Freedom from singularities, such as holes, in the underlying manifold.
Connection Fundamental geometrical object for prescribing parallelism structures on a manifold.
Curvature Tensor field carrying information about the noncommutativity of parallel transport.
Differential form Section of a bundle of antisymmetric tensors.
Geodesic General geometrical equivalent of a straight line in Euclidean geometry.
Metric tensor Fundamental tensor field for prescribing metrical geometry on a manifold.
Parallel transport Process available through a connection for the movement of geometrical entities along curves in the manifold.
Section "Slice" of a bundle that chooses one bundle point over each point of the base manifold.
Tensor field Section of a tensor bundle; similarly, vector fields.
Torsion Tensor field carrying information about the asymmetry of a connection.

MANIFOLD GEOMETRY is the latest contribution to one of the oldest branches of mathematics. It exploits modern abstract algebra and abstract analysis in the study of geometrical spaces, called manifolds, that are natural generalizations of spaces studied by Euclid. To handle the greater complexity, modern geometers created superstructures called bundles over manifolds. Through these bundles, generalizations of differential and integral calculus

allow a geometrization of diverse problems in abstract analysis and in physical field theories; they also allow the use of algebraic topological methods of cohomology to study global properties. Very elegant formulations of the concepts of parallelism and curvature arise from the geometrical object called a connection. This distinguishes the geodesic curves that generalize to a manifold the Euclidean notion of a straight line and also gives a test for completeness of the manifold.

I. PRELIMINARY NOTIONS

A. Sets and Maps

We shall accept as sufficient for our purposes the intuitive notion of a set as a collection of distinct and definite elements. Likewise, a map from a set X to a set Y is a rule that prescribes precisely one element of set Y to be related to each element of set X. Then we indicate it by a diagram like

$$f\colon X \to Y : x \mapsto f(x)$$

and speak of the map f with domain X that sends a typical element x in X to the element $f(x)$ in Y. A convenient way to view f is in terms of its graph, that is, the set of elements $(x, f(x))$ lying in the $X \times Y$ space. Every map $f : X \to Y$ defines another map bringing subsets of Y back to subsets of X

$$f^{\leftarrow}\colon \operatorname{sub} Y \to \operatorname{sub} X : B \mapsto \{x \in X \mid f(x) \in B\}$$

Usually, the sets X and Y that we shall be interested in will have some geometric structure and our maps will in some sense respect this structure. Typical problems reduced to geometrical essentials are:

1. Given sets X and Y with particular geometric structures, show that a map $f\colon X \to Y$ satisfying certain requirements exists.
2. Extend the domain of a given map.
3. Given a map $f\colon X \to Y$ and some geometric structure on Y, what geometric structure is induced on (i.e., pulled back to) X?
4. Given a map $f\colon X \to Y$ and some geometric structure on X, what geometric structure is coinduced on (i.e., pushed forward to) Y?

B. Algebra and Analysis

It turns out that two mathematical areas are intertwined in the precise formulation and solution of geometrical problems. On the one hand we need analysis to deal with concepts of nearness and change; on the other hand we need algebra to organize our calculus procedures and to represent symmetries in the geometry. Our first task is to describe the appropriate parts of analysis and algebra to explain the above use of the term "geometric structure" on a set and to explain the kinds of requirements that are often imposed on maps that arise in geometrical contexts. We shall attempt to present them together in a geometrical setting with examples. In order to have a fund of examples we need to assume a working knowledge of certain mathematical raw materials, abbreviations, and basic tools.

We shall suppose some familiarity with elementary algebra and calculus on the real and complex number fields. The analysis on them depends on the modulus or absolute value function $a \to |a|$, which yields a distance function d between two numbers a and b by

$$d(a, b) = |a - b|$$

This satisfies the geometrical requirements:

$d(a, b) = d(b, a)$	(symmetry)
$d(a, b) = 0$ if and only if $a = b$	(positive definiteness)
$d(a, c) \leqq d(a, b) + d(b, c)$	(triangle inequality)

The algebra of the real and complex numbers depends on each of them having double group structures, one under addition with identity zero and one under multiplication (excepting zero) with identity one. Recall that a group is a set with a binary operation that is associative, has identity, and admits inverses.

C. Notations

Symbol	Meaning
$\mathbb{N}$	the set of natural numbers, i.e., 1, 2, 3, ...
$\mathbb{Z}$	the set of integer numbers, i.e., 0, ± 1, ± 2, ...
$\mathbb{Q}$	the set of rational numbers, i.e., p/q for $q \neq 0$, p, q in $\mathbb{Z}$
$\mathbb{R}$	the set of real numbers, i.e., the complete geometrical line
$\mathbb{C}$	the set of complex numbers, i.e., $x + iy$ for $i = \sqrt{-1}$, x, y in $\mathbb{R}$
E^2	the real plane, i.e., two-dimensional real space
E^n	n-dimensional real space for n in $\mathbb{N}$: $E^1 = \mathbb{R}$
$\mathbb{R}^n$	the standard n-dimensional real vector space
$[0, 1]$	the interval $\{x \in \mathbb{R} \mid 0 \leqq x \leqq 1\}$
$[0, 1)$	the interval $\{x \in \mathbb{R} \mid 0 \leq x < 1\}$
$x \in X$	x is an element of set X
$A \subseteq B$	A is a set, all of whose elements are also elements of B; a *subset* of B
sub B	the set of all subsets of set B
$A = B$	$A \subseteq B$ and $B \subseteq A$; A and B are the same set
$\emptyset$	the empty set; the set with no elements
$A \cap B$	the set of elements in both subsets A and B; their *intersection*

$A \cup B$	the set of elements in one or both of subsets A or B; their *union*
$\{x \in X \mid P(x)\}$	the set of elements in X for which statement $P(x)$ is true
$\Rightarrow$	implies; then
$\Leftrightarrow$	implies and is implied by; if and only if
$A \times B$	the set of ordered pairs (a, b) such that a is in A and b is in B
$X \to Y$	a map from set X to set Y
$X \twoheadrightarrow Y$	a map from set X *onto* (the whole of) set Y
$X \simeq Y$	an equivalence of structures (iso/homeo/ diffeomorphism)
$f \circ g$ or fg	the composite map; do g then f

EXAMPLE. Denote by G the multiplicative group of unit modulus complex numbers, that is,

$$G = \{z \in \mathbb{C} \mid |z| = 1\} = \{e^{i\theta} \mid \theta \in \mathbb{R}\}$$

This group *acts on* the set of complex numbers $\mathbb{C}$ in a simple way:

$$\alpha\colon G \times \mathbb{C} \to \mathbb{C} : (e^{i\theta}, w) \mapsto e^{i\theta} w$$

that is simply to rotate each complex number $w = re^{i\phi}$ by the angle θ to give $e^{i\theta} \times re^{i\phi} = re^{i(\theta+\phi)}$, for each choice of ϕ (Fig. 1). This α is a continuous action since

1. Nearby θ values send a fixed w to nearby numbers in $\mathbb{C}$;
2. Nearby w numbers are sent to nearby numbers by a fixed θ.

Consider seeking a real map (i.e., a function)

$$f\colon \mathbb{C} \to \mathbb{R}$$

that is invariant under the above group action. Evidently we want commutativity of a diagram

$$\begin{array}{ccc} G \times \mathbb{C} & \xrightarrow{\alpha} & \mathbb{C} \\ {\scriptstyle p_2}\downarrow & & \downarrow{\scriptstyle f} \\ \mathbb{C} & \xrightarrow{f} & \mathbb{R} \end{array} \quad \text{that is}$$

$$\begin{array}{ccc} (\theta, w) & \longmapsto & e^{i\theta} w \\ \downarrow & & \downarrow \\ w & \longmapsto & ? \end{array} \quad \text{on elements}$$

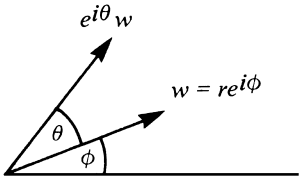


FIGURE 1 Rotation through angle θ.

Hence we seek f satisfying, for all θ and w, the condition

$$f(e^{i\theta} w) = f(w)$$

This requirement is simply rotational symmetry; that is, $f(w)$ is independent of the angular position of w. One such function is the modulus, defined for all $w = x + iy$ by

$$f(x + iy) = (x^2 + y^2)^{1/2}$$

Try repeating this example for the same group when acting as a rotation about a point other than the origin.

Whenever we have a group action on a set we wish to know the subsets left invariant by the action—if there are any. In our example they do exist and in fact they are the circles with center the origin; we call such subsets the orbits of the group action. They always form a partition of the set carrying the action into disjoint nonempty subsets. Orbits that consist of single points are called *fixed points* (the origin is one in the example). Orbits need not look like planetary orbits; for example, the action on the Euclidean plane E^2 given for the additive group of real numbers by

$$\mathbb{R} \times E^2 \xrightarrow{\alpha} E^2 : (a, (x, y)) \mapsto (x + a, y)$$

has no fixed point and orbits the horizontal lines. This latter group action is simply translation by a real number along the x axis, but in fact any other direction (or all directions) could have been used. Indeed, if we combine all rotations, translations, and reflections into one collection we get the Euclidean group, so called because it leaves invariant the essentials of Euclidean geometry, namely lengths and angles.

D. Geometrical Spaces

Euclidean geometry is the usual geometry of real n-space E^n and it is algebraically easy to handle because E^n is an affine space; this simply means that relative to any choice of origin it is equivalent to a vector space. Non-Euclidean n-dimensional geometries arise from the smooth patching together of pieces of E^n without regard for the preservation of their algebraic structures. Such patchwork structures are called n-manifolds, and they are our principal domain spaces in general geometry. Familiar examples of 2-manifolds are a sphere and a torus; evidently each has some residual local similarity to pieces of the Euclidean plane E^2 from which they are synthesized, but globally they are very different. One geometrical difference is already apparent: in the Euclidean plane the angle sum of any triangle is 180°, but it is easy to see that on a sphere we can find a triangle with sides the arcs of three great circles and having an angle sum of 270°. In fact, this angular excess is a manifestation of the presence of curvature. The formal way to handle it in general geometry is via an entity called a *connection*, which governs the definition of

parallelism on a manifold: going due north on a sphere is a curve that propagates in a parallel fashion on the curved surface, that is, maintains its direction as closely as it can. Intuitively, a connection structure provides a general geometer with a tool somewhat like the parallel rulers used by navigators.

Once we can say what we mean by the "same direction" at two different points it is meaningful to pose problems like: what is the rate of change in that direction of some entity defined on the space? By means of a connection we obtain a differential operator that precisely measures all such rates of change in an invariant (sometimes called covariant) way, that is, independently of the manner in which we choose to label the directions and points. The operator involved here is the covariant derivative and it replaces ordinary derivatives in non-Euclidean geometries, so allowing the representation of very intricate systems of differential equations such as those used in electrical engineering or for physical field theories.

Probably the most important differential operator on a manifold is the exterior derivative; this comes free with the manifold: no extra structure need be assumed. In quite a precise way it generalizes the gradient operation on scalar functions and the curl operation on vector functions that are used in vector calculus on E^3. In that sense it turns out that curvature is the "curl of the connection." We shall see how the exterior derivative characterizes some intrinsically topological properties of manifolds by means of de Rham cohomology theory.

Perverse though it may seem, no sooner do mathematicians set up a nice machinery than there arises an interest in how and why the machinery could be made to seem inadequate or defective. Clearly, if well-founded logically, then the theory (e.g., general geometry) will not have intrinsic defects. However, near the "edges" of its validity some interesting problems arise. For example, general geometry can be asked to model the environment of a physical singularity, like a black hole. This means devising a geometrical space in which something goes wrong; namely, particles disappear if they follow certain paths. It is an interesting recent result that in such a model the singularity seems to be stable: we cannot make it go away by perturbing the geometry, as done for example in quantization procedures.

II. GEOMETRICAL SPACES

A. Euclidean Spaces

The simplest Euclidean space is the one-dimensional case of the real line, E^1. As is almost always the case in mathematical processes of generalization, there are two major steps: from one to more than one, and from finitely many to infinitely many. Thus the step from E^1 to E^2, the real plane, is intuitively difficult at first meeting; but the step on, to E^4 or E^n for any finite n, is simple because the algebra and analysis are essentially unchanged. The further step to infinite dimensions is also difficult because definitions change; the kind of hurdle that arises is like the transition for polynomial functions to power series. In fact, we shall not study infinite-dimensional spaces explicitly but they will arise incidentally.

So for our purposes, a good intuitive grasp of the geometry of the real plane E^2 is a prime asset. By suppressing coordinates and other manifestations of its "twoness" we shall be able to use the algebra and analysis for higher-dimensional situations. We summarize the geometry as follows:

1. *Points* of E^2 are ordered 2-tuples of real numbers like $p = (p_1, p_2)$.
2. There is a vector difference between pairs of points, a map, diff: $E^2 \times E^2 \to \mathbb{R}^2 : (p, q) \mapsto q - p = (q_1 - p_1, q_2 - p_2)$, and this correspondence is the best possible in the sense that if we hold one of the points fixed then it is a one-to-one correspondence between all points of E^2 and all vectors of $\mathbb{R}^2$; reversing the points reverses the vector.
3. The parallelogram law holds for vector differences of points.
4. There is a distance function defined between pairs of points, the length (or norm) of their vector difference:

$$\text{dist}(p, q) = \|\text{diff}(p, q)\|$$

We use property (1) for labeling the points of the space: in E^2 each point has two coordinates, so E^2 and $\mathbb{R}^2$ are equivalent (in bijective correspondence) as sets. The vector difference map gives us two things:

1. The set of all directions at each point; this is the tangent space at the point.
2. A simple algebraic cohesion among points that facilitates the solution of problems by trigonometric methods.

Recall that in $\mathbb{R}^2$ we have the dot product between two vectors given by

$$u \cdot v = (u_1, u_2) \cdot (v_1, v_2) = u_1 v_1 + u_2 v_2$$

This also determines the length (or norm) of a vector u by

$$\|u\| = \sqrt{u \cdot u}$$

and angle θ between nonzero vectors u and v by

$$u \cdot v = \|u\| \|v\| \cos\theta$$

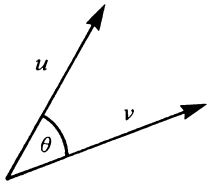


FIGURE 2 Angle between two vectors.

(see Fig. 2). Calculus on E^2 depends on the usual process of partial differentiation applied to functions of two variables taking values in $\mathbb{R}^n$. There are in particular two operators that readily generalize to E^n: grad and div.

For a differentiable real-valued map

$$f\colon E^2 \to \mathbb{R} : (x, y) \mapsto f(x, y)$$

its gradient is the vector-valued map

$$\text{grad } f\colon E^2 \to \mathbb{R}^2 : (x, y) \mapsto \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$$

grad f at $(x, y) \in E^2$ being viewed as a vector in the "tangent space" to E^2 at (x, y), that is, in the space of directions there.

For a differentiable tangent space-valued map

$$\phi\colon E^2 \to \mathbb{R}^2 : (x, y) \mapsto (\phi_1(x, y), \phi_2(x, y))$$

its divergence is the real-valued map

$$\text{div } \phi\colon E^2 \to \mathbb{R} : (x, y) \mapsto \frac{\partial \phi_1}{\partial x} + \frac{\partial \phi_2}{\partial y}$$

Of course we can apply div to grad f; then we get the *Laplacian*

$$\Delta f\colon E^2 \to \mathbb{R} : (x, y) \mapsto \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

Typical problems that arise in geometrical form are:

Given $X \subseteq E^2$ find $f\colon X \to \mathbb{R}$ satisfying $\Delta f = 0$ and subject to certain boundary conditions on f.

Given $X \subseteq E^2$ and $\phi\colon X \to \mathbb{R}$ find a curve

$$c\colon [0, 1] \to E^2 : t \mapsto (x(t), y(t))$$

beginning at (x_0, y_0), that is, $c(0) = (x_0, y_0)$, with tangent vector there given by $\dot{c}(0) = (u_0, v_0)$ and satisfying

$$\ddot{c} = \text{grad } \phi \quad \text{(at } t \text{ and } c(t) \text{ respectively)}$$

Here $\dot{c}(t) = (dx\,dt, dy/dt), \ddot{c}(t) = (d^2x/dt^2, d^2y/dt^2)$.

B. Non-Euclidean Spaces

We can obtain more general spaces, called *manifolds*, in two ways:

1. Keep the same point set (e.g., as in E^2) but use nonstandard definitions for lengths and angles of vectors in the tangent spaces.
2. Construct a new set of points (e.g., by joining pieces of E^2), then add some geometry.

In both cases we would want to meet certain compatibility conditions to avoid abrupt (i.e., not smooth) geometrical changes from place to place. In method 1 we already have all tangent spaces and may vary how we geometrize them in a smoothly changing manner. In method 2 we have only the tangent spaces for each small piece of Euclidean space so we have to make sure that they fit smoothly together, unambiguously overlapping at the joins; then we can introduce smoothly changing measures of lengths and angles for vectors over the whole space.

Method 1 would allow us to construct 2-manifolds that have the appearance of bent and twisted planes, for example, a saddle-shaped surface. However, method 1 would not allow the construction of a sphere or torus. Either of these would involve some joining together (called *identification*) of certain points of E^2. For example, to obtain a torus we begin with a rectangle in the plane, join together two edges to form a tube, then join the two open ends together. The joining process can be visualized through paper models, though this and the smoothing process at the joins can be made mathematically precise. There is a subtle point here. When we appeal intuitively to paper models it is implicit that they are viewed in three-dimensional space, and so at each point of such a model surface there is a well-defined tangent space of directions in the surface. It is actually the tangent plane at the point. In the formal mathematical process the tangent spaces can be constructed without leaving the surface. The same is true in general for n-manifolds constructed by joining and smoothing out pieces of E^n. However, there is an important theorem due to Whitney which says that we could, if we wish, view any such n-manifold as embedded in a Euclidean space of sufficiently high dimension. We saw that the torus 2-manifold needed E^3; in general an n-manifold may need E^{2n+1}.

One final point before getting to technical definitions. In geometry we often speak of "the unit circle" or "the n-sphere" when in fact there are many ways to describe these spaces. For example S^1, the unit circle, can be viewed as the set of all unimodular complex numbers or perfectly equivalently (i.e., up to structural isomorphism) as the set of vectors of unit norm in $\mathbb{R}^2$. We often do not bother to make distinctions between them.

III. MANIFOLDS AND BUNDLES

A. Manifolds

An n-manifold is a set M of points with the following properties:

1. M can support the notion of continuous functions on it; (we take M to be Hausdorff with countable base).
2. M is the union of a collection $\{U_\alpha \mid \alpha \in A\}$ of its subsets; (the collection is an open cover for M).
3. For each α in the indexing set A there is a continuous equivalence between U_α and E^n; (homeomorphisms $\phi_\alpha\colon U_\alpha \to E^n$ give coordinates).
4. The change of coordinate maps are smoothly differentiable; ($\phi_\alpha \circ \phi_\beta^{-1}$ is a diffeomorphism if $U_\alpha \cap U_\beta \neq \emptyset$).

We call the collection $\{(U_\alpha, \phi_\alpha) \mid \alpha \in A\}$ an atlas of charts for M. Properties 1–4 are clear enough but it is worth noting that 4 is meaningful because maps like $\phi_\alpha \circ \alpha_\beta^{-1}$ go between pieces of E^n on which we suppose differentiability to be well understood—namely, calculus on $\mathbb{R}^n$ attached to each point. It is property 4 that enables us to say what we mean by a map between two manifolds being differentiable. Let M' be an m-manifold with atlas $\{(V_\lambda, \psi_\lambda) \mid \lambda \in B\}$. Then a map

$$f\colon M \to M'$$

is called *differentiable* if and only if the composite maps $\psi_\lambda \circ f \circ \phi_\alpha^{-1}$ are differentiable wherever $fU_\alpha \cap V_\lambda \neq \emptyset$, for any $\alpha \in A$ and $\lambda \in B$. The reason for this is apparent from the diagram

$$\begin{array}{ccc} U_\alpha & \xrightarrow{f} & V_\lambda \cap fU_\alpha \\ {\scriptstyle \phi_\alpha^{-1}}\uparrow & & \downarrow{\scriptstyle \psi_\lambda} \\ E^n & \xrightarrow[\psi \circ f \circ \phi_\alpha^{-1}]{} & E^m \end{array}$$

Evidently, we are borrowing for manifolds the already known property of differentiability of maps from E^n to E^m. Property 4 is spoken of as giving a *differentiable* or *smooth structure* to M.

A similar trick of borrowing is employed to define T_xM, the *tangent space* to M at $x \in M$. If $x \in U_\alpha$ then we certainly have a nice vector space tangent at $\phi_\alpha(x) \in E^n$ and it is isomorphic to $\mathbb{R}^n$. The problem is that we may also have $x \in U_\beta$. In that case $\phi_\alpha \circ \phi_\beta^{-1}$ and $\phi_\beta \circ \phi_\alpha^{-1}$ have derivatives that are actually isomorphisms between the tangent spaces at $\phi_\alpha(x)$ and $\phi_\beta(x)$; in fact, they will appear as invertible Jacobian matrices with entries the partial derivatives, from calculus on $\mathbb{R}^n$. Such isomorphisms are used to take equivalence classes of E^n-tangent spaces for synthesizing T_xM. Then the process can be rounded off nicely because the derivative of a map between manifolds actually becomes a linear map between tangent spaces.

The totality of all tangent spaces to an n-manifold M is actually itself a $2n$-manifold with point set

$$TM = \bigcup_{x \in M} T_xM = \{(x, u) \mid x \in M, u \in T_xM\}$$

and atlas $\{(TU_\alpha, T\phi_\alpha) \mid \alpha \in A\}$, where $TU_\alpha = \cup_{x \in U_\alpha} T_xM$ and $T\phi_\alpha$ is the derivative of ϕ_α. We call this manifold the tangent bundle to M and it is noteworthy that it comes free with M; no further structure was needed. There is a natural smooth map onto M:

$$p\colon TM \twoheadrightarrow M : (x, u) \mapsto x$$

The *fiber* of such a map over x is

$$p^{\leftarrow}\{x\} = \{y \in TM \mid p(y) = x\}$$

which evidently coincides with T_xM and therefore looks like $\mathbb{R}^n$ for each x, though there is no unique isomorphism $T_xM \simeq \mathbb{R}^n$ actually determined for each x since each chart gives a different one. The tangent bundle is an example of an important class of structures over manifolds, the vector bundles.

B. Bundles

A *vector bundle with fiber* F (a vector space) over a manifold M is a smooth surjection $p\colon X \twoheadrightarrow M$ such that every $x \in M$ has a neighborhood U for which there is a diffeomorphism $p^{\leftarrow}U \simeq U \times F$, and isomorphism on fibers. This implies that the zero vectors in fibers are joined smoothly together and the set of all of them looks like a copy of M embedded in X. In the case of the tangent bundle over an n-manifold, $F = \mathbb{R}^n$ and the local triviality condition is

$$TU_\alpha \simeq U_\alpha \times \mathbb{R}^n$$

Thus a vector bundle over M consists of M together with copies of a vector space smoothly fitted over all points of M. We interpret the tangent bundle over M as the set of all points and directions in M.

If we have a vector bundle with fiber F over M then we may be able to use its construction to obtain another vector bundle with fiber the dual space F^* consisting of all linear functionals defined on F. One case when this is always possible is for the tangent bundle TM; we replace each tangent space T_xM by its dual $(T_xM)^*$ consisting of real-valued linear maps on T_xM. The resulting bundle is T^*M, the contangent bundle. Recall that just as we can dual vector spaces so we can dual linear maps between them by sending linear f to linear f^* where

$$\left.\begin{array}{c} F \xrightarrow{f} G \\ G^* \xrightarrow{f^*} F^* \end{array}\right\} \quad \text{with } f^*\colon \alpha \mapsto \alpha f$$

So the dual map actually goes in the reverse direction. For a finite-dimensional vector space F we may identify $(F^*)^*$, the double dual, with F itself by the natural isomorphism

$$F \simeq (F^*)^*\colon v \mapsto \hat{v}$$

where

$$v\colon F^* \to \mathbb{R} : \alpha \mapsto \alpha(v)$$

EXAMPLE. Take $M = S^1 = \{z \in \mathbb{C} \mid |z| = 1\}$, the 1-sphere. Then $TS^1 = S^1 \times \mathbb{R}$, a cylinder easily visualizable by considering each tangent line to S^1 to be represented by a vertical copy of $\mathbb{R}^1$ so that one such can be fitted to each point. Thus the tangent bundle to S^1 is a trivial product.

On the other hand, we can find another vector bundle over S^1 with fiber $\mathbb{R}^1$. This is the Möbius bundle in which the different copies of $\mathbb{R}^1$ attached to points of S^1 actually rotate through 180° as we pass once around the circle. So this bundle is not a trivial product; it is a twisted product and easily investigated by making a paper model. It is worth adding that the tangent bundle to the 2-sphere S^2 is not a trivial product, it is twisted. This is rather a deep result but quite believable when you consider the problems of stacking tangent planes on a sphere. On the other hand, there is the normal vector bundle on S^2, which is a trivial vector bundle with fiber $\mathbb{R}^1$.

Bundles can be made with fibers that are manifolds with structures other than vector spaces, for example, groups. The great importance of bundles is that they allow a crucial generalization of the concept of a map. To see this, consider the two vector bundles with fiber $\mathbb{R}^1$ over S^1 given above. Now a continuous map from S^1 to $\mathbb{R}^1$ is simply a continuous choice of a vector in $\mathbb{R}^1$ for each point in S^1. Given such a map we could draw its graph on the cylinder bundle $S^1 \times \mathbb{R}^1$ above, as a continuous curve that projects onto the whole of S^1. Equally, given such a curve then it determines uniquely a continuous map from S^1 to $\mathbb{R}^1$ by sending each point of S^1 to the vector in $\mathbb{R}^1$ cut by the curve above that point. Now consider the twisted bundle. If we draw on this a continuous curve that projects onto S^1 then we do not obtain a unique continuous map from S^1 to $\mathbb{R}^1$ because the $\mathbb{R}^1$ in which we are trying to obtain values changes continuously as we go around S^1. The precise formulation of our generalization of a map is the following.

A section of a bundle $X \overset{p}{\twoheadrightarrow} M$ with fiber F over M is a smooth map $\sigma\colon M \to X$ such that $p\sigma$ is the identity map on M. Such a section is sometimes referred to as an X-field on M. So σ smoothly chooses at each point of M an element in the fiber over that point. Thus, if $X = M \times F$, the trivial product bundle, then all sections have the form

$$\sigma\colon M \to M \times F\colon a \mapsto (a, f(a))$$

where

$$f\colon M \to F\colon a \mapsto f(a)$$

is a map from M to F.

Conversely, if X is not a trivial product bundle then there will be some $a \in M$ with neighborhoods U and V such that $a \in U \cap V$ with local triviality

$$\lambda\colon p^{\leftarrow}U \simeq U \times F \qquad \text{and} \qquad \mu\colon p^{\leftarrow}V \simeq V \times F$$

Therefore we have three diffeomorphisms

$$\begin{array}{ccc} & p^{\leftarrow}U \cap p^{\leftarrow}V & \\ \overset{\lambda}{\swarrow} & & \overset{\mu}{\searrow} \\ (U \cap V) \times F & \underset{\mu\lambda^{-1}}{\simeq} & (U \cap V) \times F \end{array}$$

Now, a section $\alpha\colon M \to X$ can be locally decomposed by means of λ and μ and $U \cap V$ thus:

$$\begin{array}{ccccc} & & & & (U \cap V) \times F \to F \\ & & & \overset{\lambda}{\nearrow} & \\ U \cap V & \overset{\sigma}{\to} & p^{\leftarrow}U \cap p^{\leftarrow}V & & \\ & & & \underset{\mu}{\searrow} & \\ & & & & (U \cap V) \times F \to F \end{array}$$

which on elements becomes of the form

$$\begin{array}{ccc} & & (a, \lambda(a)) \mapsto \lambda(a) \\ & \swarrow & \\ a \overset{\sigma}{\mapsto} \sigma(a) & & \\ & \searrow & \\ & & (a, \mu(a)) \mapsto \mu(a) \end{array}$$

Now, as long as λ and μ differ, then σ does not define any unique map from $U \cap V$ to F.

C. Combining Vector Bundles

We can obtain new vector bundles from old by exploiting two processes available for vector spaces: their direct sum $\oplus$ and their tensor product $\otimes$. Consider two finite-dimensional real vector spaces F and G. Then we can represent $F \oplus G$ and $F \otimes G$ in terms of any bases $\{f_1, \ldots, f_n\}$ for F and $\{g_1, \ldots, g_m\}$ for G as follows:

$$F \oplus G \text{ has basis } \{f_1, \ldots, f_n, g_1, \ldots, g_m\}$$
$$\text{dimension } n + m$$

$$F \otimes G \text{ has basis } \{f_i \otimes g_j \mid i = 1, \ldots, n;\ j = 1, \ldots, m\}$$
$$\text{dimension } nm$$

Thus, $F \oplus G$ arises from the disjoint union set of vectors in F and G (identifying the two zero vectors) and $F \otimes G$ arises from the product set of vectors, $F \times G$. The structure of $F \oplus G$ is clear enough; for example, if F is the (x, y) plane in $\mathbb{R}^3$ and G is the z axis then $F \oplus G = \mathbb{R}^2 \oplus \mathbb{R}^1 \simeq \mathbb{R}^3$. Less clear is $F \otimes G$, but in concrete applications we shall see a specific role provided for the basis vectors $f_i \otimes g_j$. With bases $\{\hat{i}, \hat{j}\}$ for $\mathbb{R}^2$ and $\{\hat{k}\}$ for $\mathbb{R}^1$ we have a basis $\{\hat{i} \otimes \hat{k}, \hat{j} \otimes \hat{k}\}$ for $\mathbb{R}^2 \otimes \mathbb{R}^1$ and so $\mathbb{R}^2 \otimes \mathbb{R}^1 \simeq \mathbb{R}^2$.

Both products $\oplus$ and $\otimes$ can be extended in an obvious way to linear maps between spaces. For example, denote by $L(F;J)$ the vector space of linear maps from vector space F to vector space J. Then with $f \in L(F;J)$ and $g \in L(G;K)$ we obtain new linear maps:

$$\begin{aligned} f \oplus g &: F \oplus G \to J \oplus K \\ &: x \oplus y \mapsto f(x) \oplus g(y) \\ f \otimes g &: F \otimes G \to J \otimes K \\ &: x \otimes y \mapsto f(x) \otimes g(y) \end{aligned}$$

In particular, $F^* = L(F;\mathbb{R})$ and a very useful equivalent formulation of $L(F;J)$ itself is obtained:

$$L(F;J) \simeq F^* \otimes J$$

For example $(\mathbb{R}^3)^* \otimes \mathbb{R}^2$ can therefore be interpreted as the space of linear maps from $\mathbb{R}^3$ to $\mathbb{R}^2$, namely the space of 2×3 matrices.

There is a third composition, also based on the tensor product, available for a vector space F with itself. This is the alternating or exterior product $F \wedge F$. In terms of the basis $\{f_1, \ldots, f_n\}$ for F:

$$\begin{aligned} F \wedge F \text{ has basis } &\{f_i \otimes f_j - f_j \otimes f_i \mid i < j\} \\ &\text{and so dimension } \tfrac{1}{2}(n^2 - n) \end{aligned}$$

We usually write $f_i \wedge f_j = \frac{1}{2}(f_i \otimes f_j - f_j \otimes f_i)$. Hence $\mathbb{R}^2 \wedge \mathbb{R}^2$ has basis $\{\hat{\imath} \otimes \hat{\jmath} - \hat{\jmath} \otimes \hat{\imath}\}$ and so $\mathbb{R}^2 \wedge \mathbb{R}^2 \simeq \mathbb{R}^1$. Observe that we can only have $(F \wedge F) \simeq F$ if $\frac{1}{2}(n^2 - n) = n$, that is, if $n = 3$. This fact is closely related to the existence on $\mathbb{R}^3$ only of the vector cross product $\times$ defined by

$$\hat{\imath} \times \hat{\jmath} = \hat{k} \qquad \hat{\jmath} \times \hat{k} = i \qquad \hat{\imath} \times \hat{k} = -\hat{\jmath}$$

For, we can make an isomorphism

$$\mathbb{R}^3 \wedge \mathbb{R}^3 \simeq \mathbb{R}^3 \times \mathbb{R}^3$$

$$\text{by mapping linearly } \begin{cases} \hat{\imath} \wedge \hat{\jmath} \mapsto \hat{\imath} \times \hat{\jmath} \\ \hat{\imath} \wedge \hat{k} \mapsto \hat{\imath} \times \hat{k} \\ \hat{\jmath} \wedge \hat{k} \mapsto \hat{\jmath} \times \hat{k} \end{cases}$$

Once again we can use the product $\wedge$ on spaces to give a product on linear maps. For $f, g \in L(F;J)$ we obtain

$$f \wedge g : F \wedge F \to J \wedge J : x \wedge y \mapsto f(x) \wedge f(y)$$

These vector space processes can be performed smoothly over a manifold to the fibers of vector bundles. In particular, we have automatically the following vector bundles derived from the tangent bundle TM and cotangent bundle T^*M to a smooth manifold:

$$\begin{aligned} T_0^2 M &= TM \otimes TM && \text{with fiber } \mathbb{R}^n \otimes \mathbb{R}^n \\ T_2^0 M &= T^*M \otimes T^*M && \text{with fiber } (\mathbb{R}^n)^* \otimes (\mathbb{R}^n)^* \\ T_1^1 M &= TM \otimes T^*M && \text{with fiber } (\mathbb{R}^n) \otimes (\mathbb{R}^n)^* \\ A^2 M &= T^*M \wedge T^*M && \text{with fiber } (\mathbb{R}^n)^* \wedge (\mathbb{R}^n)^* \end{aligned}$$

It is clear enough how to get more factors with $\otimes$; less clear is how to get $A^{r+s}M$ from $A^r M$ and $A^s M$. We use the alternating operator (hence the A in $A^2 M$)

$$A_k : F^* \otimes F^* \otimes \cdots \otimes F^* \to A^k F^* : w \mapsto w_A$$

where

$$\begin{aligned} w_A &: F \times F \times \cdots \times F \to \mathbb{R} : (v_1, \ldots, v_k) \\ &\mapsto \frac{1}{k!} \sum_\tau \operatorname{sgn}(\tau) w\big(v_{\tau(1)}, v_{\tau(2)}, \ldots, v_{\tau(k)}\big) \end{aligned}$$

τ runs over all $k!$ permutations of $\{1, 2, \ldots, k\}$, and $\operatorname{sgn}(\tau)$ is ± 1 according as τ is an even or an odd permutation. Then we can define

$$\begin{aligned} (A^r F^*) \wedge (A^s F^*) &= A^{r+s} F^* \\ &= A_{r+s}(A^r F^* \otimes A^s F^*) \end{aligned}$$

which agrees with what we gave before for $r = s = 1$, and the exterior product turns out to be associative and distributive but anticommutative:

$$w \wedge v = (-1)^{rs} v \wedge w, \quad w \in A^r F^*, \quad v \in A^s F^*$$

It follows that if $\dim F = \dim F^* = n$ then

$$\dim A^r F^* = \binom{n}{r} = \frac{n!}{(n-r)!r!}$$

and therefore

$$\dim A^r F^* = \binom{n}{r} = \binom{n}{n-r} = \dim A^{n-r} F^*$$

so in particular

$$\begin{aligned} \dim A^n F^* &= \dim A^0 F^* = 1 \\ \dim A^k F^* &= 0 \quad \text{for } k > n \end{aligned}$$

All of this exterior algebra can be carried over to apply smoothly over a manifold M and hence give rise to $A^r M$, the alternating tensor bundles of differential r-forms or r-form bundles. The collection of all r-forms for all $r = 0, 1, \ldots$ on an n-manifold M is denoted ΛM and the exterior product gives ΛM the structure of a Grassmann or exterior algebra. The collection of all r-forms on M is denoted $\Lambda^r M$ so ΛM is the disjoint union of these over all $r = 0, 1, \ldots$. We call the $T_s^r M$ the tensor bundles and we usually adopt the conventional notation

$$T_0^0 M = A^0 M = M \times \mathbb{R}$$

for the trivial product vector bundle with fiber $\mathbb{R}^1$ over M. Much of general geometry and its applications is concerned with sections of these bundles. For example, we may point out that

1. A metric structure is determined by a section of $T_2^0 M$.
2. A connection (i.e., parallelism) structure is determined by a section of $A^2 M$.
3. Curvature of a connection is determined by a section of $T_3^1 M$.
4. On an n-dimensional configuration space of particles: the kinetic energy is a section of $T_2^0 M$ and the Lagrangian is a section of $A^n M$.

Notation

It is important to distinguish between a bundles $X \twoheadrightarrow M$ over M and its space of sections, which sometimes is denoted $\chi(X)$ or $\Gamma(X)$. However, it is common to speak of, for example, the bundle $A^r M$ of r-forms on M when properly $A^r M \xrightarrow{p} M$ is the bundle and the collection of its sections is quite different, denoted $\Lambda^r M$ by us; precisely

$$\Lambda^r M = \{\text{smooth } w\colon M \to A^r M \mid pw = 1_M\}$$

Similarly, we use $\Upsilon_s^r M$ to denote the collection of all sections of $T_s^r M$, that is, the (r, s)-tensor fields:

$$\Upsilon_s^r M = \left\{\text{smooth } v\colon M \to T_s^r M \;\middle|\; pv = 1_M\right\}$$

D. Manifolds with Boundary

The alert reader may have noticed that our definition of an n-manifold disqualifies, for example, a unit cylinder and the closed unit disk from being 2-manifolds. They simply fail to be locally like E^2 at their boundary edges; instead the edge points are homeomorphic to half-spaces like

$$\{(x, y) \in E^2 \mid y \geqq 0\}$$

Evidently such entities are likely to be needed in geometry and physics, so we shall allow such edge points in future. Spaces that have them are called *manifolds with boundary* and we shall denote the boundary of such an M by ∂M. In fact, if M is an n-manifold then ∂M is an $(n-1)$-manifold.

EXAMPLES

a. $M = \{x \in \mathbb{R}^3 \mid \|x\| \leqq 1\} =$ closed 3-ball B^3
   $\partial M = \{x \in \mathbb{R}^3 \mid \|x\| = 1\} \simeq$ 2-sphere S^2
b. Similarly, $\partial B^n \simeq S^{n-1}$ for $n \leq 1$
c. $M = [0, 1] \times S^1 =$ unit cylinder
   $\partial M = [0] \times S^1 \cup \{1\} \times S^1 \simeq$ two disjoint circles

IV. CALCULUS OF SECTIONS

A. Module of Sections

Intuitively we viewed a section of a bundle X over M as a copy of M lifted vertically up into the manifold X in such a way that each point of M goes to the fiber over itself. For this to be possible at all we need the projection of X to be surjective onto M; this we denote by $X \xrightarrow{p} M$. In a precise technical sense a section $\sigma\colon M \to X$ is actually a lift of the identity map $1_M\colon M \to M$ since we want Fig. 3 commuting, namely such that $p(\sigma(x)) = x$, that is, $\sigma(x) \in p^{\leftarrow}\{x\}$ for all $x \in M$. Vector bundles always have one smooth such section, the zero section which sends each $x \in M$ to the zero vector in the fiber $p^{\leftarrow}\{x\}$. Some vector bundles (e.g., the trivial product bundles) have a full set of linearly independent sections; such sections, of course, can never cross the zero section. Other vector bundles (e.g., the Möbius bundle over the circle, or TS^2, the tangent bundle to the ordinary sphere) do not have even one continuous section that never meets the zero section.

The fact that sections take values in fibers means that over any point in M we may use available algebraic structure in the fiber on sections passing through it. Now, the fibers are joined smoothly together, so not surprisingly we find that their algebraic structures change, but only smoothly, so we can apply much of the algebra to sections themselves. In particular, the set of sections of a vector bundle over a manifold is a vector space (infinite-dimensional) with pointwise addition and multiplication by numbers in each fiber. Furthermore, we can actually allow multiplication by numbers that vary smoothly from point to point, that is by sections of the trivial scalar bundle; this means that the set of sections of a vector bundle over a manifold is a module over the *ring* of smooth scalar-valued maps on the manifold. Thus, if σ, τ are two sections of the real vector bundle $X \twoheadrightarrow M$ and $\lambda\colon M \to \mathbb{R}$ is a smooth map, then also $\sigma + \tau$ and $\lambda\sigma$ are sections of $X \twoheadrightarrow M$ with

$$\sigma + \tau\colon M \to X : x \to \sigma(x) + \tau(x)$$

$$\lambda\sigma\colon M \to X : x \to \lambda(x)\sigma(x)$$

Much of geometrical analysis and its applications depend on the study of sections of the tensor bundles; these

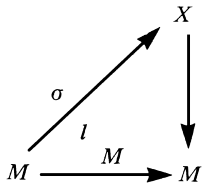


FIGURE 3 Commutative diagram for a section.

are the tensor fields. As we have seen, they are built up from sections of the tangent and cotangent bundles via the tensor product. Accordingly, we shall begin by a study of tangent and cotangent vector fields.

B. Tangent Vector Fields

From the canonical structure of TM inherent in the differentiable structure on an n-manifold M, it turns out that for each chart

$$\phi\colon U \to E^n : x \mapsto (x_1, x_2, \ldots, x_n)_x$$

the tangent space $T_x M$ has a basis $(\partial_1, \partial_2, \ldots, \partial_n)_x$, where each ∂_i is the partial differential operator $\partial/\partial x_i$ for real functions defined on neighborhoods of $\phi(x)$. Hence, for all tangent vector fields $\sigma\colon M \to TM$, on the domain of such a chart, σ appears in the form

$$\sigma\colon U \to TM : x \mapsto \left(\sigma^1\partial_1 + \sigma^1\partial_2 + \cdots + \sigma^n\partial_n\right)_x$$

with each σ^i a real function on U. A change of chart induces a change of basis and corresponding change of components of σ, both in agreement with the usual transformation law for partial derivatives. Viewing ΥM, the space of sections of TM, as a module, it is convenient to think of the map $x \mapsto (\partial_1, \partial_2, \ldots, \partial_n)_x$ as giving a basis for this module in the neighborhood U of x.

EXAMPLE. Let $M = S^2$ be the unit sphere in $\mathbb{R}^3$, explicitly:

$$M = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$$

Define two charts (U, ϕ), $(V, \tilde{\phi})$ for

$$U = \{(x, y, z) \in M \mid z > 0\}$$
(the "northern hemisphere")
$$V = \{(x, y, z) \in M \mid x > 0\}$$
(the "eastern hemisphere")

by

$$\phi\colon U \to \mathbb{R}^2\colon (\mathrm{x}, \mathrm{y}, \mathrm{z}) \mapsto (x, y) = (x_1, x_2), \text{ say}$$
$$\tilde{\phi}\colon V \to \mathbb{R}^2 : (\mathrm{x}, \mathrm{y}, \mathrm{z}) \mapsto (y, z) = (\tilde{x}_1, \tilde{x}_2), \text{ say}$$

Then on $U \cap V = \{(x, y, z) \in M \mid x > 0, z > 0\}$ (the NE quadrant) we have alternative coordinates: the (x_i) or the $(\tilde{x}_i)$. These determine corresponding local basis sections:

$$(\partial_1, \partial_2) = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$$
$$(\tilde{\partial}_1, \tilde{\partial}_2) = \left(\frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)$$

Evidently these are related (invertibly) by

$$\tilde{\partial}_1 = \partial_2 \qquad \text{and} \qquad \tilde{\partial}_2 = \frac{\partial x}{\partial z}\partial_1 + \frac{\partial y}{\partial z}\partial_2$$

with $x = (1 - y^2 - z^2)^{1/2}$ and $y = (1 - x^2 - z^2)^{1/2}$. Also on $U \cap V$, we could introduce angular coordinates by, say,

$$\tilde{\phi}\colon U \cap V \to \mathbb{R}^2 : (\mathrm{x}, \mathrm{y}, \mathrm{z}) \to (\theta, \psi) = (\hat{x}_1, \hat{x}_2)$$

where θ and ψ are the angles defined by

$$\left.\begin{aligned} x &= \cos\psi\cos\theta \\ y &= \cos\psi\sin\theta \\ z &= \sin\psi \end{aligned}\right\} \begin{array}{l} \text{so } \psi = 0 \text{ on the equator and } \pi/2 \text{ at} \\ \quad \text{the north pole} \\ \text{and } \theta = 0 \text{ in the positive } x \text{ direction} \\ \theta = \pm\pi/2 \text{ in the } \pm y \text{ directions} \end{array}$$

Again we get a local basis section

$$(\hat{\partial}_1, \hat{\partial}_2) = \left(\frac{\partial}{\partial\theta}, \frac{\partial}{\partial\psi}\right)$$

which is related to the original choice (∂_1, ∂_2) by

$$\hat{\partial}_1 = \frac{\partial x}{\partial\theta}\partial_1 + \frac{\partial}{\partial\theta}$$
$$\hat{\partial}_2 = \frac{\partial x}{\partial\psi}\partial_1 + \frac{\partial y}{\partial\psi}\partial_2$$

In fact, if we wish to exploit the sphericity of S^2 then the chart with (θ, ψ) coordinates is very convenient. For instance,

$$U \cap V \to TM : (\theta, \psi) \mapsto \cos\psi\,\hat{\partial}_1$$

locally models an east–west wind on the earth, decaying from the equator to the north pole. Similarly,

$$U \cap V \to TM : (\theta, \psi) \mapsto -\cos\psi\,\hat{\partial}_2$$

models a north wind.

A local flow about $x_0 \in M$ for a vector field v on M is a map for some neighborhood U of x_0 and positive ε,

$$\Phi\colon U \times (-\varepsilon, \varepsilon) \to M$$

such that $(\forall x \in U)$

a. $\Phi(x, 0) = x$.
b. $c_x\colon (-\varepsilon, \varepsilon) \to M : t \mapsto \Phi(x, t)$ is a curve with tangent vector $\dot{c}_x(t) = v(c(t))$.

Intuitively we think of Φ as a family of curves that "join up the arrows" of the vector field v. We call each curve c_x an integral curve of v through x.

EXAMPLE. Take U, V, and M as in the previous example. Then a local flow on S^2 for the vector field

$$v\colon U \cap V \to TS^2 : (\theta, \psi) \mapsto \cos\psi\,\hat{\partial}_2$$

is given by

$$\Phi\colon U \cap V \times (-\varepsilon, \varepsilon) \to M : (\theta, \psi, t) \mapsto (\alpha(t), \beta(t))$$

where the real functions, α, β must satisfy

$$\frac{d\alpha}{dt} = \cos\psi, \qquad \frac{d\beta}{dt} = 0$$
$$(\alpha(0), \beta(0)) = (\theta, \psi)$$

Hence $\alpha(t) = \theta + t\cos\psi$, $\beta(t) = \psi$, and the integral curves are parts of circles of latitude. There is a nice existence and uniqueness theorem for smooth local flows of smooth vector fields on manifolds. It says that through each point there is one and only one integral curve and $\Phi_t\colon x \mapsto \Phi(x, t)$ satisfies $\Phi_{t+s} = \Phi_t \circ \Phi_s$ when t, s, $(t+s) \in (-\varepsilon, \varepsilon)$. For our flow on S^2 this is satisfied for small enough s, t since $\Phi_t\colon (\theta, \psi) \mapsto \theta + t\cos\psi$. Moreover, each Φ_t *is a diffeomorphism* from U to $\Phi_i U$.

EXAMPLE. To find integral curves and hence a local flow for a vector field. We consider $M = E^2$ with its standard chart. Take the vector field

$$v\colon E^2 \to TE^2 : (x, y) \mapsto \partial_x + x^2\partial_y$$

It is often convenient to use (x, y) as coordinate labels instead of (x^1, x^2); then we denote their induced basis fields by $\partial_x = \partial_1$ and $\partial_y = \partial_2$. We take the open subset of E^2 given by

$$U = \left\{(x, y) \in E^2 \mid |x + 1| < \tfrac{1}{4}, |y| < \tfrac{1}{4}\right\}$$

We seek for each $a \in U$ an integral curve

$$c_a\colon \left(-\tfrac{1}{2}, \tfrac{1}{2}\right) \to E^2$$
$$\text{with } c_a(0) = a \text{ and } \dot{c}_a(t) = v(c_a(t))$$

The differential equation expands into

$$\dot{c}_a(t) = \dot{c}^1(t)\,\partial_x + \dot{c}^2(t)\,\partial_y = \partial_x + (c^1(t))^2\,\partial_y$$

Hence $\dot{c}^1(t) = 1$ and $\dot{c}^2(t) = (c^1(t))^2$. Therefore $c^1(t) = k + t$ and $c^2(t) = \frac{1}{3}(k+t)^3 + l$. Taking $a = (\alpha, \beta)$ as the initial point, we find

$$c_a(t) = \left(\alpha + t, \beta - \tfrac{1}{3}\alpha^3 + \tfrac{1}{3}(\alpha + t)^3\right)$$

Accordingly, the local flow is given by

$$\Phi_s\colon (\alpha, \beta) \mapsto \left(\alpha + s, \beta - \tfrac{1}{3}\alpha^3 + \tfrac{1}{3}(\alpha + s)^3\right)$$

and we easily check that

$$\Phi_{t+s} = \Phi_t\Phi_s\colon (\alpha, \beta) \mapsto \left(\alpha + s + t,\ \beta - \tfrac{1}{3}\alpha^3 + \tfrac{1}{3}(\alpha + s + t)^3\right)$$

C. Cotangent Vector Fields

Given an n-manifold M and a chart

$$\Phi\colon U \to E^n : x \mapsto (x_1, x_2, \ldots, x_n)_x$$

we obtain a set of local tangent vector fields

$$\partial_i\colon U \to TU : x \mapsto (\partial_i)_x \quad i = 1, 2, \ldots, n$$

that generate all tangent vector fields over U. Now, an element of T^*M is a dual vector to some tangent vector on M, that is, an element of

$$(T_xM)^* = L(T_xM; \mathbb{R}) \quad \text{for some } x \in M$$

Also, if $(b_1, \ldots, b_n)$ is a basis for T_xM then it is easy to show that $(c^1, \ldots, c^n)$ is a basis for $(T_xM)^*$ with

$$c^j\colon T_xM \to \mathbb{R} : \lambda_1 b_1 + \cdots + \lambda_n b_n \mapsto \lambda_j \qquad j = 1, 2, \ldots, n$$

Hence in terms of the Kronecker symbol

$$c^j(b_i) = \delta_i^j = \begin{cases} 1 & \text{if } i = j, \quad i, j = 1, 2, \ldots, n \\ 0 & \text{if } i \neq j \end{cases}$$

and we call $(c^1, \ldots, c^n)$ the *dual basis* to the basis $(b_1, \ldots, b_n)$. Clearly we can extend this process to the fields $(\partial_1, \ldots, \partial_n)$ and obtain their duals $(dx^1, \ldots, dx^n)$ with

$$dx^j(\partial_i) = \delta_i^j, \quad i, j = 1, 2, \ldots, n$$

This notation is standard and should not be confused with the unfortunate usage of dx^j in older texts to mean a small increment in coordinate x^j. The correct way to view any df is as the gradient of a real-valued function f on M. With respect to our chart diffeomorphism

$$\Phi\colon U \to E^n : x \mapsto (x_1, \ldots, x_n)_x$$

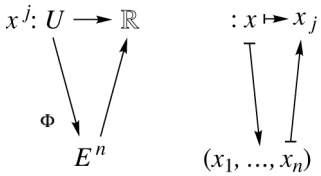
we get a restriction of f to U. Now we define

$$df\colon U \to T^*M : x \mapsto \left(f_1\,dx^1 + \cdots + f_n\,dx^n\right)_x$$

where for each $i = 1, \ldots, n$

$$f_i\colon E^n \to \mathbb{R} : (x_1, \ldots, x_n) \mapsto \frac{\partial f \circ \Phi^{-1}}{\partial x^i}$$

So in particular if we choose f to be x^j, that is,



then

$$dx^j\colon U \to T^*M : x \mapsto \left(\frac{\partial x_j}{\partial x_1}dx^1 + \cdots + \frac{\partial x_j}{\partial x_n}dx^n\right)_x$$

But $\partial x^j/\partial x^i = \delta_i^j$ so $dx^j(x) = (dx^j)_x$, $j = 1, 2, \ldots, n$.

We have been careful so far to put subscript x on various entities to indicate that they vary with x. This is not always necessary and where it is clear whether we mean, for instance, a vector at x or a field about x, we shall omit such subscripts. There is another very useful notational trick: the summation convention attributed to Einstein. This is simply to sum over repeated upper and lower indices. For example, if A is an $n \times n$ matrix with entries a^i_j then we can write trace $A = a^i_i$. Thus the identity matrix has entries δ^i_j and its trace is $\delta^i_i = n$. We shall use the same convention with more than one symbol present, for instance,

$$\begin{aligned} f_1\,dx^1 + f_2\,dx^2 + \cdots + f_n\,dx^n &= f_i\,dx^i \\ v^1\partial_1 + v^2\,\partial_2 + \cdots + v^n\,\partial_n &= v^i\,\partial_i \\ h^{11}\,\partial_1 \otimes \partial_1 + h^{12}\,\partial_1 \otimes \partial_2 + \cdots + h^{nn}\,\partial_n \otimes \partial_n & \\ = h^{ij}\,\partial_i \otimes \partial_j & \end{aligned}$$

D. Commutator or Lie Bracket

Given two tangent vector fields u, v with expressions $u = u^i\,\partial_i$ and $v = v^j\,\partial_j$ in terms of local partial derivatives defined by some chart (U, Φ), then we obtain their commutator vector field

$$[u, v]\colon U \to TM : x \mapsto \left(u^i \partial_i v^j - v^i \partial_i u^j\right)\partial_j$$

Independently of any chart we can define the commutator through its action as a derivation on any smooth real function $f\colon M \to \mathbb{R}$ by

$$[u, v]\,(f) = u(v(f)) - v(u(f))$$

A feel for what the commutator measures can be obtained by examining the flows of u and v. Suppose that these flows are, respectively, Φ and Ψ with

$$\begin{aligned} \Phi_t\colon U &\to M : x \mapsto \Phi(x, t) \\ \Psi_s\colon U &\to M : x \mapsto \Psi(x, s) \end{aligned}$$

Then it turns out that on U

$$\begin{aligned} \Phi_i \circ \Psi_s &= \Psi_s \circ \Phi_t \\ \text{if and only if } [u, v] &= \mathbf{0} \end{aligned}$$

Thus, $[u, v]$ measures the failure of the two flows to commute. It follows all vector fields v satisfy the self commuting property

$$[v, v] = \mathbf{0}$$

then the flow Φ of V reflects this in the identity $\Phi_t \circ \Phi_s = \Phi_s \circ \Phi_t = \Phi_{s+t}$ wherever all are defined on U.

E. Products of Fields

We have seen how $\otimes$ and $\oplus$ can be used in vector spaces, and it is easy to see how these operations can be performed smoothly over a manifold, so giving corresponding products on fields. Locally with respect to some chart $\phi\colon U \to E^n$, suppose that we have basis fields $\{\partial_1, \ldots, \partial_n\}$ for tangent and $\{dx^1, \ldots, dx^n\}$ for cotangent vector fields. Then we obtain, for example, the basis fields

$$\begin{aligned} &\{\partial_i \otimes \partial_j \mid i, j = 1, \ldots, n\} && \text{for } T^2_0 U \\ &\{dx^i \otimes dx^j \mid i, j = 1, \ldots, n\} && \text{for } T^0_2 U \\ &\{dx^i \wedge dx^j \mid 1 \leqq i < j \leqq n\} && \text{for } \Lambda^2 U \\ &\left\{dx^{i_1} \wedge dx^{i_2} \wedge \ldots \wedge dx^{i_r} \mid \right. \\ &\quad \left. 1 \leqq i_1 < i_2 < \ldots < i_r \leqq n\right\} && \text{for } \Lambda^r U \end{aligned}$$

To accommodate the structure of such spaces as $\Lambda^r U$ we modify the summation convention by enclosing the lower set of repeated indices in brackets to indicate that the summation is only over increasing sequences. Thus, for example, we would write for $w \in \Lambda^2 E^3$

$$\begin{aligned} w &= w_{12}\,dx^1 \wedge dx^2 + w_{13}\,dx^1 \wedge dx^3 + w_{23}\,dx^2 \wedge dx^3 \\ &= w_{(ij)}\,dx^i \wedge dx^j \end{aligned}$$

EXAMPLE. Consider the simple case $M = E^2$. Then we can illustrate two important fields with respect to the standard chart.

(1) $g\colon M \to T^0_2 M : (x_1, x_2) \mapsto \delta_{ij}\,dx^i \otimes dx^j$,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

This defines the usual metric structure on E^2, that is, the standard dot product on each tangent space $T_x M$:

$$\begin{aligned} g\left(u^k\,\partial_k, v^m\,\partial_m\right) &= \delta_{ij}\,dx^i \otimes dx^j\left(u^k\,\partial_k, v^m\,\partial_m\right) \\ &= \delta_{ij}\,u^k v^m\,dx^i \otimes dx^j(\partial_k, \partial_m) \\ &= \delta_{ij}\,u^k v^m\,\delta^i_k\,\delta^j_m \\ &= \delta_{ij}\,u^i v^j = u^1 v^1 + u^2 v^2 \end{aligned}$$

(2) $\omega\colon M \to \Lambda^2 M : (x_1, x_2) \mapsto dx^1 \wedge dx^2$

This defines the usual geometrical measure on E^2, that is, the standard parallelogram area mapped out by pairs of vectors in each tangent space $T_x M$:

$$\begin{aligned} \omega\left(u^k\,\partial_k, v^m\,\partial_m\right) &= \tfrac{1}{2}(dx^1 \otimes dx^2 - dx^2 \otimes dx^1) \\ &\quad \times\left(u^k\,\partial_k, v^m\,\partial_m\right) \\ &= \tfrac{1}{2}(u^1 v^2 - u^2 v^1) \end{aligned}$$

Both (1) and (2) generalize to E^n: g has the same form but we sum over $i, j = 1, \ldots, n$; $\omega = dx^1 \wedge \ldots \wedge dx^n$. Any change of chart induces corresponding changes in their local expressions.

F. Exterior Derivative

There is an extraordinarily useful derivative on differential forms defined locally by

$$\begin{aligned} d\colon \Lambda^r M &\to \Lambda^{r+1} M \\ : w_{(i_1 \ldots i_r)} dx^{i_1} \wedge \ldots \wedge dx^{i_r} &\mapsto dw_{(i_1 \ldots i_r)} \\ & dx^{i_1} \wedge \ldots \wedge dx^{i_r} \end{aligned}$$

where $dw_{i_1 \ldots i_r} = \partial_j w_{i_1 \ldots i_r}\, dx^j$, the gradient of the component function. Surprisingly, this exterior derivative d is uniquely determined by the requirements:

a. $d\colon \Lambda^r M \to \Lambda^{r+1} M$ linearly
b. $d\colon \Lambda^0 M \to \Lambda^1 M : f \mapsto df$
c. $w \in \Lambda^r M$, $v \in \Lambda^s M$
$\Rightarrow d(w \wedge v) = dw \wedge v + (-1)^r w \wedge dv$
d. $d^2 = 0$

Here of course we are using $\Lambda^r M$ to denote the space of sections of the bundle.

EXAMPLE. For a real function $f \in \Lambda^0 M$, with respect to some chart

$$\begin{aligned} df &= \partial_i f\, dx^i \in \Lambda^1 M \\ d^2 f &= d(\partial_i f) \wedge dx^i \\ &= \partial_j \partial_i f\, dx^j \wedge dx^i \\ &= 0 \quad \text{since } \partial_j \partial_i f = \partial_i \partial_j r \end{aligned}$$

It is precisely $d^2 = 0$ when applied to $M = E^3$ that gives the familiar identities of ordinary vector calculus:

$$\begin{aligned} \text{curl grad } f &= 0 \\ \text{div curl } v &= 0 \end{aligned}$$

Now, we know that on an n-manifold M, $\Upsilon^* M = A^1 M$ is modeled on $(\mathbb{R}^n)^*$, which has dimension n. Hence, for $r > n$, $\Lambda^r M$ contains only zero forms. Accordingly, the exterior derivative determines a chain complex of the spaces of forms viewed as modules over real functions:

$$\begin{array}{ccccccccccc} 0 \to & \Lambda^0 M & \xrightarrow{d} & \Lambda^1 M & \xrightarrow{d} & \Lambda^2 M \to \cdots \to & \Lambda^n M & \xrightarrow{d} & 0 \\ & \| & & \| & & & & & \\ & M \times \mathbb{R} & & \Upsilon_1^0 M & & & & & \end{array}$$

We always denote the zero module by 0. Properly, we should put indices on each d to indicate its domain, but often they are omitted. At each stage in all such sequences of structure-preserving maps we have a standard terminology, which can be illustrated in Fig. 4. In $\Lambda^r M$ we have two distinguished substructures:

1. r-forms sent to zero in $\Lambda^{r+1} M$, the *kernel* of

$$d\colon \Lambda^r M \to \Lambda^{r+1} M$$

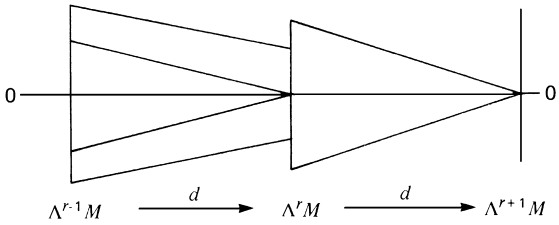


FIGURE 4 Chain complex of modules.

2. r-forms arriving from $\Lambda^{r-1} M$, the *image* of

$$d\colon \Lambda^{r-1} M \to \Lambda^r M$$

We usually abbreviate those to $\ker d$ and $\text{im } d$, and if we wish to emphasize precisely where they come from we write

$$\ker d = Z^r(M, d) \quad \text{and} \quad \text{im } d = B^r(M, d)$$

Note that it is a consequence of $d^2 = 0$ that

$$B^r(M, d) \subseteq Z^r(M, d)$$

and a consequence of the linearity of d that $B^r(M, d)$ is a submodule of $Z^r(M, d)$ and both are submodules of $\Lambda^r M$. When equality occurs we say that the sequence is exact at $\Lambda^r M$; this means that

$$\text{if } dw = 0 \text{ then } w = dv \text{ for some } v \in \Lambda^{r-1} M$$

There is a standard way to compare submodules of a module; just as for subgroups of groups and linear subspaces of vector spaces, we can take the quotient

$$H^r(M, d) = Z^r(M, d)/B^r(M, d)$$

We call $H^r(M, d)$ the rth de Rham cohomology module of M. Amazingly, it gives information about the connectivity of M by indicating the presence of $(r+1)$-dimensional holes.

EXAMPLE. To get a feel for the quotienting process, suppose that $\Lambda^r M$ looks like $\mathbb{R}^4$ and that

$$Z^r(M, d) \text{ is the 3-dimensional subspace } \mathbb{R}^3$$

$$B^r(M, d) \text{ is the } (x, y) \text{ plane in } Z^r(M, d)$$

Then

$$\begin{aligned} H^r(M, d) &= Z^r(M, d)/B^r(M, d) \\ &= \{[(x, y, z)]\} \end{aligned}$$

where $[(x, y, z)] = \{(x, y, z) \in Z^r(M, d) \mid (x, y) \in B^r(M, d)\}$. Hence $H^r(M, d)$ has one point, $[(x, y, z)]$, for each value of z and therefore is isomorphic to $\mathbb{R}^1$. Geometrically we view the quotient of $\mathbb{R}^3$ by its (x, y) subspace as the space consisting of all planes parallel to the (x, y) plane. That is, we stop distinguishing between points in the same horizontal plane; the set of such planes is isomorphic to $\mathbb{R}^1$, the z axis in $\mathbb{R}^3$.

G. Cohomology and Exactness

There are some special cases for an n-manifold M:

a. $B^0(M,d)=0$, so $H^0(M,d)\simeq Z^0(M,d)\simeq\mathbb{R}^1$ (constant f) if M is connected.
b. $Z^n(M,d)=0$, so $H^n(M,d)=0$.
c. If $Z^r(M,d)=B^r(M,d)$ then $H^r(M,d)=0$.

The cohomology module $H^r(M,d)$ measures departure from exactness of d at $\Lambda^r M$. There are traditional names for elements:

$Z^r(M,d)$ consists of *closed* r-forms, w with $dw=0$
$B^r(M,d)$ consists of *exact* r-forms, w with $w=dv$

From the quotient process it turns out that

$$u\in H^r(M,d)\Rightarrow u=w+B^r(M,d)$$

for some

$$w\in Z^r(M,d)$$

that is: two equivalent representatives of u in $Z^r(M,d)$ differ only by some $dv\in B^r(M,d)$.

A famous result of Poincaré is the following:

$$\text{if } M\simeq E^n \text{ then } H^k(M,d)=0 \text{ for all } k>0$$

Intuitively, there are no holes in Euclidean space or in any manifold that is diffeomorphic to it.

The dimension of $H^r(M,d)$ is called the rth Betti number of M; it is a topological invariant, so two manifolds that are homeomorphic (not necessarily diffeomorphic) have the same Betti numbers. The Betti numbers are finite for compact manifolds M and then they determine the Euler characteristic

$$\chi(M)=\sum_{r=0}^{\dim M}(-1)^r\dim H^r(M,d)$$

which, like the Betti numbers, can be computed by algebraic topological methods.

EXAMPLE. Take $M=S^1$, the unit circle. This is connected, so we find $H^0(S^1,d)\simeq\mathbb{R}^1$, and it is one-dimensional, so $H^r(S^1,d)=0$ for $r\geqq 2$. We find $H^1(S^1,d)$; intuitively, it should not be zero because S^1 actually surrounds a two-dimensional hole.

$$0\to\Lambda^0S^1\xrightarrow{d}\Lambda^1S^1\to 0$$

$$H^1(S^1,d)=Z^1(S^1,d)/B^1(S^1,d)$$

But $Z^1(S^1,d)=\Lambda^1S^1=\{\text{smooth } w\colon S^1\to A^1S^1=T^*S^1\}$ and $B^1(S^1,d)=\{df\mid f\in\Lambda^0S^1=\text{smoot maps } S^1\to\mathbb{R}\}$. Hence

$$H^1(S^1,d)=\{w+B^1(S^1,d)\mid w\in\Lambda^1S^1\}$$

Now Λ^1S^1 is generated by any nowhere-zero section of A^1S^1. Denote by δ^1 the element of TS^1 defined by using an angle θ for coordinate on S^1. (Properly our coordinates should take values in $\mathbb{R}^1$ but we identify 0 and 2π to save using two charts explicitly.) Let ω be the dual of δ_1; then ω is nowhere zero because δ_1 is nowhere zero. Also, by definition, for any $h\partial_1\in\Upsilon S^1$

$$\omega(h\partial_1)=h\quad\text{since }\omega(\partial_1)=1$$

Hence, any 1-form v on S^1 is given by

$$v\colon S^1\to A^1S^1:\theta\mapsto\lambda(\theta)\omega$$

for some real function λ.

Now ω is not exact (though a common notation for it is $d\theta$) since it is not the differential of any smooth real function. For suppose $\omega=df$, then we require $df(\partial_1)=1$ and hence

$$f\colon S^1\to\mathbb{R}:\theta\mapsto f(\theta)$$

must be continuous and smooth with $df/d\theta=1$ everywhere. Apparently $f(\theta)=\theta$ would satisfy this differential equation, but the problem is $f(0)\neq f(2\pi)$ so it is not continuous. Therefore $\omega+B^1(S^1,d)$ generates $H^1(S^1,d)$ and so

$$H^1(S^1,d)\simeq\mathbb{R}^1$$

which has dimension 1, corresponding to one 2-hole in S^1.

H. Derivatives of Smooth Maps

A smooth map $f\colon M\to M'$ between two manifolds preserves smooth things. Now, since the prototypes of smooth things are the tangent structures, it is not surprising that such a map induces a smooth map, the derivative of f between the tangent bundles; we denote it by $Df\colon TM\to TM'$, though sometimes Df is written Tf. This smooth map is actually between two fiber bundles so, as we would wish, it preserves fibers; moreover, since tangent bundles are vector bundles it is actually linear between tangent spaces. Hence at each $x\in M$, Df restricted to T_xM is a linear map

$$\begin{array}{ccc} D_xf\colon T_xM & \longrightarrow & T_{f(x)}M' \\ \simeq & & \simeq \\ \mathbb{R}^n & \xrightarrow[[f_{ij}]]{} & \mathbb{R}^m \end{array}$$

which, on choosing charts about x and $f(x)$, appears locally as the Jacobian matrix $[f_{ij}]$ of partial derivatives of components of f with respect to coordinates at x. As always, there is a dual to the linear map D_xF, going the other way:

$$\begin{array}{ccc} D_x F^*\colon T_{f(x)}M^* & \longrightarrow & T_x M^* \\ \simeq & & \simeq \\ \mathbb{R}^m & \xrightarrow[[f_{ij}]^t]{} & \mathbb{R}^n \end{array}$$

and its corresponding matrix representative is just the transpose of that for $D_x f$. As with the $D_x f$, their duals fit together smoothly to give a vector bundle map $(Df)^*\colon T^*M' \to T^*M$.

From Df and $(Df)^*$ it is easy to obtain induced vector bundle maps for all of the tensor bundles. So

$$\text{smooth } f\colon M \to M' \Rightarrow \text{smooth } \begin{cases} D_s^r f\colon T_s^r M \to T_s^r M' \\ \Lambda^r f\colon A^r M' \to A^r M \end{cases}$$

We often abbreviate $D_s^r f$ to f_* and $A^r f$ to f^* when we apply them to sections of these bundles.

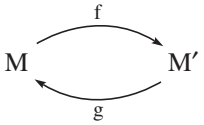
It is of fundamental importance that

$$f^*\colon \Lambda^r M' \to \Lambda^r M \qquad \begin{array}{l}\text{preserves closedness and} \\ \text{exactness of } r\text{-forms}\end{array}$$

In consequence f^* induces a linear map on cohomology:

$$H^r(f^*)\colon H^r(M', d) \to H^r(M, d)$$

Now certainly if f has a smooth inverse (i.e., f is a diffeomorphism) then each $D_x f$ and $(D_x f)^*$ would be an isomorphism and we would expect f^* to be an isomorphism of modules giving an isomorphism in cohomology. In fact, $H^r(f^*)$ is an isomorphism if f is any "smooth deformation" of a diffeomorphism. Precisely, M and M' will have the same cohomology if they have the same homotopy type, that is, if there are smooth maps such that $g \circ f$ can be smoothly deformed into (is *homotopic to*) the identity 1_M and $f \circ g$ is similarly homotopic to $1_{M'}$. The allowed smooth deformations are contracting or stretching, but without cutting or joining or introducing any holes or corners.



Having the same cohomology obviously means having the same Betti numbers.

EXAMPLE. S^1 and the punctured plane $\mathbb{R}^2\backslash\{(0,0)\}$ are of the same homotopy type because we have

$$f\colon S^1 \to \mathbb{R}^2\backslash\{(0,0)\} : x \mapsto x$$

$$g\colon \mathbb{R}^2\backslash\{(0,0)\} \to S^1 : x \mapsto x/\|x\| \quad \text{(usual norm)}$$

Then $g \circ f : S^1 \to S^1 : x \mapsto x$ is the identity. Also

$$f \circ g\colon \mathbb{R}^2\backslash\{(0,0)\} \to \mathbb{R}^2\backslash\{(0,0)\} : x \mapsto x/\|x\|$$

is a smooth deformation of the identity by the homotopy

$$F_t : \mathbb{R}^2\backslash\{(0,0)\} \to \mathbb{R}^2\backslash\{(0,0)\}$$
$$: x \mapsto (1-t)x + tx/\|x\|$$

Evidently, F_0 is the identity on $\mathbb{R}^2\backslash\{(0,0)\}$, F_1 is $f \circ g$, and F_t varies smoothly with t in the interval $[0, 1]$. In fact, the same homotopy holds good for the same map between S^n and $\mathbb{R}^{n+1}\backslash\{(0,0)\}$ ($\mathbb{R}^{n+1}$ with the origin removed). We conclude that for $0 \leqq r \leqq n$

$$H^r(S^n, d) \simeq H^r(\mathbb{R}^{n+1}\backslash\{(0,0)\}, d) \quad \text{for } n \geqq 1$$

I. Using Cohomology Data

We saw previously that $H^1(S^1, d) \simeq \mathbb{R}^1$, since there is a closed 1-form that is not exact; we now know that there must also be a closed 1-form that is not exact on the punctured plane. Equivalently, there is a smooth vector field that is not the gradient of any function, for example, rotation about the origin with velocity proportional to distance from the origin.

Table I gives the nonzero de Rham cohomology of some familiar spaces. Each entry indicates the dimension of the corresponding module of nonexact closed forms. Recall that we agreed to allow manifolds with boundary.

We can immediately infer the following, for example:

1. On a torus $S^1 \times S^1$ we can find $u, v \in \Lambda^1(S^1 \times S^1)$ such that u and v are independently directed cotangent fields and neither is a gradient field; moreover, there exists $w \in \Lambda^2(S^1 \times S^1)$ that is not

TABLE I Betti Number of Some Familiar Spaces

Space M	Dimension of $H^r(M, d)$ (Betti numbers)				
	$r=0$	$r=1$	$r=2$	$\cdots$	$r=n$
Circles S'	1	1			
Torus $S' \times S'$	1	2	1		
Sphere S^2	1	0	1		
n-sphere S^n	1	0	0	$\cdots$	1
2-sphere $+ m$ handles	1	$2m$	1		
Projective plane $\mathbb{R}P^2$	1				
$\mathbb{R}P^n$ (n even)	1				
$\mathbb{R}P^n$ (n odd)	1	0	0	$\cdots$	1
$\mathbb{R}^n$	1				
Punctured torus	1				
Klein bottle	1	1			
2-sphere $+ m$ discs replaced by Möbius strips	1	$m-1$			
n-ball B^n	1				
Punctured n-ball $B^n\backslash\{0\}$	1	0 $\cdots$			1

expressible as $\lambda\,du + \mu\,dv$ for any real functions λ, μ.

2. On a sphere S^2 there is no nongradient cotangent field: $u \in \Lambda^1 S^2 \Rightarrow u = df$ for some $f\colon S^2 \to \mathbb{R}$; also, there is a nonzero $w \in \Lambda^2 S^2$, which is consequently not of the form du for $u \in \Lambda^1 S^2$ since all such have $u = df$ so $du = d^2 f = 0$.
3. If M is $\mathbb{R}^n$ (or the open n-ball, $\{x \in \mathbb{R}^n \mid \|x\| < 1\}$, to which it is diffeomorphic) then every closed r-form $w \in \Lambda^r M$ is expressible as $w = du$ for some $u \in \Lambda^{r-1} M$; hence the partial differential equation $dw = 0$ has a solution.

V. METRIC GEOMETRY

In an n-manifold M we often have to measure two kinds of things: the "volume" of a piece of M or the "size" of vectors from some bundle over M. The special kind of measure function appropriate for a manifold is a volume form.

A. Volume Forms

A volume form on an m-manifold M is a nowhere-zero n-form $\mu \in \Lambda^n M$. Since $A^n M$ is a vector bundle [with fiber $(\mathbb{R}^n)^*$] there may be no such volume form.

On E^n with the standard coordinates there is a volume form namely $dx^1 \wedge dx^2 \wedge \ldots \wedge dx^n$, as we have pointed out before. Now every point of M has a neighborhood that is diffeomorphic to E^n so in each such neighborhood we could construct a local volume form by pulling it back from E^n; the problem is simply that we may not be able smoothly to join up these local expressions into a global volume form on M. A sufficient condition for effecting the joining up is easily motivated:

If (U, ϕ), (V, ψ) are two charts about $x \in U \cap V$ with coordinates $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, respectively, then

$$dx^1 \wedge \ldots \wedge dx^n = J_{\phi\psi}\, dy^1 \wedge \ldots \wedge dy^n$$

where $J_{\phi\psi}$ is the Jacobian determinant of the change of coordinates

$$J_{\phi\psi} = \det(\partial x_i / \partial y_j)$$

Plainly, a good start is if $J_{\phi\psi}$ is an everywhere positive function for enough overlapping charts to cover M; if it has this property then M is called *orientable*. In fact, being orientable is necessary and sufficient for M to admit a volume form, since from the outset we assumed that our manifolds have a Hausdorff topology with a countable base, the latter being just the property needed to ensure the proper local joining of local volume forms. Actually, integration can also be performed on manifolds that are not orientable by passing to density forms, but we shall not pursue this.

If μ is a volume form on M then so is $f\mu$ for any nowhere-zero smooth real function f, so such an f must stay the same sign on each component piece of M. Since $A^n M$ can have only one independent section, then volume forms can only differ by such a factor f. We say μ and $f\mu$ are equivalent choices of orientation if f is everywhere positive and opposite choices if f is everywhere negative.

Given a volume form μ on M and an atlas $\{(U_\alpha, \phi_\alpha) \mid \alpha \in A\}$ then it is possible to define an integral of an n-form $w \in \Lambda^n M$ by means of a technical device called a *partition of unity*. This effectively shares w out fairly among the charts involved in any overlapping. Then each ϕ_α^* carries the appropriately weighted fraction of w to $\Lambda^n E^n$, where integration is as usual, and it remains to sum up the various weighted contributions from overlapping charts. For this to give a finite answer it is necessary that the sum converge, and for this we require M to have finite volume or, if not, then w must become zero on all but a finite region of M. The technical condition is that w has compact support. The integral of w is denoted $\int_M w$ and in particular the volume of M is $\int_M \mu$.

A fundamental application of this process is in Stokes' theorem:

$$\int_M dv = \int_{\partial M} v$$

for $v \in \Lambda^{n-1} M$ with compact support

where ∂M is the boundary of an n-manifold with boundary. In this general context, integration by parts has the expression

$$\int_M df \wedge v = \int_{dM} fv - \int_M f\, dv$$

where $f \in \Lambda^0 M$.

EXAMPLE. Let M be a closed, simply connected region in E^2. Take $v \in \Lambda^1 M$ to be given in standard coordinates by

$$v = \tfrac{1}{2}(x\,dy - y\,dx)$$

Then

$$\begin{aligned} dv &= \tfrac{1}{2}(dx \wedge dy - dy \wedge dx) \\ &= dx \wedge dy \\ \int_M dv = \int_{\partial M} v \Rightarrow \int_M dx \wedge dy &= \frac{1}{2}\int_{\partial M} (x\,dy - y\,dx) \end{aligned}$$

But $dx \wedge dy$ is the usual volume form for E^2 so $\int_M dx \wedge dy$ is just the area enclosed by curve ∂M. In particular, if ∂M is the ellipse L with

$$L = \{(x = a\cos\theta,\, y = b\sin\theta) \in \mathbb{E}^2 \mid 0 \leqq \theta \leqq 2\pi\}$$

then

$$\begin{aligned} x\,dy - y\,dx &= (ab\cos^2\theta + ba\sin^2\theta)\,d\theta \\ &= ab\,d\theta \end{aligned}$$

and it follows that the area of the ellipse is

$$\int_M dx \wedge dy = \frac{1}{2}\int_{\partial M} ab\,d\theta = \pi ab$$

Evidently the volume form on E^3 is $dx \wedge dy \wedge dz$ and its restriction to the two-dimensional submanifold E^2 which is the (x, y) plane is simply $dx \wedge dy$. Now, the ellipse itself is a one-dimensional submanifold of E^2 and, like the circle, it supports a nowhere-zero 1-form $d\theta$. The standard volume form on the ellipse L is actually $r\,d\theta$, where $r^2 = x^2 + y^2$, and therefore the circumference of the ellipse is

$$\begin{aligned} \int_L r\,d\theta &= \int_0^{2\pi} (a^2\cos^2\theta + b^2\sin^2\theta)^{1/2}\,d\theta \\ &= a\int_0^{2\pi} (1 - e^2\sin^2\theta)^{1/2}\,d\theta \end{aligned}$$

the familiar elliptic integral with e the eccentricity $(1 - b^2/a^2)^{1/2}$.

B. Metric Tensors

A metric tensor on an n-manifold M is an element $g \in T_2^0 M$, that is, a section of $T_2^0 M$

$$g\colon M \to T_2^0 M : x \mapsto g_x \in T_x M^* \otimes T_x M^*$$

where $g_x\colon T_x M \times T_x M \to \mathbb{R} : (u, v) \mapsto g_x(u, v)$ satisfies the conditions

1. Symmetry: $g_x(u, v) = g_x(v, u)$.
2. Nondegeneracy: if $g_x(u, v) = 0$ for all $v \in T_x M$ then $u = 0$.

If $(\partial_1, \ldots, \partial_n)_x$ is a basis for $T_x M$ then its dual, $(dx^1, \ldots, dx^n)$, is a basis for $(T_x M)^*$ and so g_x is expressible in the form

$$g_x = \left(g_{ij}\,dx^i \otimes dx^j\right)_x$$

Then the symmetry condition imposes symmetry on the matrix of numbers $(g_{ij})_x$ and nondegeneracy makes its determinant nonzero.

Usually we drop the subscript x indicating the variability of g and its coordinate expression with position in M.

EXAMPLE. Take $M = E^2$ so each tangent space is isomorphic to $\mathbb{R}^2$ as the set of directions at the point. The simplest choice of metric tensor is given by the usual inner product

$$\begin{aligned} g_x\colon T_x M \times T_x M &\to \mathbb{R} \\ : \left(u^i\,\partial_i, v^j\,\partial_j\right) &\mapsto u^1 v^1 + u^2 v^2 \end{aligned}$$

and so here $g = \delta_{ij}\,dx^i \otimes dx^j$. This induces on E^2 all of the usual Euclidean geometry, including the usual volume form; it is easily generalized to E^n.

Another metric tensor on E^2 is given by

$$\begin{aligned} \eta_x\colon T_x M \times T_x M &\to \mathbb{R} \\ : \left(u^i\,\partial_i, v^j\,\partial_j\right) &\mapsto u^1 v^1 - u^2 v^2 \end{aligned}$$

and so this is expressed as

$$\eta = \eta_{ij}\,dx^i \otimes dx^j \quad \text{where } \eta_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

This induces Minkowski geometry on E^2, as used in relativity.

From the example we see that the Euclidean metric tensor satisfies a stronger condition than 2. It is $2'$. Positive definiteness: $g_x(u, v) = 0$ if and only if $u = 0$. This imposes on the matrix $(g_{ij})_x$ that its eigenvalues all be of one sign.

A metric tensor satisfying condition $2'$ is called a *Riemannian metric*; one satisfying only 2 is called an *indefinite metric* or a *pseudo-Riemannian metric*. Riemannian metric geometry includes all of the usual curved surfaces and much else; pseudo-Riemannian geometry includes space–time structures where angles and lengths are interpreted very differently from those of common experience.

Since the simplest possible choice of metric tensor gives the whole of standard Euclidean geometry on E^n, we can expect that the structure implied by any metric tensor is very rich indeed. This is the case, but before investigating it we should consider existence questions. Now, an n-manifold consists of copies of E^n smoothly joined together and we have a standard metric tensor on each patch of E^n, so we are once again faced with patching them together in a consistent way. The same trick works as for volume forms: we use a partition of unity to share out fairly the contributions of each overlapping copy of E^n. Thus, all of our manifolds admit a Riemannian metric. However, whereas the module $\Lambda^n M$ of volume forms is one-dimensional and so any two differed only by a scalar function, the module $\Upsilon_2^0 M$ of sections of $T_2^0 M$ is n^2-dimensional and no particular Riemannian metric is

distinguished in general. Of course, if M is itself a submanifold of some manifold M (especially E^m for some $m \geqq n$) that has a metric tensor g, then the restriction of g to M is a metric tensor on M. This is an important source of examples.

We list next some of the important implications of an n-manifold M having a metric tensor g.

1. It induces a measure of length for a curve
$$c\colon [0,1] \to M : t \mapsto c(t)$$
The tangent vector to c at t is $\dot{c}(t) \in T_{c(t)}M$ and we define the length of c to be
$$\int_{t=0}^{t=1} |g(\dot{c}(t), \dot{c}(t))|^{1/2}\, dt$$
In coordinates, if $c(t) \in U$ for some chart (U, ϕ) then $\phi(c(t)) = (c^1(t), \ldots, c^n(t))$ and
$$\dot{c}(t) = \dot{c}^i(t)\, \partial_i \quad \text{with } \dot{c}^i = dc^i/dt.$$
Hence $g(\dot{c}, \dot{c}) = g_{ij}\dot{c}^i\dot{c}^j$.
2. There is a dual metric tensor $g^* \in \Upsilon_0^2 M$ with
$$g_x^*\colon T_xM^* \times T_xM^* \to \mathbb{R} : (\alpha, \beta) \mapsto g_x^*(\alpha, \beta)$$
with coordinate expression
$$g_x^* = \left(g^{ij}\, \partial_i \otimes \partial_j\right)_x \quad \text{and} \quad (g^{ij}) = (g_{ij})^{-1}$$
as matrices, well defined since $\det g_{ij} \neq 0$.
3. There is an isomorphism $\Upsilon_1^0 M \simeq \Upsilon_0^1 M$ of tangent and cotangent fields:
$$g_x^b\colon T_xM \to T_xM^* : v \mapsto v^*$$
$$\text{with } v^*(u = g_x(u, v))$$
which appears in coordinates as
$$g_x^b\colon v^i\, \partial_i \mapsto g_{ij}v^i\, dx^j$$
with inverse
$$g_x^{\#}\colon w_i\, dx^i \mapsto g^{ij}w_i\, \partial_j$$
4. These isomorphisms in 2 can be tensor-producted to give isomorphisms
$$\Upsilon_s^r M \simeq \Upsilon_m^k M \quad \text{for all } r + s = k + m$$
Consequently, we can switch among tensor fields to suit our convenience when we have a metric tensor.
5. If M is an orientable manifold then a choice of an oriented atlas (Jacobian everywhere positive on overlaps) yields a unique volume form μ_g determined locally by
$$\mu_g = |\det g_{ij}|^{1/2}\, dx^1 \wedge \cdots \wedge dx^n$$
This is nowhere zero since nondegeneracy ensures $\det g_{ij} \neq 0$. Observe that this does agree with the standard volume form on E^n, where $\det g_{ij} = 1$ everywhere.
6. If M is oriented then there is a Hodge dual isomorphism on differential forms:
$$^*\colon \Lambda^r M \to \Lambda^{n-r} M : w \mapsto {}^*w$$
To explain the construction of this isomorphism suppose that $(e^1, \ldots, e^n)_x$ is any ordered basis in the orientation for T_xM^* and ordered bases for $\Lambda^r M$ and $\Lambda^{n-r}M$ are given locally by
$$\{e^{i_1} \wedge e^{i_2} \cdots e^{i_r} \mid 1 \leqq i_1 < \cdots < i_r \leqq n\}$$
$$\{e^{j_1} \wedge e^{j_2} \cdots e^{j_{n-r}} \mid 1 \leqq j_1 < \cdots < j_{n-r} \leqq n\}$$
Then with $g_{ij} = g(e^i, e^j)$
$$^*w_{(i_1 \cdots i_r)} e^{i_1} \wedge \cdots \wedge e^{i_r} = |\det g_{ij}|^{1/2} w^*_{(j_1 \cdots j_{n-r})} e^{j_1} \wedge \cdots \wedge e^{j_{n-r}}$$
where
$$w^*_{j_1 \cdots j_{n-r}} = g^{k_1 i_1} g^{k_2 i_2} \cdots g^{k_r i_r} w_{i_1 \cdots i_r}\, \mathrm{sgn}(k \to i)$$
and $\mathrm{sgn}(k \to i)$ is the sign of the permutation
$$(i_1, \ldots, i_r, j_1, \ldots, j_{n-r}) \mapsto (k_1, \ldots, k_r, j_1, \ldots, j_{n-r})$$
Intuitively, ${}^*w$ is the "complement" of w in the volume form μ_g determined by g; in particular, we find
$$^*1 = \mu_g \qquad (1 \text{ is the constant unit real function on } M,\ 1 \in \Lambda^0 M)$$
$$^*\mu_g = (-1)^\nu \qquad (\nu \text{ is the number of negative eigenvalues of } g)$$
The appearance of * is simplest if we choose the oriented base $(e^1, \ldots, e^n)$ for T_xM^* to be orthonormal, that is, $g^*(e^i, e^j) = \pm\delta^i_j$ so $|\det g_{ij}| = 1$. In this case
$$^*(e^{i_1} \wedge \cdots \wedge e^{i_r}) = (e^{j_1} \wedge \cdots \wedge e^{j_{n-r}})$$
for any even permutation $(i_1, \ldots, i_r, j_1, \ldots, j_{n-r})$ of $(1, \ldots, n)$, and we extend the isomorphism to all r-forms by requiring it to be linear.

Some justification for the presence of $|\det g_{ij}|$ in the volume form can be seen by recalling that (g_{ij}) is the local $(n \times n)$ matrix expression of a linear map $T_xM \to T_xM^*$. For such matrix maps from $\mathbb{R}^n$ to $\mathbb{R}^n$ the determinant measures precisely the factor of volume change under the map: a unit n-cube is sent to an n-box of volume $|\det g_{ij}|$.

EXAMPLE. The equations of the elecromagnetic field on a space-time manifold can be very neatly expressed in terms of the electromagnetic 2-form $F \in \Lambda^2 M$ and $\dim M = 4$.

Locally, for a basis of 1-form fields (ω^i),

$$F = F_{(ij)}\,\omega^i \wedge \omega^j$$

If we suppose that the ω^i are mutually orthogonal unit fields, then the metric tensor components (g_{ij}) are the eigenvalues of g lying along the diagonal.

The Hodge dual isomorphism gives ${}^*(\omega^i \wedge \omega^j) = \omega^m \wedge \omega^k$, where (i, j, m, k) is an even permutation of $(1, 2, 3, 4)$. So ${}^*(\omega^1 \wedge \omega^2) = \omega^3 \wedge \omega^4$ and so forth. Similarly, ${}^*(\omega^1 \wedge \omega^2 \wedge \omega^3) = \omega^4$ and so forth.

Physical theory leads to the following equations for F in regions that contain negligible amounts of matter:

$$dF = 0 \quad {}^*d^*F = J \qquad (\dagger)$$

where J is the current density. These equations correpond to the usual Maxwell's equations through the vector calculus correspondences:

$${}^*d \equiv \text{curl} \quad \text{and} \quad {}^*d^* \equiv \text{divergence}$$

Conservation of charge is expressed by

$$\text{div}\, J = 0$$

This is automatically satisfied when there is negligible matter since it becomes

$${}^*d^{**}d^*F = 0 \quad \text{because } d^2 = 0$$

However, in the presence of matter, Eq. (†) becomes

$$dA = 0 \qquad {}^*d^*B = J \quad \text{for some } A, B \in \Lambda^2 M$$

with A and B related by some transformation, perhaps linear. Once again, $d^2 = 0$ ensures conservation of charge.

Locally, $J = \rho J_i \omega^i$, where ρ is the chrage density. Then over a compact spacelike submanifold S of M we can measure the total charge Q_N and find that

$$\begin{aligned} Q_N &= \int_S \rho\omega^i \wedge \omega^2 \wedge \omega^3 = \int_S {}^*J \\ &= \int_S d^*B = \int_{\partial S} {}^*B \end{aligned}$$

VI. CONNECTION GEOMETRY

Given a tangent vector field v: $M \to TM$ it is quite likely that we shall be interested in measurings its rate of change over M. Now, v is a smooth map between smooth manifols so it induces the derivative

$$Dv\colon TM \to T(TM)$$

between the corresponding tangent bundles. Unfortunately, this is not useful as a measure of the rate of change of v on M because Dv takes values in $T(TM)$, not in TM. A connection gets us from $T(TM)$ to TM in orderly fashion.

A. Linear Connection

There are more general connections than the one that we shall use; we are interested in those that have particular significance for the structures we already have on M.

A (linear) connection on M is a splitting of the vector bundle $T(TM)$ into a direct sum of a horizontal part HM and a *vertical* part VM, with HM isomorphic to TM as vector bundles. The motivation is clear: given any vector $u \in TM$ and field v: $M \to TM$ then

$$Dv(u) \in T(TM) \simeq HM \oplus VM$$

and we interpret the rate of change of v in the direction u as the projection of the HM part onto TM. Equivalently, we define a connection on M to ba a map ∇ that assigns to each $u \in TM$ and $v \in \Upsilon_0^1 M$ a vector $\nabla_u v$ in $T_x M$ (where $u \in T_x M$) such that

a. ∇_u is linear over $u \in T_x M$.
b. $\nabla_u v$ is linear over $v \in \Upsilon_0^1 M$.
c. $\nabla_u f v = u(f)v(x) + f(x)\,\nabla_u v$ if f: $M \to \mathbb{R}$.
d. If $w, v \in \Upsilon_0^1 M$ then so is $\nabla_w v$: $x \mapsto \nabla_{w(x)} v$.

We view $\nabla_u v$ as the rate of change of the field v in the direction of the vector u at $x \in M$ and call it the covariant derivative of v with respect to u.

By exploiting the linearity properties we can easily obtain coordinate expressions. Given a chart (U, ϕ) with cordrinates $(x_1, \ldots, x_n)$ about x and basis fields $(\partial_1, \ldots, \partial_n)$ for tangent vector fields about x, any $u \in T_x M$ is of the form $u = (u^i \partial_i)_x$ and near x any vector field w is of the form $w = w^j\,\partial_j \in \Upsilon_0^1 U$. Now we must have

$$\begin{aligned} \nabla_u w &= \nabla_{u i_{\partial_i}} \left(w^j\,\partial_j\right) \in T_x M \\ &= u^i \nabla_{\partial_i} \left(w^j\,\partial_j\right) \quad \text{by a} \\ &= u^i \left(\left(\partial_i\, w^j\right)\partial_j + w^j\,\nabla_{\partial_i}\,\partial_j\right)_x \quad \text{by b and c} \end{aligned}$$

Hence ∇ is completely determined on U by specification of $\nabla_{\partial_i}\,\partial_j \in \mathrm{T}_0^1 U$. But any such field is of the form

$$\nabla_{\partial_i}\,\partial_j = \Gamma_{ij}^k\,\partial_k \quad \text{for some } \Gamma_{ij}^k\colon U \to \mathbb{R}$$

So locally ∇ appears as an n^3 array of smooth real functions on U. These functions are traditionally called the *Christoffel symbols* of the connection, and they will change smoothly from chart to chart. Substitution of $[(\partial y^m/\partial x^k)\,\hat{\partial}_m]$ for (∂_i) through a change of chart coordinates from $(x_1, \ldots, x_n)$ to $(y_1, \ldots, y_n)$ can be traced trough the above steps to obtain expressions for

$$\nabla_{\hat{\partial}_i}\,\hat{\partial}_j = \hat{\Gamma}_{ij}^k\,\hat{\partial}_k$$

Then the $\hat{\Gamma}^k_{ij}$ can be related to the Γ^k_{ij} through the partial derivatives $(\partial y^m/\partial x^k)$. This is one way to see that the Christoffel symbols are not components of any tensor field.

We easily extend ∇ to give covariant derivative of arbitrary tensor fields by defining for $u \in T_x M$

a. $\nabla_u f = u(f)$ for $f \in \Upsilon^0_0 M$.
b. $\nabla_u(dx^i \otimes \partial_i) = 0$ for mutually dual basis fields.
c. $\nabla_u(w \otimes v) = (\nabla_u w) \otimes v + w \otimes \nabla_u v$.

From b and c with $u = \partial_j$ we get, for example,

$$\nabla_{\partial_j}(dx^i \otimes \partial_i) = \left(\nabla_{\partial_j} dx^i\right) \otimes \partial_i + dx^i \otimes \nabla_{\partial_j} \partial_i = 0$$
$$0 = \nabla_{\partial_j} dx^k \otimes \partial_k + \Gamma^k_{ji}\, dx^i \otimes \partial_k$$

therefore

$$\nabla_{\partial_j} dx^k = -\Gamma^k_{ji}\, dx^i$$

This defines a family of 1-forms called the *connection 1-forms*.

B. Parallel Transport

Given a curve c: $[0, 1] \to M$, its tangent vector field is denoted $\dot{c}$: $[0, 1] \to TM$ and in components $\dot{c} = \dot{c}^j\, \partial_j$. If $w \in \Upsilon^r_s M$ then we say that w is parallel along c if

$$\nabla_{\dot{c}} w = 0$$

If $w = w^k\, \partial_k \in \Upsilon^1_0 M$ is parallel along c with tangent vector field $\dot{c}^j\, \partial_j = d/dt$, then

$$\nabla_{\dot{c}} w = \nabla_{\dot{c}\, j_{\partial_j}} w^k\, \partial_k = 0 \in \Upsilon^1_0 M$$

so $(\dot{c}^j \partial_j w^k)\, \partial_k + \dot{c}^j w^i\, \nabla_{\partial_j} \partial_i = 0$. Hence $dw^k/dt + (dc^j/dt) w^i\, \Gamma^k_{ji} = 0 \in \mathbb{R}, k = 1, 2, \ldots, n$. This system of n linear differential equations, which represents $\nabla_{\dot{c}} w = 0$, has a unique solution

$$w\colon [0, 1] \to TM\colon t \mapsto w_t$$

for given initial value $w(0) = w_0 \in T_{c(0)} M$. Therefore ∇ and c have defined a map

$$\tau_t\colon T_{c(0)} M \to T_{c(t)} M\colon w_0 \mapsto w_t$$

This turns out to be a very good map indeed: an isomorphism of vector spaces; it is called *parallel transport* and is available for tensor spaces also. Moreover, parallel transport allows us to express the covariant derivative as a limit of a difference, because with it we can bring all vectors and tensors along c back to $c(0)$. Let $u \in T_x M$ be any vector; $w \in \Upsilon^r_s M$ and c is any curve with $c(0) = x, \dot{c}(0) = u$; then

$$(\nabla_u w)_x = \lim_{t \to 0} \frac{\tau_t^{-1}(w(c(t))) - w(c(0))}{t}$$

EXAMPLE. Consider E^1 with $\nabla_{\partial_1} \partial_1 = \lambda$, for some constant $\lambda \in \mathbb{R}$, with respect to the standard chart.

Take c: $[0, 1] \to E^1$: $t \mapsto t$, so $\dot{c}(t) = \partial_1$.
Then τ_t: $T_{c(0)} E^1 \to T_{c(t)} E^1$: $\alpha_0 \partial_1 \mapsto \alpha(t) \partial_1$ satisfies.

$$\frac{d\alpha}{dt} + \alpha\lambda = 0 \quad \text{so } \alpha(t) = \alpha_0 e^{-\lambda t}$$

Evidently $\lambda = 0$ corresponds to the usual connection since we do not usually alter the length of vectors when we move them on E^1. Any $\lambda \neq 0$ determines a non-Euclidean parallelism structure on E^1. A similar connection could be put on S^1.

EXAMPLE. To find a local expression for the parallel transport isomorphism. We consider $M = E^2$ with the standard chart and connection ∇ having constant Christoffel symbols

$$\Gamma^1_{12} = \Gamma^1_{21} = 1 \quad \text{and all other components zero}$$

Given the curve c: $[0, 1] \to E^2$: $t \mapsto (t, t^2)$ we find the parallel vector field

$$w\colon [0, 1] \to TE^2\colon t \mapsto f(t)\, \partial_1 + g(t)\, \partial_2$$

for two independent initial tangent vectors:

a. $w(0) = \partial_1$
b. $w(0) = \partial_2$

The parallel transport condition is $\nabla_{\dot{c}} w = 0$ and we are given $\dot{c}(t) = \partial_1 + 2t\, \partial_2$. Substituting

$$\dot{f}\partial_1 + \dot{g}\, \partial_2 + 2tf\, \Gamma^i_{21}\, \partial_i + g\Gamma^i_{12}\, \partial_i = 0$$
$$(\dot{f} + 2tf + g)\, \partial_1 + \dot{g}\, \partial_2 = 0$$
$$\text{so } g(t) = g(0)$$

We solve $\dot{f} + 2tf + g = 0$ for constant g to give:

Case (a): $f(t) = e^{-t^2}, g(t) = g(0) = 0$.
Case (b): $f(t) = -e^{-t^2} \int_0^t e^{x^2}\, dx = k(t)$, say, and $g(t) = 1$.

Then parallel transport along c is the isomorphism

$$\tau_t :\ T_{c(0)} E^2 \to T_{c(t)} E^2$$
$$:\ \alpha\, \partial_1 + \beta\, \partial_2 \mapsto \left(\alpha e^{-t^2} + \beta k(t)\right) \partial_1 + \beta\, \partial_2$$

or in matrix form

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \mapsto \begin{pmatrix} e^{-t^2} & k(t) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Evidently ∇ is not compatible with the usual metric tensor on E^2 because parallel transport is not an isometry, for instance,

$$\tau_t(\partial_1) = e^{-t^2}\, \partial_1$$

C. Geodesics

It is natural to turn the equation for parallel transport onto the curve itself: solve $\nabla_{\dot c}\dot c = 0$ for c.

Such a curve is called a *geodesic* for the connection ∇ because on the surface of the earth, viewed as a sphere with the usual connection, such curves are great circles and do "divide the earth." On an n-manifold we can always find geodesic curves going in all directions from a point but in general we may not be able to make them go very far. Accordingly, we usually define a *geodesic* through $x \in M$ with direction $u_0 \in T_x M$ in a manifold M with connection ∇ as any smooth $c\colon (-\varepsilon, \varepsilon) \to M$, for positive ε, with $c(0) = x$, $\dot c(0) = u_0$, and $\nabla_{\dot c(t)}\dot c = 0$ for all $t \in (-\varepsilon, \varepsilon)$. Cleary, if M has a boundary then a geodesic may not be extensible after it meets ∂M. Another type of inextensibility can occur if M is incomplete in some sense.

EXAMPLE. Let $M = E^2\backslash\{(0,0)\}$, the punctured plane with standard coordinates. Then the Euclidean connection has zero Christoffel symbols and the equation of a geodesic becomes

$$\nabla_{\dot c}\dot c = \nabla_{\dot c^i \partial_i}\dot c^j \partial_j = \ddot c^k \partial_k + \dot c^i \dot c^j \Gamma^k_{ij}\partial_k = 0$$

so $\ddot c^k = 0$, $k = 1, 2$. Hence the geodesics are straight lines, as we expect for a submanifold of the Euclidean plane, but they cannot pass through the origin. Thus, for example, the geodesic

$$c\colon (-\varepsilon, \varepsilon) \to M : t \mapsto (2 - 2t, 1 - t)$$

which begins at $(2, 1)$ in direction $-2\hat i - \hat j$, is only defined for $\varepsilon \leqq 1$.

EXAMPLE. To find geodesics (corresponding to free particle trajectories) in Schwarzschild space-time. Here we take $M = \mathbb{R} \times (E^3\backslash B)$ with $\mathbb{R}$ giving the time coordinate t and $E^3\backslash B$ being the Euclidean space outside a ball of some radius $k > 0$ centered on the origin. We give $E^3\backslash B$ the usual spherical polar coordinates (r, θ, ϕ) and view the region B as containing some spherically symmetric mass, like a star or planet. In general relativity, free particles falling under gravity follow geodesics. The appropriate metric tensor for this physical situation has components

$$(g_{ij}) = \begin{pmatrix} 1 - 2m/r & 0 & 0 & 0 \\ 0 & -(1 - 2m/r)^{-1} & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2 \sin^2\theta \end{pmatrix} \quad \text{for } r > k > 2m$$

where m is the mass of the material contained in B. Denoting the coordinates (t, r, θ, ϕ) by (x_0, x_1, x_2, x_3), we find the metric connection ∇ has only the following nonzero Christoffel symbols:

$$\begin{aligned} \Gamma^1_{11} &= -\Gamma^0_{10} = -(m/r^2)(1 - 2m/r)^{-1} \\ \Gamma^2_{12} &= \Gamma^3_{13} = 1/r, \\ \Gamma^1_{33} &= \Gamma^1_{22} = -r(1 - 2m/r)\sin^2\theta \\ \Gamma^1_{00} &= (m/r^2)(1 - 2m/r), \\ \Gamma^2_{33} &= \sin\theta\cos\theta \\ \Gamma^3_{32} &= \cot\theta, \quad \text{and} \quad \Gamma^k_{ij} = \Gamma^k_{ji} \end{aligned}$$

Geodesic curves satisfy $\nabla_{\dot c}\dot c = 0$ and we consider two cases each with parameter s given by $g(\dot c, \dot c) = 1$.

1. *Circular geodesics*: $c(s) = (t(s), r(s), \theta(s), \phi(s))$ with $\dot r(s) = 0$. We shall take the plane of one of these circular orbits to be $\theta(s) = \pi/2$. Denote differentiation with respect to parameter s by a dot; then we expand $\nabla_{\dot c}\dot c = 0$ to give the system of equations

$$\begin{aligned} \ddot t = 0 &\Rightarrow \dot t = \text{const} \\ \Gamma^1_{33}(\dot\phi)^2 + \Gamma^1_{00}(\dot t)^2 = 0 &\Rightarrow (1 - 2m/r)(m/r^2(\dot t)^2 \\ &\qquad - r(\dot\phi)^2) = 0 \\ &\Rightarrow (\dot t)^2 = r^3/m(\dot\phi)^2 \\ \ddot\phi = 0 &\Rightarrow \dot\phi = \text{const} = \text{period}/2\pi \\ &\Rightarrow \text{period } T = 2\pi/\dot\phi \end{aligned}$$

   From $g(\dot c, \dot c) = 1$ we find

$$(\dot t)^2(1 - 2m/r) - r^2(\dot\phi)^2 = 1$$

   Substitution above gives circular orbits with periods

$$T = 2\pi r(r/m - 3)^{1/2} \quad r > 3m$$

   In time units, for the sun we have $2m = 10^{-5}$ sec and $t = 1$ year $\doteq 10^7\pi$ sec, for the earth orbit, so we deduce that the implied radius is $r \doteq 500$ sec, which is what we observe. Similar data can be checked for the moon or other satellites orbiting the earth. For the earth $2m \doteq 3 \times 10^{-11}$ sec.
2. *Radial geodesics:* $c(s) = (t(s), r(s), \theta(s), \phi(s))$ with θ, ϕ constant. The geodesic equation reduces on $\theta = \pi/2$ to

$$\ddot r - m/r^2(1 - 2m/r)^{-1}(\dot r)^2 + m/r^2(1 - 2m/r)(\dot t)^2 = 0$$

   and we deduce $r = -m/r^2$. Now r measures precisely the acceleration due to gravity at distance r from the center of a spherically symmetric mass m, in agreement to first approximation with Newton's theory. We find, for example, in time units, that

$$\text{on the earth} \quad \begin{cases} 2m \doteq 3 \times 10^{-11}\,\text{sec} \\ r \doteq 2.1 \times 10^{-2}\,\text{sec} \\ \ddot{r} \doteq -3/8 \times 10^{-7} \end{cases}$$

$$\text{on the moon} \quad \begin{cases} 2m \doteq 2.5 \times 10^{-13}\,\text{sec} \\ r \doteq 5.4 \times 10^{-3}\,\text{sec} \\ \ddot{r} \doteq -14/3 \times 10^{-9} \end{cases}$$

$$\text{on the sun} \quad \begin{cases} 2m \doteq 10^{-5}\,\text{sec} \\ r \doteq 2\,\text{sec} \\ \ddot{r} \doteq -1/8 \times 10^{-5} \end{cases}$$

A manifold with connection is called *geodesically complete* if all of its geodesics can be extended to infinite parameter values (or until they meet the boundary, if M is a manifold with boundary). It is known that, with their standard connections induced from being embedded in Euclidean space, the circle, sphere, and torus are all geodesically complete. Observe that on such spaces the extension of a geodesic to infinite parameter values may involve it in repeatedly convering the same points, for some geodesics become closed curves.

As might be expected, from the fact that some sufficiently small region about $x \in M$ is diffeomorphic to E^n, we can get a geodesic going in any direction. That is, for all sufficiently small initial tangent vectors $\dot{c}(0) = u_0 \in T_x M$, there is a geodesic through $c(0) = x$. To make "sufficiently small" precise we need a norm in $T_x M$, but any norm will do equivalently well. Define

$$S_x = \{u_0 \in T_x M \mid \text{ there is a geodesic } c\colon [0,1] \to M \text{ with } c(0) = x \text{ and } \dot{c}(0) = u_0\}$$

Then there is a nice map called the *exponential map at x*

$$\exp_x : S_x \to M : u_0 \mapsto c(1) \quad \text{where } c \text{ is a geodesic with } c(0) = x \text{ and } \dot{c}(0) = u_0$$

It turns out that at every point x there is someneighborhood of the zero vector in $T_x M$ on which $\exp_x$ is a diffeomorphism onto its image. If the connection is complete then $\exp_x$ is defined on all of $T_x M$ for all $x \in M$.

D. Metric Connection

We saw that any Euclidean space has a connection, the simplest possible choice, and we implied that any subset of E^n that is a manifold will inherit a unique connection. Now this is a consequence of the following result:

Every metric tensor g determines a unique connection ∇^g, called the *metric connection* or *Levi–Civita connection*. There is one obvious condition that it should satisfy:

a. $\nabla^g_u g = 0$ for all $u \in \Upsilon^1_0 M$.

This is called being *compatible* with the metric and it effectively says that the covariant derivative will always view the metric tensor (and its dual g^*) as a constant: ∇^g factors out any variability introduced by peculiarties of g. In particular, it will make parallel transport an isometry as well as an isomorphism and covariant derivatives will commute with the isomorphisms $g^\#$ and $g^\flat$ induced by g.

The second condition is less obvious at first:

b. $\nabla^g_u v - \nabla^g_v u = [u, v]$ for all $u, v \in \Upsilon^1_0 M$.

This is called being *symmetric* because it implies that with respect to any coordinates the Christoffel symbols Γ^k_{ij} are symmetric in ij.

Conditions (a) and (b) are sufficient to select a unique connection for a manifold with metric tensor. They give a system of differential equations that locally allows the Christoffel symbols to be calculated from partial derviatives of the components of g. We find that

$$\partial_k g(\partial_i, \partial_j) = g\big(\nabla^g_{\partial_k} \partial_i, \partial_j\big) + g\big(\partial_i, \nabla^g_{\partial_k} \partial_j\big)$$

from a so

$$\partial_k\, g_{ij} = \Gamma^m_{ki} g_{mj} + \Gamma^m_{kj}\, g_{im},$$

$$\nabla^g_{\partial_i}\, \partial_j - \nabla^g_{\partial_j} \partial_i = [\partial_i, \partial_j] = 0$$

from b so $\Gamma^k_{ij} = \Gamma^k_{ji}$. Hence we use the inverse matrix of (g_{rs}) and symmetry to give

$$\Gamma^k_{ij} = \tfrac{1}{2} g^{km}(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij})$$

It can be shown that every manifold can be given a Riemannian metric that is geodesically complete. Evidently, the fact that $(g_{ij}) = (\delta_{ij})$ everywhere on E^n immediately gives zero Christoffel symbols in the standard coordinates; however, if we use other than rectilinear coordinates, then the corresponding metric tensor components will be nonconstant and some nonzero Christoffel symbols will airse. The idea is clear: if we wish to keep Euclidean geometry but describe it with curvilinear coordinates, then we shall expect any components in these coordinates (of vectors or tensors) to alter as we parallel transport them.

EXAMPLE. To find a metric connection and equations for a parallel vector field along a given curve. We take $M = (0, 2\pi) \times S^1$, an open cylinder with identity coordinate x on the interval $(0, 2\pi)$ and angular coordinate θ on the circle S^1. Consider the expression in these coordinates of the pseudo-Riemannian metric tensor

$$(g_{ij}) = \begin{bmatrix} -(1-\cos x)^2 & 0 \\ 0 & (1-\cos x)^2 \end{bmatrix} \quad \text{at } (x, \theta) \in M$$

From symmetry and compatibility of the induced metric connection ∇ with Christoffel symbols (Γ^m_{ij}) we have

$$\Gamma^m_{ij} = \tfrac{1}{2} g^{mk}\big(\partial_i g^{jk} + \partial_j g_{ij} - \partial_k g_{ij}\big)$$
$$= \tfrac{1}{2} g^{mk}{}_k \Gamma_{ij}, \quad \text{say}$$

Substitution gives

$$({}_1\Gamma_{ij}) = \begin{bmatrix} -2\sin x(1-\cos x) & 0 \\ 0 & -2\sin x(1-\cos x) \end{bmatrix}$$

$$({}_2\Gamma_{ij}) = \begin{bmatrix} 0 & 2\sin x(1-\cos x) \\ 2\sin x(1-\cos x) & 0 \end{bmatrix}$$

Hence

$$\big(\Gamma^1_{ij}\big) = \begin{bmatrix} \sin x/(1-\cos x) & 0 \\ 0 & \sin x/(1-\cos x) \end{bmatrix}$$

$$\big(\Gamma^2_{ij}\big) = \begin{bmatrix} 0 & \sin x/(1-\cos x) \\ \sin x/(1-\cos x) & 0 \end{bmatrix}$$

For the vertical-going curve $c\colon (0, 2\pi) \to M \,:\, t \mapsto (t, 0)$ a parallel vector field is $v\colon (0, 2\pi) \to TM\colon t \mapsto f(t)\partial_x + h(t)\partial_\theta$ where $\nabla_{\dot c} v = 0$. This differential equation becomes

$$\partial_x f + f\Gamma^1_{11} = 0$$
$$\partial_x h + h\Gamma^2_{12} = 0$$

So suitable f and g must satisfy

$$\partial_x f = -\frac{\sin x}{1-\cos x} f$$

and

$$\partial_x h = -\frac{\sin x}{1-\cos x} h$$

E. Torsion of a Connection

In general, a connection ∇ need not be related to a metric tensor; for example, the (constant) connection on the unit circle S^1 with Christoffel symbol $\Gamma^1_{11} = \lambda$ for fixed $\lambda \neq 0$ does not arise as the metric connection of any metric tensor on S^1. However, it is trivially symmetric. The question of whether a connection is symmetric is independent of any metric tensor and it can be formulated in terms of the *torsion* map

$$T\colon \Upsilon^1_0 M \times \Upsilon^1_0 M \to \Upsilon^1_0 M$$
$$: (u, v) \mapsto \nabla_u v - \nabla_v u - [u, v]$$

Hence T is zero if and only if ∇ is symmetric. For the usual basis fields $(\partial_1, \ldots, \partial_n)$ arising from coordinates $(x_1, \ldots, x_n)$ we have

$$T(\partial_i, \partial_j) = \nabla_{\partial i}\partial_j - \nabla_{\partial j}\,\partial_j \quad (\text{since } [\partial_i, \partial_j] = 0)$$
$$= \Gamma^k_{ij}\,\partial_k - \Gamma^k_{ji}\,\partial_k = \big(\Gamma^k_{ij} - \Gamma^k_{ji}\big)\partial_k$$

We can interpret T as a section of $T^1_2 M$ since at each $x \in M$ we have a bilinear map

$$T\colon T_x M \times T_x M \to T_x M$$

but such maps are effectively elements of $T_x M^* \otimes T_x M^* \otimes T_x M$. Hence we may view T as a section of $T^*M \otimes T^*M \otimes TM$ and locally

$$T = \big(\Gamma^k_{ij} - \Gamma^k_{ji}\big)\, dx^i \otimes dx^j \otimes \partial_k$$

The antisymmetry of T in i and j immediately suggests and interpretation of T as some king of 2-form. This is possible but the presence of the ∂_k direction means that then we view T as a vector-valued 2-form, the torsion form

$$\Theta\colon \Upsilon^1_0 M \times \Upsilon^1_0 M \to \Upsilon^1_0 M$$

which in local coordinates becomes

$$\Theta = \tfrac{1}{2}\Gamma_{ij}\, dx^i \wedge dx^j \quad \text{where } \Gamma_{ij} = \Gamma^k_{ij}\,\partial_k$$

so locally Θ takes values in $\mathbb{R}^n \simeq T_x M$.

In much the same way, we can represent any connection by a vector-valued 1-form, the connection form

$$\omega\colon \Upsilon^1_0 M \to \Upsilon^1_1 M$$

which in local coordinates becomes

$$\omega = \omega_i\, dx^i \qquad \text{where } \omega_i = \Gamma^k_{ij}\, dx^j \otimes \partial_k$$

So, locally ω takes values in $\mathbb{R}^{n^2}$, the space of $n \times n$ real matrices, which represents $T_x M^* \otimes T_x M \simeq L(T_x M, T_x M)$.

EXAMPLE. On E^2 with the standard coordinates, one connection ∇ that is not symmetric is given by the Christoffel symbols

$$\Gamma^1_{ij} = \begin{pmatrix} 1 & 8 \\ 4 & 0 \end{pmatrix} \qquad \Gamma^2_{ij} = \begin{pmatrix} 1 & 6 \\ 4 & 2 \end{pmatrix}$$

Its torsion form is $\Theta\colon \Upsilon^1_0 E^2 \times \Upsilon^1_0 E^2 \to \Upsilon^1_0 E^2$ with

$$\Theta = \big(\tfrac{1}{2}\Gamma^k_{ij}\partial_k\big)\, dx^i \wedge dx^j$$
$$= \tfrac{1}{2}(8\partial_1 + 6\partial_2)\, dx^1 \wedge dx^2$$
$$+ \tfrac{1}{2}(4\partial_1 + 4\partial_2)\, dx^2 \wedge dx^1$$
$$= \tfrac{1}{2}(4\partial_1 + 2\partial_2)\, dx^1 \wedge dx^2 \quad (\text{with values in } T_x M)$$

So

$$\Theta = (2, 1)\, dx^1 \wedge dx^2 \quad (\text{with values in } \mathbb{R}^2)$$

The connection form of this ∇ is $\omega : Y_0^1 E^2 \to Y_1^1 E^2$ with

$$\begin{aligned}\omega &= \left(\Gamma_{ij}^k dx^j \otimes \partial_k\right) dx^i \\ &= \begin{pmatrix} 1dx^1 \otimes \partial_1 & 8\,dx^1 \otimes \partial_2 \\ 4dx^2 \otimes \partial_1 & 0\,dx^2 \otimes \partial_2 \end{pmatrix} dx \\ &\quad + \begin{pmatrix} 1dx^1 \otimes \partial_1 & 6\,dx^1 \otimes \partial_2 \\ 4dx^1 \otimes \partial_1 & 4\,dx^2 \otimes \partial_1 \end{pmatrix} dx^2\end{aligned}$$

So

$$\omega = \begin{pmatrix} 1 & 8 \\ 4 & 0 \end{pmatrix} dx^1 + \begin{pmatrix} 1 & 6 \\ 4 & 2 \end{pmatrix} dx^2 \quad \text{(with values in } \mathbb{R}^{2\times 2}\text{)}$$

F. Curvature of a Connection

Intuitively we perceive that the unit sphere in E^3 has curvature but the (x, y) plane there has not; both are 2-manifolds inheriting a metric connection from Euclidean E^3. A geometer detects the presence of curvature by taking a tangent vector around closed curves by parallel transport. If upon return to the starting point the transported vector is always the same as the initial vector, then the connection used for the parallel transport is called *flat*; otherwise it is called *curved*. The amount of curvature at different points and in different directions is measured by the curvature map

$$\begin{aligned} R\colon \Upsilon_0^1 M \times \Upsilon_0^1 M &\to L\left(\Upsilon_0^1 M, T_0^1 M\right) \\ : (u, v) &\mapsto R(u, v)\end{aligned}$$

where $R(u, v)\colon \Upsilon_0^1 M \to \Upsilon_0^1 M : w \mapsto \nabla_u \nabla_v w - \nabla_v \nabla_u w - \nabla_{[u,v]} w$. Since $L(\Upsilon_0^1 M, \Upsilon_0^1 M) \simeq \Upsilon_1^1 M$, we can consider R to be a member of $L(\Upsilon_0^2 M, \Upsilon_1^1 M) \simeq \Upsilon_3^1 M$, that is, a $\binom{1}{3}$-tensor field. Observe that $R(u, v) = -R(v, u)$.

We interpret $R(u, v)$ at a point x as a map of $T_x M$ to itself that is a limiting case of parallel transport around a curvilinear parallelogram determined by $u(x), v(x) \in T_x M$. In components,

$$R(\partial_i, \partial_j) w = \nabla_{\partial_i} \nabla_{\partial_j} w - \nabla_{\partial_j} \nabla_{\partial_i} w \quad \text{since } [\partial_i, \partial_j] = 0$$

so

$$\begin{aligned} R(\partial_i, \partial_j)\partial_k &= \nabla_{\partial_i}\left(\Gamma_{jk}^m \partial_m\right) - \nabla_{\partial_j}\left(\Gamma_{ik}^m \partial_m\right) \\ &= \left(\partial_i \Gamma_{jk}^m + \Gamma_{jk}^l \Gamma_{il}^m\right)\partial_m - \left(\partial_j \Gamma_{ik}^m + \Gamma_{jk}^l \Gamma_{jl}^m\right)\partial_m \\ &= \left(\partial_i \Gamma_{ik}^m - \partial_j \Gamma_{jk}^m + \Gamma_{jk}^l \Gamma_{il}^m - \Gamma_{ik}^l \Gamma_{jl}^m\right)\partial_m \\ &= R_{ijk}^m\, \partial_m \end{aligned}$$

So as a $\binom{1}{3}$-tensor field

$$R = R_{ijk}^m \partial_m \otimes dx^i \otimes dx^j \otimes dx^k$$

Most conveniently we can interpret R as a vector-valued 2-form, the curvature form

$$\Omega\colon \Upsilon_0^1 M \to \Upsilon_1^1 M$$

which in local coordinates becomes

$$\Omega = \Omega_{(ij)}\, dx^i \wedge dx^j$$

where $\Omega_{ij} = (\partial_i \Gamma_{jk}^m - \Gamma_{ik}^l \Gamma_{jl}^m)\, \partial_m \otimes dx^k$. So, locally Ω takes values in $T_x M \otimes T_x M^* \simeq \mathbb{R}^{n^2}$, giving a matrix representing the limiting parallel transport map of vectors from $T_x M$ around a parallelogram defined by ∂_i, ∂_j.

It is easy to extend the exterior product and exterior derivative to vector-valued r-forms, just by applying the usual operations to their components. When we do this connection, curvature, and torsion forms we obtain the famous structural equations of E. Cartan, for example,

$$\Omega(u, v) = d\omega(u, v) + \tfrac{1}{2}[\omega(u), \omega(v)] \quad u, v \in \Upsilon^1 M$$

EXAMPLE. The previously mentioned Schwarzschild metric tensor on $M = \mathbb{R} \times E^3 \backslash B$ with coordinates (t, r, θ, ϕ) is of the form

$$(g_{ij}) = \begin{pmatrix} -f^2(r) & 0 & 0 & 0 \\ 0 & f^{-2}(r) & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2\theta \end{pmatrix}$$

with f a function of r

As before we let indices run 0, 1, 2, 3. Evidently, a basis of mutually orthogonal unit 1-form fields, that is, of orthonormal fields, is given by

$$(\omega^i) = (f\, dt,\ f^{-1}\, dr,\ r\, d\theta,\ r \sin\theta\, d\phi)$$

Their exterior derivatives satisfy the structural equations

$$d\omega^i = -\omega_j^i \wedge \omega^j \quad \text{where } \omega_j^i = \Gamma_{jk}^i \omega^k$$

and

$$\Omega_j^i = d\omega_j^i + \omega_k^i \wedge \omega_j^k \quad \text{where } \Omega_j^i = R_{l(ij)}^k w^i \wedge \omega^j$$

Computation of the derivatives yields the following with $\dot{f}$ denoting the derivative of f with respect to r:

$$\begin{aligned} d\omega^0 &= \dot{f}\omega^1 \wedge \omega^0 \\ d\omega^1 &= 0 \\ d\omega^2 &= (f/r)\omega^1 \wedge \omega^2 \\ d\omega^3 &= (f/r)\omega^1 \wedge \omega^3 + (\cot\theta)/r\, \omega^2 \wedge \omega^3 \end{aligned}$$

Then we deduce that the only nonzero ω_j^i are

$$\begin{aligned} &\omega_1^2 = -\omega_2^1 = f/r\omega^2 \quad \text{so } d\omega_1^2 = f\dot{f}/r\omega^1 \wedge \omega^2 \\ &\omega_1^3 = -\omega_3^1 = f/r\omega^3 \quad \text{so } d\omega_1^3 = f\dot{f}/r\omega^1 \wedge \omega^2 + f/r \\ &\qquad\qquad \cot\theta\, \omega^2 \wedge \omega^3 \\ &\omega_1^0 = \omega_0^1 = \dot{f}\omega^0 \quad \text{so } d\omega_1^0 = (\dot{f}^2 + f\dot{f})\omega^1 \wedge \omega^0 \\ &\omega_2^3 = -\omega_3^2 = (\cot\theta)/r\omega^3 \quad \text{so } d\omega_2^3 = -1/r^2\omega^2 \wedge \omega^3 \end{aligned}$$

By inspection of the second structural equation we find

$$\Omega_1^0 = (f\dot{f} + \dot{f}^2)\omega^1 \wedge \omega^0$$
$$\Omega_2^0 = f\dot{f}/r\,\omega^3 \wedge \omega^0$$
$$\Omega_1^3 = f\dot{f}/r\,\omega^1 \wedge \omega^3 + f/r^2 \cot\theta\,\omega^2 \wedge \omega^3$$
$$\Omega_1^2 = f\dot{f}/r\,\omega^1 \wedge \omega^2$$
$$\Omega_2^3 = (f^2 - 1)/r^2\,\omega^2 \wedge \omega^3$$
$$\Omega_3^0 = f\dot{f}/r\,\omega^3 \wedge \omega^0$$

Then from the definition of the curvature form we obtain the components R^k_{lij} of the Riemann curvature tensor. For example,

$$R^2_{332} = R^3_{223} = (f^2/r^2) - (1/r^2)$$

Einstein's equation in general relativity can be written

$$R^k_{ijk} = R^0_{ij0} + R^1_{ij1} + R^2_{ij2} + R^3_{ij3} = 0$$

It results in two differential equations for f, reducible to

$$f^2 + r\,(d/dr)(f^2) - 1 = 0$$

which admits the solution we encountered before:

$$f(r) = (1 - 2m/r)^{1/2}$$

VII. SINGULAR GEOMETRY

In this section we shall look, albeit briefly, at some situations where the geometry goes wrong. We have seen that if we start with a nice geometrical space, like Euclidean E^2, we can introduce singular behavior by removing points. Spaces obtained by such removal operations are manifestly incomplete. The interesting thing is that some geometrical spaces are not of this type but are incomplete; in other words some singular behavior still persists after all removed points are replaced.

A. Riemannian Completeness

It is an interesting geometrical problem to describe just what constitutes a singularity and where it is in such spaces. Geodesic completeness is one obvious criterion and it solves the problem for Riemannian manifolds: there is a geodesic singularity in a region $U \subseteq M$ if there is a geodesic that enters U and does not leave but cannot be extended. It can be shown by means of the Hopf-Rinow theorem that this criterion covers all other curves as well: a Riemannian manifold (without boundary) is complete with respect to all curves if and only if it is complete with respect to all geodesics. A sufficient condition for a Riemannian manifold to be geodesically complete is for it to be compact, for instance, a closed and bounded subset of some Euclidean E^n. Spheres and the torus are of this kind. It is not a necessary condition because E^n itself is complete but not compact. In a connected, complete Riemannian manifold any two points can be connected by a geodesic that is minimal in length relative to all joining curves between the points. Of course, in any Riemannian manifold, "small enough" regions are always complete.

B. Connection Completeness

Geodesics do not tell the whole story about completeness in non-Riemannian manifolds. It is possible to have geodesic completeness but some other curves may be incomplete in any reasonable definition. For example, it is possible to contrive a model space-time that is geodesically complete but in which an observer, in a rocket say, could with finite energy follow a trajectory that cannot be extended beyond a certain finite time. Such an observer would disappear from that universe. Most realistic cosmological models imply space-time geometries that are of this kind, simply because gravity is attractive.

We are faced with the problem: does a given curve $c\colon [0, 1) \to M$ admit a continuous extension by one point to the closed interval [0, 1]? If not, then we say that the curve is *inextensible*. This in itself need not imply a problem of incompleteness since, in the Euclidean plane,

$$c\colon [0, 1) \to E^2 : t \mapsto (0, (1 - t)^{-1} - 1)$$

begins at the origin and proceeds to cover the whole positive y axis but we cannot extend it to a domain including $t = 1$. The additional test that we need to make is for some kind of finiteness property. In the presence of a Riemannian metric we could use the length of the curve. Clearly, though inextensible the above curve is not finite in length, so we do not wish to infer from it any intrinsic incompleteness of the space E^2. In the absence of suitable measurements for length, as in pseudo-Riemannian manifolds, we need another device to test for finiteness. The natural one, which arises when we have a connection, depends on the use of parallel transport along the given curve.

C. *b*-Incompleteness

Let $c\colon [0, 1) \to M$ be a curve in manifold M with connection ∇ and choose a basis (b_i) for $T_{c(0)}M$. Then the parallel transport isomorphism of tangent spaces along the curve's path, say,

$$\tau_t\colon T_{c(0)}M \to T_{c(t)}M : u^i b_i \mapsto u^i \tau(b_i) = u^i \beta_i(t)$$

defines a basis $(\beta_i(t))$ for each $T_{c(t)}M$. Now, any n-dimensional vector space, once we have chosen a basis,

has a unique isomorphism with $\mathbb{R}^n$ simply by taking components with respect to that basis. Next, tmr^n has a natural norm

$$\|(x^i)\| = ((x^1)^2 + (x^2)^2 + \cdots + (x^n)^2)^{1/2}$$

So, given the choice of basis $(b_i) = (\beta_i(0))$ at $c(0)$ for our tangent spaces, we can use its parallel transported image to define a b length for the tangent vector

$$\dot{c}(t) = \dot{c}^i(t)\beta_i(t) \quad \text{to } c \text{ at } t$$

namely

$$\|\dot{c}(t)\|_b = \|(\dot{c}^i(t))\|$$

This gives us a b length of the curve with respect to basis (b_i) at $c(0)$, defined by

$$L_b(c) = \int_0^t \|\dot{c}(t)\|_b \, dt$$

In a space-time manifold, a choice of basis for one tangent space is effectively a choice of reference frame of directions and scale of units, that is, an observer. Clearly our b length will vary with the choice of the observer. However, what does not vary is whether it is finite or not. One choice of initial basis (b_i) at $c(0)$ is sufficient to test for finiteness of b-length with respect to any basis.

Accordingly, we say that curve c is b incomplete if it is inextensible and has finite b length. This definition only needs the presence of a connection, not a metric tensor. When we do have a Riemannian metric connection then the finiteness of b length is equivalent to the finitness of ordinary length. We say that a manifold with connection is b incomplete if it contains a b-incomplete curve. A Riemannian manifold is actually b incomplete if and only if it is geodesically incomplete, but this is not the case for pseudo-Riemannian manifolds.

When it was discovered that all realistic models of the universe, relativistic or otherwise, are likely to be incomplete, it was natural to enquire if this was merely an inadequacy of the field theory of gravity. Thus, it might be hoped than an appropriate quantum theory of gravity would be such as to average out any classical singularities. Unfortunately, there is no single quantum theory of gravity that is accepted by all. One theory, geometrical quantization, when applied to a massless Klein–Gordon scalar field on a curved space-time could not prevent the collapse of the state vector: the incompleteness found in the classical geometry could not be quantized away. More generally, it was shown that if there is b incompleteness with respect to one connection then there will be b incompleteness with respect to any nearby connection. So the singularity is stable under perturbations of the connection and unlikely to be removable by any quantum theory of gravity.

VIII. TOPOLOGY, GEOMETRY, AND PHYSICS

The 20th century enjoyed a wealth of developments from the interplay of formal mathematics and natural sciences, and in no other area is this as rich in results as that involving geometry, topology, and theoretical physics. One of the most remarkable families of recent results has been in the work of Freedman and Donaldson that led to their award of Fields Medals at the International Congress of Mathematicians in 1986. An amazing but easily comprehended result is that there are copies of $\mathbb{R}^4$ that are topologically indistinguishable from ordinary $\mathbb{R}^4$ but which have different manifold structures—namely, there are exotic 4-spaces and this happens in no other dimension.

In simple terms, we could say that whereas on all $\mathbb{R}^n$ for $n = 1, 2, 3, 5, 6, \ldots$, there is only one way to set up calculus, in the case of $\mathbb{R}^4$ there are infinitely many different ways to do it. It was already known that there were exotic 7-spheres (but none of lower dimension), but the new results led to the existence of exotic closed 4-manifolds. Many compact (topological) 4-manifolds cannot be given a differentiable structure and those that do admit differentiable structures may allow infinitely many. Interestingly, this was proved by using the methods of Yang-Mills theory from physics. The obstructions to differentiable structures arise from the theory of connections associated with the Yang-Mills equations and instantons. Thus, Donaldson gave an application of physical gauge field theory to geometrical topology.

Cosmology has provided rich areas of application for the curved pseudo-Riemannian geometry needed for general relativity theory and it has generated very precise expriments to detect the physical consequences. In general relativity, spacetime has its curvature controlled by the matter distribution and the curvature controls how freely gravitating bodies will move.

The usual model for a simple, homogeneous isotropic spacetime is $\mathbb{R}^4$ with the Friedmann-Robertson-Walker metric tensor, given by the arclength expression

$$ds^2 = c^2\,dt^2 - a(t)^2\left(\frac{dr^2}{1-kr^2} + r^2(\sin^2\theta\, d\phi^2 + d\theta^2)\right).$$

Here, $k = 1, 0, -1$ depending on whether the universe is closed, flat, or open, respectively. In the case $k = 1$, space is represented at each instant t by a sphere of radius $a(t)$.

Geodesics represent the trajectories of particles free from all influences other than gravitational interactions with the matter in the universe. Photons follow null geodesics, which lie on the boundary of local null cones $(ds^2 = 0)$ determined by the constancy of the speed of light. Free material particles follow timelike geodesics but even those subject to acceleration are nevertheless

constrained to lie locally in the interior ($ds^2 > 0$) of null cones of directions.

In the Friedmann-Robertson-Walker cosmological model, the wavelength of electromagnetic radiation emitted from a distant source at time t and observed at time t_0 has redshift relative to the local wavelength given by:

$$\frac{\lambda_{\text{source}}}{\lambda_{\text{local}}} = 1 + z = \frac{a(t_0)}{a(t)}.$$

Free, initially thermal radiation energy remains thermal during the expansion of the universe, with temperature proportional to $a(t)^{-1}$ and the relative expansion rate of the spacelike surfaces is given by

$$\left(\frac{1}{a}\frac{da}{dt}\right)^2 = H^2 = \frac{8}{3}\pi G\rho + \frac{k^2}{a^2} + \frac{1}{3}\Lambda$$

Here, G is the gravitational constant, H is Hubble's constant at time t, Λ is the cosmological constant and ρ is the mean matter density. The critical mean density that corresponds to $k = 0$ when $\Lambda = 0$, is given at the present epoch when $H = H_0$, by

$$\rho_{\text{crit}} = \frac{3H_0^2}{8\pi G}$$

A recent trend has been to allow Λ to differ from zero, since there is evidence that the present mean matter density is less than the critical mean density ρ_{crit}.

A Hot Big Bang some 12 Giga years ago, at $t = 0$ in the Friedmann-Robertson-Walker model, followed by an adiabatic expansion controlled by general relativity, is the broad scenario most widely accepted by cosmologists. It accounts well for the relative abundance of the lightest nuclides and for the observable microwave background radiation. However, there are some intriguing difficulties. The matter in typical galaxies arose from fluctuations within the first year after the Big Bang, because later the scales would have been too large for causal effects to occur. This gives rise to the question of the origin and nature of the fluctuations in the hot dense early phase; "cosmic inflation" is a currently favored way to answer this question. This notion involves an initial period of exponentially accelerating expansion lasting $\sim 10^{-32}$ sec, caused by a positive cosmological constant, which physicists associate with a hypothetical scalar field or "inflaton." Such an inflation period would precede the normal evolution of $a(t)$. A special case of interest is when $k = 0$, the mean matter density is zero $\rho = 0$, and the cosmological constant Λ is positive in Friedmann-Robertson-Walker spacetime. This corresponds to the de Sitter cosmological model, which happens also to be the limiting scenario for all indefinitely expanding models with $\Lambda > 0$. In de Sitter spacetime, the cosmic inflation corresponds to $a(t) \propto e^{\sqrt{\Lambda/3}t}$.

Rather surprisingly, general relativistic cosmology coupled with deep space observations lead cosmologists to conclude that most of the matter in galaxies and clusters is dark matter and so is invisible as far as electromagnetic emission or absorption is concerned. The matter is known to be there through gravitational effects and it is dominant at all scales of structure larger than galactic cores, but it is unclear what form it takes. Inference from interpreting gravitational effects on infrared spectral redshifts indicates the presence of supergalactic sheetlike clusters containing about 60% of all galaxies. The remaining galaxies seem to be about equally shared between dense filamentary conglomerations ("walls") and sparse filaments. This disposition of matter leaves large voids in the observable univese and the distribution of the sizes of these voids is also an active research area.

Quasars and active galactic nuclei need some source of energy and one possibility is for this to consist of massive black holes, of size 10^7–10^9 solar masses. The simplest geometry for an otherwise empty spacetime containing an isolated black hole with mass M is that of the Schwarzschild model. There we have

$$ds^2 = (1 - 2M/r)\,dt^2 - \frac{dr^2}{1 - 2M/r} + r^2(\sin^2\theta\,d\phi^2 + d\theta^2).$$

The event horizon consists of the surface at $r = 2M$, from which neither photons nor material particles can escape to the outside and through which any nearby particles will be drawn toward the central singularity at $r = 0$. In fact, quantum theory may allow a tunneling escape of matter and a consequential reduction of mass as a result of pair production just outside the event horizon, if only one of the pair is drawn in. Infalling matter swept up by a black hole in a galactic core could generate radiation energy through its acceleration toward the event horizon.

SEE ALSO THE FOLLOWING ARTICLES

ALGEBRA, ABSTRACT • ALGEBRAIC GEOMETRY • COSMOLOGY • LOOP GROUPS • TOPOLOGY, GENERAL

BIBLIOGRAPHY

Arnold, V., Atiyah, M., Lax, P., and Mazur, B., eds. (1999). "Mathematics: Frontiers and Perspectives," Am. Math. Soc., Providence, RI.

Beem, J. K., and Ehrlich, P. E. (1981). "Global Lorentzian Geometry," Dekker, New York.

Beem, J. K., Ehrlish, P. E., and Easley, K. L. (1996). "Global Lorentzian Geometry," Second ed., Marcel Dekker, New York.

Berger, M. (1999). "Riemannian Geometry During the Second Half of the Twentieth Century," Am. Math. Soc., Providence, RI.

Bott, R., and Tu, L. W. (1982). "Differential Forms in Algebraic Topology," Springer-Verlag, New York.

Choquet-Bruhat, Y., DeWitt-Morette, C., and Dillard-Bleick, M. (1982). "Analysis, Manifolds and Physics," 2nd ed., North-Holland, Amsterdam.

Cordero, L. A., Dodson, C. T. J., and de Leon, M. (1989). "Differential Geometry of the Frame Bundles," D. Reidel, Dordrecht.

Dekel, A., and Ostriker, J. P., eds. (1999). "Formation of Structure in the Universe," Cambridge Univ. Press, Cambridge.

Dodson, C. T. J. (1988). "Categories, Bundles and Spacetime Topology," 2nd ed., Kluwer, Dordrecht.

Dodson, C. T. J., and Parker, P. (1997). "A Users' Guide to Algebraic Topology," Kluwer, Dordrecht.

Dodson, C. T. J., and Poston, T. (1991) "Tensor Geometry," Graduate Texts in Mathematics 130, Second ed., Springer-Verlag, New York.

Donaldson, S. K. (1987). The geometry of 4-manifolds, Proc. Int. Congress of Mathematicians, Berkeley, 1986, pp. 43–54, Am. Math. Soc., Providence, RI.

Gompf, R. E., and Stipsicz, A. L. (1999). "4-Manifolds and Kirby Calculus," Am. Math. Soc., Providence, RI.

Gray, A. (1998). "Modern Differential Geometry of Curves and Surfaces," Second ed., CRC Press, Boca Raton, FL.

Peebles, P. J. E. (1993). "Principles of Physical Cosmology," Princeton Univ. Press, Princeton.

Sternberg, S. (1983). "Lectures on Differential Geometry," 2nd ed., Chelsea, New York.

Thurston, W. P. (1997). "Three-Dimensional Geometry and Topology," Princeton Univ. Press, Princeton.

Willmore, T. J. (1982). "Total Curvature in Riemannian Geometry," Ellis Horwood, Chichester.



Mathematical Logic

Yiannis N. Moschovakis
University of California, Los Angeles,
and University of Athens

I. Propositional Logic, $\mathbb{PL}$
II. First-Order Logic, $\mathbb{FOL}$
III. Gödel's Incompleteness Theorem
IV. Computability
V. Recursion and Programming
VI. Alternative Logics
VII. Set Theory

GLOSSARY

Church-Turing thesis Claim that every computable function can be computed by a Turing machine.
Computability theory Study of computable functions on the natural numbers.
Continuum hypothesis Conjecture that there are only two sizes of infinite sets of real numbers.
Database Finite, typically relational structure.
First-order logic Mathematical model of the part of language built up from the propositional connectives and the quantifiers.
Incompleteness phenomenon Gödel's discovery, that sufficiently strong axiomatic theories cannot decide all propositions which they can express.
Model theory Study of formal definability in first-order structures.
Paradox Counterintuitive truth.
Peano arithmetic Axiomatic theory of natural numbers.
Proof theory Study of inference in formal systems independently of their interpretation.
Propositional connectives The linguistic constructs "and," "not," "or," and "implies."
Quantifiers The linguistic constructs "there exists" and "for all."
Turing machine Mathematical model of computing device with unbounded memory.
Unsolvable problem A problem whose solution requires a non-existent algorithm.

NARROWLY CONSTRUED, mathematical logic is the study of *definition* and *inference* in mathematical models of fragments of language, especially the *first-order logic* fragment. Logic has made critical contributions to the foundations of science, especially through the work of Kurt Gödel, and it also has numerous applications. For *set theory* and *theoretical computer science*, these applications are so important, that parts of these fields are normally included in the modern, broad conception of the discipline.

I. PROPOSITIONAL LOGIC, $\mathbb{PL}$

Each logic $\mathbb{L}$ has a *syntax* which delineates the grammatically correct linguistic expressions of $\mathbb{L}$, a *semantics* which assigns meaning to the correct expressions, and a structured *system of proofs* which specifies the rules by which some $\mathbb{L}$-expressions can be inferred from others.

There are other words to describe these things: *formal language* is sometimes used to describe a plain syntax, *formal system* often identifies a syntax together with an inference system (but without an interpretation), and *abstract logic* has been used to refer to a syntax together with an interpretation, leaving inference aside. It is, however, a fundamental feature of logic that it draws clean distinctions and studies the connections among these three aspects of language. We explain them first in the simplest example of the "logic of propositions," which is part of many important logics.

A. Propositional Syntax

The *symbols* of $\mathbb{PL}$ are the *connectives*

$$\neg \text{ (not)} \quad \& \text{ (and)} \quad \vee \text{ (or)} \quad \rightarrow \text{ (implies, if-then)}$$

the two parentheses '(', ')', and an infinite list of (formal) propositional variables P_0, P_1, P_2, ... which intuitively stand for declarative propositions, things like "John loves Mary" or "3 is a prime number." It has only one category of grammatically correct expressions, the *formulas*, which are *strings* (finite sequences) of symbols defined inductively by the following conditions:

1. Each P_i is a formula.
2. If A and B are formulas, then so are the expressions

$$\neg A \quad (A \,\&\, B) \quad (A \vee B) \quad (A \rightarrow B)$$

For example, if P and Q are propositional variables, then $(P \rightarrow Q)$ and $(P \vee \neg P)$ are formulas, which we read as "if P then Q" and "either P or not P."

The inductive definition gives a precise specification of exactly which strings of symbols are formulas, and also insures that each formula is either *prime*, i.e., just a variable P_i, or it can be constructed in exactly one way from its simpler *immediate parts*, by one of the connectives. This makes it possible to prove properties of formulas and to define operations on them by *structural induction* on their definition.

More propositional connectives can be introduced as "abbreviations" of formula combinations, e.g.,

$$A \leftrightarrow B \equiv ((A \rightarrow B) \,\&\, (B \rightarrow A))$$

$$A \vee B \vee C \equiv (A \vee (B \vee C)).$$

TABLE I Truth Value Semantics

A	B	$\neg A$	$(A \,\&\, B)$	$(A \vee B)$	$(A \rightarrow B)$
1	1	0	1	1	1
1	0	0	0	1	0
0	1	1	0	1	1
0	0	1	0	0	1

B. Propositional Semantics

If B stands for some *true* proposition, then $\neg B$ is false, independently of the "meaning" or internal structure of B. This is an instance of a general *Compositionality Principle for* $\mathbb{PL}$: The truth value of a formula depends only on the truth values of its immediate parts. The semantics of $\mathbb{PL}$ comprise the rules for computing truth values, and they can be summarized in Table I, where 1 stands for "truth" and 0 for "falsity." By the first line of this table, for example, if A and B are both true, then $\neg A$ is false while $(A \,\&\, B)$, $(A \vee B)$, and $(A \rightarrow B)$ are all true. Notice that if A is false, then $(A \rightarrow B)$ is reckoned to be true no matter what the truth value of B, so that "if the moon is made of cheese, then $1 + 1 = 5$" is true (on the plausible assumption that the moon is not made of cheese). This *material implication* assumed by Propositional Logic has been attacked as counterintuitive, but it agrees with mathematical practice and it is the only useful interpretation of implication which accords with the Compositionality Principle.

Using these rules, we can construct for each formula A a *truth table* which tabulates its truth value under all assignments of truth values to the variables. For example, the truth table for $(Q \rightarrow P)$ consists of the first three columns of Table II while the first two and the last column give the truth table for $(P \rightarrow (Q \rightarrow P))$.

If n variables occur in a formula A, then the truth table for A has 2^n rows and determines an n-ary *bit function* v_A, with arguments and values in the two-element set $\{1, 0\}$. By the *Definitional Completeness Theorem*, every n-ary bit function is v_A for some A, so that the formulas of $\mathbb{PL}$ provide definitions (or "symbolic representations") for all bit functions.

TABLE II Truth Table

P	Q	$(Q \rightarrow P)$	$(P \rightarrow (Q \rightarrow P))$
1	1	1	1
1	0	1	1
0	0	1	1
0	1	0	1

A formula A is a *semantic consequence of* a set of formulas T (or T-*valid*) if every assignment to the variables which *satisfies* (makes true) all the formulas in T also satisfies A. We write

$$T \models A \Leftrightarrow A \text{ is } T\text{-valid},$$

and $\models A$, in the important special case when T is empty, in which case A is called a *tautology*. A formula A is *satisfiable* if some assignment satisfies it, i.e., if $\neg A$ is not a tautology. Let

$$A \sim B \Leftrightarrow \{A\} \models B \text{ and } \{B\} \models A$$
$$\Leftrightarrow \models A \leftrightarrow B,$$

and call A and B *equivalent* if $A \sim B$. Equivalent formulas define the same bit function, and they can be substituted for each other without changing truth values. Clearly

$$(A \rightarrow B) \sim (\neg A \vee B),$$

so that the implication connective is superfluous. In fact, every formula is equivalent to one in *disjunctive normal form*, i.e., a disjunction $A_1 \vee \cdots \vee A_k$ where each A_i is a conjunction of variables or negations of variables (*literals*).

C. Applications to Circuits

Each formula A with n variables can be realized by a *switching circuit* $C(A)$ with n *inputs* and one *output*, so that $C(\mathsf{P}_i)$ consists of just one input-output edge, $C(A \,\&\, B)$ is constructed by joining $C(A)$ and $C(B)$ with an *and-gate*, etc. Figure 1 exhibits the circuit for $((\mathsf{P}_1 \,\&\, \mathsf{P}_2) \rightarrow \mathsf{P}_3)$ using the equivalent formula without implications, so that only $\neg$-, &-, and $\vee$-gates are required. These are restricted circuits, of *fan-in* (maximum number of edges into a node) 2 and *fan-out* 1, but the *Definitional Completeness Theorem* implies that every n-ary bit function can be computed by some *formula circuit* $C(A)$.

There are basically two useful measures of circuit complexity, and both of them are faithfully mirrored in formulas. The *number of gates* of $C(A)$ is exactly the number of connectives in A and measures *size complexity* (construction cost), while the *depth* of $C(A)$, which measures the *time complexity* of computation, is exactly the *rank* of A, defined inductively so that $\mathrm{rk}(\mathsf{P}_i) = 1$, $\mathrm{rk}(A \,\&\, B) = \max(\mathrm{rk}(A), \mathrm{rk}(B)) + 1$ and similarly for the other connectives. One can now use natural manipulations of formulas to construct circuits which compute a given bit function with minimum size or time complexity, or to establish optimality results for the computation of bit functions by appealing to the formula representations of the circuits which realize them. For example, using disjunctive normal forms, one sees immediately that (if we do not care about cost), *every n-ary bit function can be computed by an unbounded fan-in circuit in no more than 3 time units*. There is, in general, a substantial *trade-off* between the size and time complexity of the circuits which compute a given bit function.

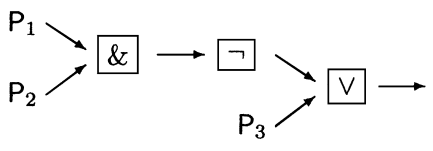


FIGURE 1 The circuit for $(\neg((\mathsf{P}_1 \,\&\, \mathsf{P}_2) \vee \mathsf{P}_3)$.

D. The Satisfiability Problem

The assertion that "$C(A)$ and $C(B)$ never give the same output on the same inputs" means precisely that "$(A \leftrightarrow \neg B)$ is a tautology," so that to detect that A and B do not have this *safety property* we need to determine whether the formula $\neg(A \leftrightarrow \neg B)$ is *satisfiable*.

Because of such natural formulations of "error detection" for circuits relative to given specifications, it is very important to find efficient algorithms for determining whether a given formula is satisfiable. The problem is of *non-deterministically polynomial time complexity* (*NP*), because it can be resolved by guessing ("non-deterministically") some assignment and then verifying that it satisfies A in a number of steps which is bounded by a polynomial in the length of A; and it is *NP*-complete, i.e., every *NP*-problem can be "reduced" to it by a polynomial reduction. This is a basic result of S. Cook, who introduced the complexity class *NP*, showed that it contains a large number of important problems, and asked if it coincides with the (seemingly) smaller class P of "feasible," *deterministically polynomial time* problems. The question whether $P = NP$ is the fundamental open problem of complexity theory; it amounts simply to the question whether the satisfiability problem can be solved by a deterministic, polynomial algorithm.

E. Propositional Inference

A *proof* of a formula A from a set of hypotheses T is any finite sequence

$$A_0, A_1, \ldots, A_{n-1}, A$$

which ends with A, and such that each A_i is either in T, or a $\mathbb{PL}$-*axiom*, or follows from previously listed formulas by a *rule of inference*. To make this notion precise we need to specify a set of $\mathbb{PL}$-axioms and rules of inference; and for these to be useful, it should be that they are few and easy to understand, and that the formulas provable from T are exactly the T-tautologies.

We need just one, binary *inference rule*:

$$\frac{A \qquad (A \to B)}{B} \text{ (Modus Ponens)}$$

This is *sound*, i.e., $\{A, (A \to B)\} \models B$, so that if A and $(A \to B)$ are both T-tautologies, then so is B.

An axiom is any instance of the following *axiom schemes*, where A, B, and C are arbitrary formulas and we have omitted several parentheses which pedantry would require:

(1) $A \to (B \to A)$
(2) $(A \to B) \to ((A \to (B \to C)) \to (A \to C))$
(3) $A \to (B \to (A \,\&\, B))$
(4) $(A \,\&\, B) \to A$ (4′) $(A \,\&\, B) \to B$
(5) $A \to (A \vee B)$ (5′) $B \to (A \vee B)$
(6) $(A \to C) \to ((B \to C) \to ((A \vee B) \to C))$
(7) $(A \to B) \to ((A \to \neg B) \to \neg A)$
(8) $\neg\neg A \to A$

These are all tautologies, and so every formula provable from T is T-valid. We write

$$T \vdash A \Leftrightarrow \text{there is a proof of } A \text{ from } T,$$

and it is not hard now to establish the basic

Soundness and Completeness Theorem for $\mathbb{PL}$. *For all sets* T *and any* A,

$$T \models A \Leftrightarrow T \vdash A.$$

F. Boolean Algebras

A *Boolean algebra* is a set $\mathcal{B}$ with at least two, distinct elements 0 and 1, a unary *complementation operation* $'$, and binary *infimum* $\cap$ and *supremum* $\cup$ operations such that certain properties hold. The standard example is the set $\mathcal{P}(M)$ of all subsets of some nonempty set M, with $0 = \emptyset$, $1 = M$ and the usual complementation, intersection and union operations, which for a singleton M gives the two-element set $\{1, 0\}$ of truth values; but there are others, e.g., the set of all finite and cofinite subsets of some infinite set, the set of all "closed and open" subsets of a topological space, etc.

Each formula A with n variables defines an n-ary function on every Boolean algebra $\mathcal{B}$, simply by letting the propositional variables range over $\mathcal{B}$ and replacing $\neg$, &, and $\vee$ and $\to$ by $'$, $\cap$, $\cup$, and $\Rightarrow$ respectively, where

$$x \Rightarrow y = x' \cup y$$

on $\mathcal{B}$. Now the axioms for a Boolean algebra insure that *every propositional axiom defines a function with constant value* 1—in fact the particular choice of axiomatization for Boolean algebras (and there are many) is quite irrelevant as long as this fact obtains; and then the Completeness Theorem implies that *two formulas* A *and* B *define the same* n*-ary operation on all Boolean algebras exactly when* $A \sim B$, i.e., when A and B define the same bit function.

Boolean algebras have many important applications in mathematics (to measure theory, among other things), and they are the subject of the classical *Stone Representation Theorem* which identifies them all (up to isomorphism) with sub-algebras of powerset algebras. In logic they are mostly used through the "nonstandard" *Boolean semantics* of this subsection, which extend to richer logics and provide a powerful tool for *independence* (unprovability) results.

II. FIRST-ORDER LOGIC, $\mathbb{FOL}$

Consider the claim:

If everybody has a mother, and every mother loves her children, then everybody is loved by somebody.

It is certainly true, it has the "linguistic form" of many similar (more substantial) claims in mathematics, and it appears to be *true by virtue of its form* and not because of any special properties of the words "mother," "love," etc. First-Order Logic makes it possible to express complex assertions of this type and to show that they are true *by logic alone*. The symbolic expression of this one will be

$$[(\forall x)(\exists y)M(x, y) \,\&\, (\forall x)(\forall y)[M(x, y) \to L(y, x)]] \to (\forall x)(\exists y)L(y, x),$$

give-or-take a few parentheses and brackets which will be required to make the syntax completely precise.

A. First-Order Syntax

The symbols of $\mathbb{FOL}$ are the propositional connectives, the parentheses, the *quantifiers*

$$\forall \text{ (for all)} \qquad \exists \text{ (there exists)}$$

the comma ',', the identity symbol '=', an infinite list $\mathsf{v}_0, \mathsf{v}_1, \ldots$ of *individual variables* which will denote arbitrary objects in some domain, and for each $n = 0, 1, \ldots$, two infinite lists of *function* and *relational symbols*

$$\mathsf{f}^n_0, \mathsf{f}^n_1, \ldots, \qquad \mathsf{P}^n_0, \mathsf{P}^n_1, \ldots,$$

which will stand for n-ary functions and relations on the objects.

There are two categories of grammatically correct expressions in $\mathbb{FOL}$, *terms* and *formulas*, defined recursively by the following conditions.

T1. Each variable v_i is a term.
T2. If $t_1, \ldots, t_n$ are terms, then (the string) $\mathsf{f}_i^n(t_1, \ldots, t_n)$ is also a term. When $n = 0$, we write simply f_i^0.
F1. If $t_1, \ldots, t_n$ are terms, then the expressions

$$t_1 = t_2 \quad \mathsf{P}_i^n(t_1, \ldots, t_n)$$

are formulas, the latter written simply P_i when $n = 0$.
F2. If A and B are formulas, then so are the expressions

$$\neg A \quad (A \,\&\, B) \quad (A \vee B) \quad (A \to B)$$

F3. If A is a formula, then so are the expressions

$$(\forall \mathsf{v}_i)A \quad (\exists \mathsf{v}_i)A$$

Notice that by the notational convention in F1, all $\mathbb{PL}$-formulas are also $\mathbb{FOL}$-formulas.

This logic is called *first order* because quantification is only allowed over individuals; if we add formula formation rules

$$\left(\forall \mathsf{P}_i^n\right)A \quad \left(\exists \mathsf{P}_i^n\right)A$$

we obtain the formulas of *second-order logic*, $\mathbb{SOL}$.

Consider the simple formula

$$(\exists \mathsf{v}_2)\left(\neg \mathsf{v}_2 = \mathsf{v}_1 \,\&\, \mathsf{P}_1^1(\mathsf{v}_2)\right). \tag{1}$$

Its "translation" into English by the reading of the symbols we have introduced is

some object other than v_1 has the property P_1^1

which is exactly how we would translate the result of substituting v_3 for v_2 in it,

$$(\exists \mathsf{v}_3)\left(\neg \mathsf{v}_3 = \mathsf{v}_1 \,\&\, \mathsf{P}_1^1(\mathsf{v}_3)\right).$$

This is because both occurrences of v_2 in Eq. (1) are *bound* by the quantifier $\exists \mathsf{v}_2$, just as the occurrences of x are bound by the dx in $\int_0^1 x^2 dx$ and can be replaced by y without changing the meaning of the definite integral. On the other hand, the occurrence of v_1 in Eq. (1) is *free*, because it is not within the scope of any quantifier, and so the interpretation of v_1 clearly affects the meaning of Eq. (1).

Using the same simple example, consider the results of substituting $\mathsf{f}_0^1(\mathsf{v}_3)$ and $\mathsf{f}_0^1(\mathsf{v}_2)$ for v_1 in Eq. (1),

$$(\exists \mathsf{v}_2)\left(\neg \mathsf{v}_2 = \mathsf{f}_0^1(\mathsf{v}_3) \,\&\, \mathsf{P}_1^1(\mathsf{v}_2)\right),$$

$$(\exists \mathsf{v}_2)\left(\neg \mathsf{v}_2 = \mathsf{f}_0^1(\mathsf{v}_2) \,\&\, \mathsf{P}_1^1(\mathsf{v}_2)\right).$$

The first of these says of $\mathsf{f}_0^1(\mathsf{v}_3)$ what Eq. (1) says of v_2, but the second says that "something is not a fixed point of f_0^1 and has property P_1^1," which is quite different—evidently because the variable v_2 in $\mathsf{f}_0^1(\mathsf{v}_2)$ is "caught" by the quantifier $\exists \mathsf{v}_2$. The first is a *free substitution* (causing no confusion) while the second is not. We will denote the result of substituting the term t for the free occurrences of the variable x in some formula A by

$$A\{x :\equiv t\}$$

and we will tacitly assume that all substitutions are free.

Formulas of $\mathbb{FOL}$ are too messy to write down, and so we often resort to "informal descriptions" of them like the example about mothers loving their children above, recipes, really, from which the full, grammatically correct formula could (in principle) be constructed.

B. First-Order Semantics

Whether Eq. (1) is true or false depends on the object v_1, on the function f_0^1, on the property P_1^1, and (most significantly) on the range of objects over which we interpret the existential quantifier—where do we search for things which may or may not satisfy P_1^1?

To interpret the formulas of $\mathbb{FOL}$ we must be given a *domain* D and an *interpretation* $\imath$, a function which assigns an object $\imath(\mathsf{v}_i)$ in D to each individual variable, an n-ary function $\imath(\mathsf{f}_i^n)$ on D to each n-ary function symbol f_i^n, and an n-ary relation $\imath(\mathsf{P}_i^n)$ on D to each P_i^n. Using these, first we extend inductively $\imath$ to all terms by

$$\imath\left(\mathsf{f}_i^n(t_1, \ldots, t_n)\right) = \left(\imath\left(\mathsf{f}_i^n\right)\right)(\imath(t_1), \ldots, \imath(t_n)),$$

so that $\imath(t)$ is some object in D. To assign truth values to formulas, define first, for each variable x and d in D, the *update*

$$\jmath = \imath\{x := d\},$$

which agrees with $\imath$ on all function and relation symbols, and also on all individual variables, except that $\jmath(x) = d$. With the help of this basic operation, we can state in Table III the classical *Tarski truth conditions* which determine the truth of formulas relative to a fixed domain D and an interpretation $\imath$. The *truth value* of a formula A relative to an interpretation $\imath$ is 1 if $\imath \models A$ and 0 otherwise, and the *Compositionality Principle* extends to $\mathbb{FOL}$ in a straightforward manner and implies the following basic fact: the truth value of A relative to $\imath$ depends only on the

TABLE III The Tarski Truth Conditions

$\imath \models t_1 = t_2 \Leftrightarrow \imath(t_1) = \imath(t_2)$
$\imath \models \mathsf{P}_i^n(t_1, \ldots, t_n) \Leftrightarrow (\imath(\mathsf{P}_i^n))(\imath(t_1), \ldots, \imath(t_n))$
$\imath \models \neg A \Leftrightarrow \imath \not\models A$
$\imath \models (A \,\&\, B) \Leftrightarrow \imath \models A$ and $\imath \models B$
$\imath \models (A \vee B) \Leftrightarrow \imath \models A$ or $\imath \models B$
$\imath \models (A \to B) \Leftrightarrow \imath \not\models A$ or $\imath \models B$
$\imath \models (\forall \mathsf{v}_i)A \Leftrightarrow$ for all d in D, $\imath\{\mathsf{v}_i := d\} \models A$
$\imath \models (\exists \mathsf{v}_i)A \Leftrightarrow$ for some d in D, $\imath\{\mathsf{v}_i := d\} \models A$

values of ι on the function and relation symbols which occur in A, and on the values $\iota(x)$ for the individual variables which occur *free* in A.

The Tarski conditions do nothing more than translate formulas into English, in effect identifying $\mathbb{FOL}$ with a precisely formulated, small but very expressive fragment of natural language.

C. Structures

A *vocabulary* (or *signature*) is any finite sequence $\sigma = \{\mathsf{f}_1, \ldots, \mathsf{f}_k, \mathsf{P}_1, \ldots, \mathsf{P}_l\}$ of function and relation symbols, and $\mathbb{FOL}(\sigma)$ is the part of $\mathbb{FOL}$ whose formulas involve only the function and relation symbols of σ. The idea is to think of $\mathsf{f}_1, \ldots, \mathsf{f}_k$ and $\mathsf{P}_1, \ldots, \mathsf{P}_l$ as constants, denoting fixed functions and relations on some set D, and to use the formulas of $\mathbb{FOL}(\sigma)$ to study definability in *structures*

$$M = (D_M, f_1, \ldots, f_k, P_1, \ldots, P_l)$$

of vocabulary σ, where the *universe* D_M of M is any nonempty set, and $f_1, \ldots, f_k$, $P_1, \ldots, P_l$ are functions and relations which can be assigned to the vocabulary symbols, e.g., such that f_i is n-ary if f_i is n-ary.

An *M-assignment* is any function α from the variables to D_M, and it extends naturally to an interpretation α_M by the association of f_i with f_i and P_i with P_i; the standard notation for *structure satisfaction* is

$$M, \alpha \models A \Leftrightarrow \alpha_M \models A.$$

Formulas of $\mathbb{FOL}(\sigma)$ with no free variables are called *sentences* and (by the Compositionality Principle) they are simply true or false in every σ-structure, without reference to any assignment. They define properties of structures. We write

$$\begin{aligned} M \models A \Leftrightarrow\ & \text{for any (and hence all) } \alpha, \\ & M, \alpha \models A \quad (A \text{ a sentence}), \end{aligned}$$

and if $M \models A$, we say that M *satisfies* A or *is a model of* A.

While sentences define properties of structures, formulas with free variables can be used to define relations on structures. If, for example, A has at most one free variable x, we set

$$R_A(d) \Leftrightarrow M, \alpha\{x := d\} \models A,$$

where α is any assignment, since its only relevant value is updated in this definition. In the same way, formulas with n free variables define n-ary *relations* on σ-structures, the *first-order definable* relations of M. A function $f : D_M^n \to D_M$ is first-order definable if its graph

$$G_f(x_1, \ldots, x_n, w) \Leftrightarrow w = f(x_1, \ldots, x_n)$$

is first-order definable. Some examples:

A *directed graph* is a structure $G = (D, E)$, where E is a binary "edge" relation on the set of "nodes" G, and it is a *graph* (undirected) if it satisfies the sentence

$$(\forall x)(\forall y)[E(x, y) \to E(y, x)].$$

Complete graphs (cliques) are characterized by the sentence

$$(\forall x)(\forall y)E(x, y),$$

while "diameter ≤ 2" is defined by

$$(\forall x)(\forall y)[x = y \vee E(x, y) \vee (\exists z)[E(x, z) \,\&\, E(z, y)]].$$

Finite directed and undirected graphs are used to model many notions in computer science, e.g., circuits.

A *semigroup* (monoid) with identity is a structure $(S, e, \cdot)$ where the identity e is some specified member of S, $\cdot$ is a binary "multiplication" on S, and the following sentences are true:

$$(\forall x)(\forall y)[x \cdot (y \cdot z) = (x \cdot y) \cdot z],$$

$$(\forall x)(x \cdot e = x \,\&\, e \cdot x = x).$$

Here and in the sequel we write $t_1 \cdot t_2$ rather than the pedantically correct $\cdot(t_1, t_2)$.

In addition to semigroups, there are *groups*, *rings*, *fields*, and *ordered fields*, *vector spaces*, and any number of other structures which are the stuff of "abstract" algebra. These classes of structures are all characterized by first-order axioms, and the use of methods from logic is becoming increasingly important in their study.

Two structures M_1 and M_2 are *isomorphic* if some one-to-one correspondence between their universes carries the functions and relations of M_1 to those of M_2. *Isomorphic structures satisfy the same first-order sentences*, but the converse is not true, as we will see in Section II.F.

D. Databases

In the most general terms, a *database* is just a finite structure, typically *relational*, i.e., without functions, only relations. "Finite" does not mean "small" or "simple," and in the interesting applications databases are huge structures of large and complex vocabularies, with basic relations such as "x is an employee born in year n," "y is the supervisor of x," etc. Properties of structures are usually called *queries* in database theory, and one of the main tasks in the field is to develop representations for databases which support fast algorithms for *updating*, entering new information in the base, and *data testing*, determining the truth or falsity of queries. As it happens, updating and data testing for first-order queries can be done very efficiently, and so database systems, including the industry standard SQL make heavy use of methods from first-order logic.

Motivated by database theory, a good deal of research has been done since the 1970s in *Finite Model Theory*, the mathematical and logical study of finite structures. For a rather surprising, basic result, let

$$\begin{aligned}&\mathrm{Prob}_\sigma[M \models A : |D_M| = n]\\ &\quad = \text{the proportion of } \sigma\text{-structures}\\ &\qquad\qquad \text{of size } n \text{ which satisfy } A,\end{aligned}$$

where structures are counted "up to isomorphism."

The $\mathbb{FOL}$ 0-1 Law. For each sentence A of $\mathbb{FOL}(\sigma)$ in a relational vocabulary, either

$$\lim_{n\to\infty} \mathrm{Prob}_\sigma[M \models A : |D_M| = n] = 1,$$

or

$$\lim_{n\to\infty} \mathrm{Prob}_\sigma[M \models A : |D_M| = n] = 0,$$

i.e., either A or $\neg A$ is asymptotically true.

More advanced work in this area is concerned primarily with the algorithmic analysis of queries on finite structures, especially in logics richer than $\mathbb{FOL}$.

E. Arithmetic

Most basic is the structure of arithmetic

$$N = (\mathbb{N}, 0, 1, +, \cdot),$$

where $\mathbb{N} = \{0, 1, \ldots\}$ is the set of (non-negative) *natural numbers* and $+$ and $\cdot$ are the operations of addition and multiplication. The first-order definable relations and functions on N are called *arithmetical*, and they obviously include addition, multiplication, and the ordering on $\mathbb{N}$, which is defined by the formula

$$x \leq y \equiv (\exists z)[x + z = y].$$

By a basic lemma of Gödel, if a function f is determined from arithmetical functions g and h by the equations

$$\begin{cases} f(0, \vec{x}) = g(\vec{x}) \\ f(y+1, \vec{x}) = h(f(y, \vec{x}), y, \vec{x}), \end{cases} \tag{2}$$

then f is also arithmetical. Thus *exponentiation* x^y is arithmetical, with $g(x) = 1, h(w, y, x) = w \cdot x$, and, with some work, so is the function $p(x)$ which enumerates the prime numbers,

$$p(0) = 2,\ \ p(1) = 3,\ \ p(2) = 5, \ldots.$$

In fact, the scheme of *Primitive Recursion* (2) is the basic method by which functions are introduced in number theory, so that, with some work, all fundamental number theoretic relations and functions are arithmetical, and all celebrated theorems and open problems of the theory of numbers are expressed by first-order sentences of N. These include the Prime Number Theorem, Fermat's Last (Wiles') Theorem, and the (still open) question whether there exist infinitely many twin pairs of prime numbers.

F. Model Theory

The mathematical theory of structures starts with the following basic result:

Compactness and Skolem-Löwenheim Theorem. If every finite subset of a set of sentences T has a model, then T has a countable model.

For an impressive application, let (in the vocabulary of arithmetic)

$$\Delta_0 \equiv 0, \quad \Delta_{m+1} \equiv (\Delta_m + 1),$$

so that the *numeral* Δ_m is about the simplest term which denotes the number m, add a constant $\mathbf{c}$ to the language, and let

$$\begin{aligned}T = &\{A : N \models A\}\\ &\cup \{\Delta_0 \leq \mathbf{c}, \Delta_1 \leq \mathbf{c}, \Delta_2 \leq \mathbf{c}, \ldots\}.\end{aligned}$$

Every finite subset S of T has a model, namely

$$N_S = (\mathbb{N}, 0, 1, +, \cdot, m),$$

where the object m which interprets $\mathbf{c}$ is some number bigger than all the numerals which occur in formulas of S. So T has a countable model

$$N_T = (\bar{\mathbb{N}}, \bar{0}, \bar{1}, \bar{+}, \bar{\cdot}, c),$$

and then $\bar{N} = (\bar{\mathbb{N}}, \bar{0}, \bar{1}, \bar{+}, \bar{\cdot})$ is a structure for the vocabulary of arithmetic which satisfies all the first-order sentences true in the "standard" structure N but is not isomorphic with N—because it has in it some object c which is "larger" than all the interpretations of the numerals Δ_0. $\Delta_1, \ldots$. It follows that, with all its expressiveness, First-Order Logic does not capture the isomorphism type of complex structures such as N.

These *nonstandard* models of arithmetic were constructed by Skolem in the 1930s. Later, in the 1950s, Abraham, Robinson constructed by the same methods *nonstandard models of analysis*, and provided firm foundations for the classical Calculus of Leibnitz with its infinitesimals and "infinitely large" real numbers.

Model Theory has advanced immensely since the early work of Tarski, Abraham Robinson and Malcev. Especially with the contributions of Shelah in the 1970s and, more recently, Hrushovsky, it has become one of the most mathematically sophisticated branches of logic, with substantial applications to algebra and number theory.

G. First-Order Inference

The proof system of First-Order Logic is an extension of that for Propositional Logic, first by *identity axioms* which insure that $=$ is an equivalence relation and a *congruence*

for all function and relation symbols, e.g., for unary function symbols,

$$(\forall x)(\forall y)[x = y \to f(x) = f(y)].$$

In addition, there are two axioms for the quantifiers,

$$A\{x :\equiv t\} \to (\exists x)A \quad (\forall x)A \to A\{x :\equiv t\},$$

assuming that the term substitutions are free; and there are two new inference rules,

$$\frac{C \to A}{C \to (\forall x)A} \qquad \frac{A \to C}{(\exists x)A \to C}$$

which can be used only when the variable x is not free in C. Proofs from a set T of $\mathbb{FOL}(\sigma)$ sentences are defined exactly as for $\mathbb{PL}$, and we set again

$$T \vdash A \Leftrightarrow \text{there is a proof of } A \text{ from } T.$$

Notice that without the restriction on the quantifier rules, the sequence

$$P(x) \to P(x), \quad P(x) \to (\forall x)P(x),$$
$$(\exists x)P(x) \to (\forall x)P(x)$$

would be a proof of $(\exists x)P(x) \to (\forall x)P(x)$, which is, obviously, not valid. With the restriction, however, for every structure M, if every M-assignment satisfies the hypothesis of either new rule, then every M-assignment satisfies the conclusion, so that the quantifier inference rules are *sound*.

H. Gödel's Completeness Theorem

A *model* of a set of sentences T in $\mathbb{FOL}(\sigma)$ is any structure M which satisfies every A in T, in symbols

$$M \models T \Leftrightarrow \text{for all } A \text{ in } T,\ M \models A.$$

We also write

$$T \models A \Leftrightarrow \text{for all } M,$$
$$M \models T \Rightarrow M \models A,$$

which extends to $\mathbb{FOL}(\sigma)$ the *semantic consequence* relation of $\mathbb{PL}$. From the comments above:

Soundness Theorem for $\mathbb{FOL}$. If $T \vdash A$, then $T \models A$.

The fundamental fact about First-Order Logic is the converse of this result:

Completeness of $\mathbb{FOL}$. If $T \models A$, then $T \vdash A$.

It may be argued that the semantic consequence relation $T \models A$ captures the intuitive notion *A follows from the assumptions in T by logic alone*, in the sense that it insures that A is true whenever all the hypotheses in T are true, independently of the meaning of the function and relation symbols. Granting that and considering the strong expressibility of First-Order Logic discussed in Section II.C above, we may then argue further that the Completeness Theorem answers definitively (for science) the ancient question of *what follows from what by logic alone*: a proposition A follows from certain assumptions T as a matter of logic (and independently of the facts), if A and T can all be expressed faithfully as $\mathbb{FOL}(\sigma)$ assertions about some σ-structure M, and $T \vdash A$. On this view, it is hard to overemphasize the importance of this result for the foundations of mathematics and science.

Incidentally, there is an obvious extension of the Tarski conditions to Second-Order Logic, e.g.,

$$\iota \models \left(\forall \mathsf{P}_i^n\right)A \Leftrightarrow \text{for all } n\text{-ary } P \text{ on } D,$$
$$\iota\left\{\mathsf{P}_i^n :\equiv P\right\} \models A.$$

However, there is no useful Completeness Theorem for $\mathbb{SOL}$, as we will see in Section IV.F.

I. Proof Theory

If Model Theory is the study of semantics independently of inference, then Proof Theory can be viewed as the mathematical investigation of formal proofs independently of interpretation. This has always been one of the most active research areas of logic, and it has been invigorated in recent years by its substantial applications to computer science, including *automated deduction*, an important component of *artificial intelligence*. Key to these applications—and the basic result of Proof Theory—is the *Extended Normal Form Theorem* of Gentzen, whose somewhat weaker (but simpler) Herbrand version is fairly easy to describe.

There are four Herbrand inference rules, and they apply to n-ary disjunctions

$$A_1 \vee \cdots \vee A_n.$$

Two of them are *structural*, and they clearly preserve meaning: you can interchange the order of the disjuncts, or delete one of two occurrences of the same disjunct. The other two are *quantifier rules*,

$$\frac{A_1 \vee \cdots \vee A_n\{x :\equiv t\}}{A_1 \vee \cdots \vee (\exists x)A_n} \qquad \frac{A_1 \vee \cdots \vee A_n}{A_1 \vee \cdots \vee (\forall x)A_n}\ *$$

where the $*$ indicates that the $\forall$-rule can only be used if the variable x is not free in its conclusion. The result applies only to sentences without identity and in *prenex normal form*, i.e., looking like

$$(\mathsf{Q}_1 x_1) \cdots (\mathsf{Q}_n)B$$

where each Q_i is $\forall$ or $\exists$ and B is quantifier-free.

Herbrand's Theorem. Every provable $=-$free sentence A of $\mathbb{FOL}(\sigma)$ in prenex form can be derived from a provable quantifier-free disjunction by the four Herbrand rules.

The restriction to prenex sentences is not essential, because every formula can be converted to an equivalent prenex one by the application of simple rules which can be added to the system.

The theorem asserts (in part) that every provable sentence A has a "normal" proof, in which only formulas of "quantifier rank" no greater than A occur. This is a powerful tool for proof-theoretic studies. As for applications, all automated deduction systems use Herbrand-like inference systems (or their Gentzen variants), and the programming language PROLOG is based entirely on this idea.

The proof of Herbrand's Theorem is constructive: an algorithm is defined, which computes for each proof Π of a prenex sentence A a Herbrand proof Π', and then it is shown by simple, combinatorial arguments that Π', indeed, proves A. The additional, effective content is significant for the foundational applications of the theorem (for example to consistency proofs), and also in the applications to automated deduction.

It should be emphasized that the simplistic slogans "Model Theory = no inference" and "Proof Theory = no semantics" are often honored in the breach: like the Completeness Theorem, most fundamental results of logic are about connections between truth and proof, and some of the deepest results in one part of the discipline depend on methods and ideas from the other.

III. GÖDEL'S INCOMPLETENESS THEOREM

Having established that $\mathbb{FOL}$ proves all *logical truths*, it is natural to ask if it can also prove—from some natural set of axioms—all *mathematical truths*. This is not possible, by Gödel's fundamental result, whose special case for *arithmetical truths* we discuss in this section.

A. The Incompleteness of Peano Arithmetic

The classical *Peano axioms* for arithmetic comprise the properties of the successor

$$x+1 \neq 0 \qquad x+1 = y+1 \to x = y, \tag{3}$$

the recursive definitions of addition and multiplication,

$$\begin{cases} x+0 = x \\ x+(y+1) = (x+y)+1, \end{cases} \tag{4}$$

$$\begin{cases} x\cdot 0 = 0 \\ x\cdot(y+1) = x\cdot y + x, \end{cases} \tag{5}$$

and the *Induction Axiom* which cannot be expressed fully in First-Order Logic. Its Second-Order Logic version is

$$(\forall P)[(P(0)\ \&\ (\forall x)(P(x) \to P(x+1))) \to (\forall x)P(x)],$$

and the best we can do in $\mathbb{FOL}$ is to adopt the Axiom Scheme

$$(A\{y :\equiv 0\}\ \&\ (\forall x)(A\{y :\equiv x\} \to A\{y :\equiv x+1\})) \to (\forall x)A\{y :\equiv x\}. \tag{6}$$

The set PA of (first-order) Peano axioms is obtained by taking the correctly spelled versions of all the formulas in (3)–(6) and adding enough universal quantifiers in front of them so that they become sentences. This is a very strong set of axioms, it can prove all simple properties of numbers and most of their deep properties too—although proving a theorem from PA is harder than proving it using, say, methods from analysis, and number theorists distinguish and value "elementary proofs" in PA.

Gödel's First Incompleteness Theorem. There is a sentence $\mathbf{g}$ in $\mathbb{FOL}(0, 1, +, \cdot)$, such that $N \models \mathbf{g}$ but $\mathsf{PA} \not\vdash \mathbf{g}$.

One's first thought is that we can overcome this "incompleteness phenomenon" by strengthening PA, perhaps add Gödel's own $\mathbf{g}$ to it, or use the Second-Order Logic version of the Induction Axiom along with a suitable axiomatization of Second-Order Logic. None of this helps: Gödel's fundamental discovery is that first-order truth in N (and every other sufficiently rich structure) simply cannot be presented usefully as an "axiomatic theory." We will make this precise in a more general version of the Incompleteness Theorem in the next section.

B. Coding (Gödel numbering)

The basic ingredients of the proof of the Incompleteness Theorem are *coding* and *self-reference*.

In analytic geometry we "code" (represent) points in the plane by pairs of real numbers, their coordinates, so we can translate geometrical questions into algebraic problems and solve them by calculation. Gödel's basic idea is to code the syntactic objects of $\mathbb{FOL}(0, 1, +, \cdot)$—terms, formulas, proofs—by natural numbers, so that their properties are translated into properties of numbers, which can then be expressed in $\mathbb{FOL}(0, 1, +, \cdot)$ and (perhaps) proved in PA.

Since all syntactic objects are strings of symbols, if we view a proof $A_1, \ldots, A_{n-1}$ as a sequence of formulas separated by commas, it is enough to code strings, and we can do this in (at least) one simple-minded way: we enumerate the symbols of the language

$\neg$	$\&$	$\vee$	$\to$	(	)	$\forall$	$\exists$	,	$=$	0	1	$+$	$\cdot$	v_0	v_1	$\cdots$
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	$\cdots$

and we set

$$[a_0 a_1 a_2 \cdots a_m] = 2^{n_0} 3^{n_1} 5^{n_2} \cdots p(m)^{n_m},$$

where n_i is the code of the symbol a_i and $p(i)$ is the ith prime number. For example, the (correctly spelled) prime formula

$$+(\mathsf{v}_1, 0) = \mathsf{v}_0$$

has the horrendously large code

$$2^{13}3^5 5^{16}7^9 11^{11}13^6 17^{10}19^{15}.$$

The size of codes is irrelevant: what matters is that every string of symbols (and hence every term, formula and proof) has a code from which it can be reconstructed, by the *Unique Factorization Theorem* for numbers; and (more significantly) that PA is powerful enough to express and prove simple properties of formulas and proofs, thus translated into properties of numbers. For example, if Δ_n is the numeral denoting n, as above, then PA can prove all true, basic relations among numerals, e.g.,

$$m + n = k \Rightarrow \mathsf{PA} \vdash \Delta_m + \Delta_n = \Delta_k.$$

Less trivially, the basic (coded) *proof relation*

$\mathsf{Proof}_{\mathsf{PA}}(a, p) \Leftrightarrow a$ is the code of some sentence A and p is the code of a proof of A from PA

is defined by some formula $\mathsf{Proof}_{\mathsf{PA}}$ with v_1 and v_2 free, and PA can prove its basic properties, e.g.,

$$\mathsf{Proof}_{\mathsf{PA}}(a, p) \Rightarrow \mathsf{PA} \vdash \mathsf{Proof}_{\mathsf{PA}}\{\mathsf{v}_1 :\equiv \Delta_a, \mathsf{v}_2 :\equiv \Delta_p\}.$$

Similarly, the relation

$D(a, p) \Leftrightarrow a$ is the code of some formula A with only v_1 free, and p is the code of a PA-proof of $A\{\mathsf{v}_1 :\equiv \Delta_a\}$

is defined by some formula D with just v_1, v_2 free. Set

$$A \equiv (\forall \mathsf{v}_2)\neg \mathsf{D}$$

so that only v_1 is free in A, and if a is the code of A, set

$$\mathbf{g} \equiv A\{\mathsf{v}_1 :\equiv \Delta_a\}.$$

Unscrambling the definitions, **g** asserts that *there is no* PA-*proof of* $A\{\mathsf{v}_1 :\equiv \Delta_a\}$; but **g** *is* $A\{\mathsf{v}_1 :\equiv \Delta_a\}$, so that **g** claims its own unprovability; and a careful analysis of the situation shows that, indeed, **g** cannot be provable in PA, else PA would prove a contradiction. This also shows, that **g** is true.

It is not that simple, of course, and much delicate analysis and computation must be done to establish that $D(a, p)$ is arithmetical and to derive a formal contradiction from the assumption that **g** is PA-provable. Key to the proof is the "self-reference" in the definition of $D(a, p)$, which uses the coding, and the argument depends on the strength (not the weakness) of the axiomatic system PA. Coding and self-reference have become standard tools of logic since Gödel's work, and they have found substantial applications in many areas, including computer science and set theory.

IV. COMPUTABILITY

It is easy to determine whether an arbitrary equation $a_0 + a_1 x + \cdots + a_n x^n = 0$ with integer coefficients $a_0, \ldots, a_n$ has integer solutions, since every integer root must divide a_0, and so all we have to do is to test the finitely many divisors of a_0. The problem is not so easy for equations in k unknowns

$$\sum_{r_1+\cdots+r_k \le n} a_{r_1,\ldots,r_n} x_1^{r_1} x_2^{r_2} \cdots x_k^{r_k} = 0, \tag{7}$$

and it is much more interesting, in fact

to find an algorithm which determines whether Eq. (7) has a solution

is No. 10 in David Hilbert's famous 1900 list of 23 open problems in mathematics. Diophantine equations are notoriously difficult to solve, and one might suspect that no algorithm would do the job, but how can you prove such an assertion? Using ideas and techniques from Gödel's work and motivated by questions arising from it, logicians developed, in the 1930s, a tool for establishing *absolute unsolvability results* of this kind which led to some spectacular applications, including a rigorous proof of the unsolvability of Hilbert's 10th.

The most direct approach was by Turing, who reasoned that algorithms should be implemented by "mechanical devices" and introduced "abstract machines" that can perform symbolic computations some ten years before digital computers were invented.

A. Turing Machines

A Turing machine M is determined by a finite alphabet $S_M = \{s_0, \ldots, s_k\}$, a finite set $Q_M = \{q_0, \ldots, q_m\}$ of (internal) *states*, and a finite table of *transitions* of the form

$$q, s \mapsto q', s', m$$

where q, q' are states, s, s' are in S_M or the special "blank" symbol $\sqcup$, and *the move* m is -1, 0, or $+1$. No two transitions are *activated* by the same pair q, s on the left. We imagine that, at any moment, M is in some internal state q and sits in front of an infinite "tape" with symbols in some of its *cells*. The machine can only "see" the symbol s just in front of it, and does nothing (*halts*) unless one of its transitions is activated by the pair q, s; in which case

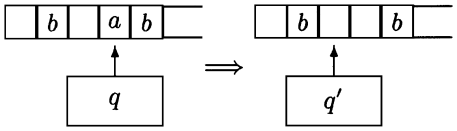
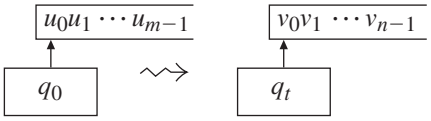


FIGURE 2 $q, a \mapsto q', \sqcup, -1$.

it switches to state q', it replaces s by s' on the tape, and it moves left (if it can), right or none-at-all, depending on whether m is $-1, +1$ or 0 (Fig. 2).

A machine M starts computing facing the leftmost cell, with an arbitrary string *input* $u = u_0 \cdots u_{m-1}$ on the tape,



and it may *diverge* (never halt), for example if $u = 11$ and M has the two transitions

$$q_0, 1 \mapsto q_0, 1, +1 \quad q_0, \sqcup \mapsto q_0, 1, +1$$

If it halts, then its *output* on u is the string $M[u] = v_0 \cdots v_{n-1}$ at the left end of the tape, until the first blank (and it is possible that $M[u]$ is empty.)

Finally, M *computes* a string function $f : S_1^* \to S_2^*$ if $S_1 \cup S_2 \subseteq S_M$ and for every string $u \in S_1^*$, $M[u] = f(u)$. By identifying each natural number n with the string $|| \cdots |$ of $n + 1$ *tallies* from the one-member alphabet $\{|\}$ (unary notation), the notion covers functions whose arguments or values are either strings or numbers. Moreover, if we code strings by numbers as above, then the transformation $u \mapsto [u]$ and its inverse can be computed by a Turing machine, so that a string function is Turing computable exactly when its "coded version" is computable, and we can safely confuse the two notions.

B. The Church-Turing Thesis

Turing argued persuasively that the symbolic computations of any "finite mechanical device" with access to unbounded memory can be simulated by one of his machines, and he has been fully justified by the subsequent developments in computers. Church had already made an equivalent (though less well justified) claim, and so the new fundamental principle carries both famous names:

The Church-Turing Thesis: A string function $f: S_1^* \to S_2^*$ is computable if and only if it can be computed by a Turing machine M on some alphabet $S_M \supseteq S_1 \cup S_2$.

The Church-Turing Thesis cannot be rigorously proved, as it identifies the intuitive, informal notion of "computability" with the precise, mathematical property of *Turing computability*. Within mathematics, it is officially a definition, much like the definitions of *arclength* or *area* in terms of integrals. But mathematical definitions are not entirely arbitrary: when we "define" the length of the circumference of a circle of radius r by an integral which computes out to $2\pi r$, we fully expect that if we draw such a circle and measure its circumference, it will turn out to be $2\pi r$, within the margin of error of our measurements. Similarly, when we prove that a certain string function f is not Turing computable, we fully expect that nobody will ever discover an algorithm which computes f, because no such algorithm exists. This is the standard method of application of the thesis.

Evidence for the Church-Turing Thesis comes from Turing's analysis, from the sixty-odd years of failed attempts to contradict it, and from the robustness of the notion of Turing computability. Many classes of functions were defined in the thirties claiming to capture the notion of "computable" from different perspectives, including Church's *λ-definable functions*, Post's *canonical systems*, the *general recursive functions* of Gödel, Herbrand, and Kleene, Kleene's *μ-recursive functions* and, in the forties, Markov's (formal) *algorithms*; each of these was proved equivalent to Turing computability, and the "simulation techniques" developed for these proofs make it seem very unlikely that some algorithm will ever be discovered which cannot be simulated by a Turing machine.

It should be emphasized, however, that the Church-Turing Thesis does not provide a rigorous definition for the notion of *algorithm*, which remains informal. Complexity results about algorithms are rigorously grounded on various so-called *computation models* which embody diverse features of actual computers. When we simulate these models by Turing machines, the time and space complexity of computations increase substantially, and so we cannot claim that the informal algorithm has been faithfully modeled. On the other hand, the time complexity increase is bound by a polynomial factor for all the known simulations, so that the class P of polynomial problems can be defined in terms of Turing machines without ambiguity.

Turing-computable functions are also called *recursive*, because of the basic Gödel-Herbrand-Kleene characterization mentioned above.

C. Unsolvable Problems

A set of strings (or *problem*) $Q \subseteq S^*$ from a finite "alphabet" S is *computable* (*recursive*, *solvable*, *decidable*) if some Turing machine M computes its *characteristic function*

$$c_Q(u) = \begin{cases} 1 & \text{if } u \in Q, \\ 0 & \text{otherwise,} \end{cases}$$

otherwise it is *unsolvable* or *undecidable*. The definitions apply to problems about natural numbers, coded in unary; to problems about $\mathbb{FOL}$-formulas, by identifying (for example) each variable v_i by a similar sequence $\mathsf{vv}\cdots\mathsf{v}$ of $i+1$ v's, so that the syntax of $\mathbb{FOL}$ is based on a finite vocabulary; and to relations (sets of n-tuples) on strings or numbers, by thinking of $u_1, \ldots, u_n$ as a single string.

Each Turing machine can be represented by a string of 0's and 1's which codes its alphabet, internal states, and transitions, and this leads to the first and most basic unsolvability result, due to Turing:

The Halting problem: It is undecidable whether an arbitrary Turing machine M halts on an arbitrary binary string u.

For the proof, Turing constructed a *universal machine* U which can simulate every other, i.e.,

$$U[\bar{M}, u] = M[u], \quad \text{if } \bar{M} \text{ is the code of } M.$$

This treatment of *programs* as *data* is, of course, routine today.

All unsolvability results are (ultimately) established by reducing the Halting Problem to them, i.e., showing that if such-and-such a function were computable, then the Halting Problem would be solvable. The proofs are often difficult and generally depend on results specific to the field in which the problem arises.

In mathematics, the problems which have been proved unsolvable include:

Hilbert's 10th: Whether a given Diophantine equation has integer solutions (Matijasevich, following work of Martin Davis, Hilary Putnam, and Julia Robinson).

The Word Problem for Groups: Whether two words denote the same element in a finitely generated, finitely presented group (P. Novikov, W. Boone).

The Homeomorphism Problem for 4-manifolds: Whether the orientable n-manifolds represented by two triangulations are homeomorphic, for $n \geq 4$ (A. Markov). This problem is solvable for 2-manifolds, by their classical representation as spheres with handles, and it is still open for 3-manifolds, pending (among other things) the resolution of the Poincaré Conjecture.

There is also a large number of unsolvable problems in computer science.

D. Undecidable Theories

A *theory* T in $\mathbb{FOL}(\sigma)$ is any set of sentences closed under consequence,

$$T \vdash A \Rightarrow A \in T.$$

The two basic examples are theories of σ-structures

$$\mathrm{Th}(M) = \{A \mid M \models A\},$$

and *axiomatic theories* of the form

$$T = \mathrm{Th}(T_0) = \{A : T_0 \vdash A\},$$

where T_0 is a decidable *set of axioms* T_0. The terminology is natural, because we would certainly demand of any "axiomatization" that it can be decided effectively whether an arbitrary sentence is an axiom.

Every decidable theory T is axiomatizable since $\mathrm{Th}(T) = T$ when T is a theory, but the converse fails, in general, and in particular for $T_0 = \emptyset$ when the vocabulary is not trivial:

Church's Theorem: If the vocabulary σ includes at least one binary function or relation symbol, then it is undecidable for a sentence A of $\mathbb{FOL}(\sigma)$ whether $\vdash A$.

A $\mathbb{FOL}(\sigma)$-theory T is *consistent* if it does not contain a contradiction $A \,\&\, \neg A$, and it is *complete* if for every sentence A, either A or $\neg A$ is in T. It is easy to verify that *every consistent, axiomatizable, complete theory is decidable*, and we can use this to formulate and prove a very general version of the Gödel Incompleteness Theorem. The key tool is the notion of *translation*.

Suppose T_1 and T_2 are theories, perhaps in different vocabularies σ_1 and σ_2—e.g., T_1 might by Th(PA), and T_2 might be some axiomatic set theory. A *translation* of T_1 into T_2 is a computable string function ρ which assigns a sentence $\rho(A)$ of $\mathbb{FOL}(\sigma_2)$ to every sentence A of $\mathbb{FOL}(\sigma_1)$ and preserves propositional logic and T_1-inference, i.e.,

$$T_2 \vdash \rho(\neg A) \leftrightarrow \neg\rho(A)$$

$$T_2 \vdash \rho(A \,\&\, B) \leftrightarrow \rho(A) \,\&\, \rho(B)$$

$$T_1 \vdash A \Rightarrow T_2 \vdash \rho(A).$$

Notice that the identity function $\rho(A) = A$ translates every theory into itself.

The Gödel Incompleteness Theorem (Rosser's form): If T is a consistent, axiomatizable theory and Peano arithmetic Th(PA) is translatable into T, then T is undecidable and hence incomplete.

In short, every consistent axiomatic system in which a reasonable amount of mathematics can be developed is undecidable and incomplete.

To state the strongest corresponding result about theories of structures, we need the simple fact that *every computable set is arithmetical*, essentially due to Gödel.

Tarski's Theorem: If $\mathrm{Th}(N)$ is translatable into $\mathrm{Th}(M)$, then $\mathrm{Th}(M)$ is not arithmetical, a fortiori it is not decidable.

To apply Tarski's Theorem, you need (in effect) to give a first-order definition of the natural numbers within the given structure. One of the first results of this type was *the undecidabilty of the theory of rational numbers* $\mathrm{Th}(\mathbb{Q}, 0, 1, +, \cdot)$ (Julia Robinson), but there are many others, and there are also many difficult open problems in this area.

On the other hand, many interesting theories are decidable, including the following:

- The theory $\mathrm{Th}(\mathbb{N}, 0, 1, +)$ of arithmetic without multiplication (Presburger).
- The theory $\mathrm{Th}(\mathbb{Q}, \leq)$. This coincides with the theory of every *dense, linear ordering without end points*.
- The theory $\mathrm{Th}(\mathbb{C}, 0, 1, +, \cdot)$ of the complex number field, which coincides with the theory of every algebraically closed field of characteristic 0 (Tarski, Abraham Robinson).
- The theory $\mathrm{Th}(\mathbb{R}, 0, 1, +, \cdot, \leq)$ of the ordered field of real numbers, which coincides with the theory of every real closed field (Tarski).

The classical result here is Tarski's decidability of the ordered field of real numbers, which (using coordinates) implies that *Euclidean geometry is decidable*, in a sense trivializing much of ancient Greek mathematics. It is still open whether the extended theory $\mathrm{Th}(\mathbb{R}, 0, 1, +, \cdot, \leq, \uparrow)$ (with $x \uparrow y = x^y$ for $x > 0$) is decidable, but there has been substantial progress in this problem with Wilkie's Theorem, that every set in $\mathbb{R}$ which is first-order definable using exponentials is a finite union of intervals.

E. The Second Incompleteness Theorem

What sorts of sentences are not provable in sufficiently strong axiomatizable theories? If $T = \mathrm{Th}(T_0)$ is axiomatizable in $\mathbb{FOL}(\sigma)$, then the (coded) proof relation

$\mathrm{Proof}_T(a, p) \Leftrightarrow a$ is the code of some sentence A in $\mathbb{FOL}(\sigma)$ and p is the code of a proof of A from T

is Turing computable, and hence arithmetical. Using this, we can construct a sentence Concis_T in the vocabulary of PA which expresses naturally the consistency of T and establish the following:

Gödel's Second Incompleteness Theorem (Rosser's form): If T is consistent, axiomatizable and ρ translates $\mathrm{Th}(\mathsf{PA})$ in T, then T cannot prove the translation $\rho(\mathsf{Consis}_T)$ of its consistency sentence.

The theorem makes it clear that we cannot axiomatize a substantial part of mathematics *in any way whatsoever* so that the consistency of the system can be established "constructively": because the (presumably simple) "constructive methods" we would be willing to use in a consistency proof should be part of the "substantial part of mathematics" we want to axiomatize. Beyond its obvious foundational significance, the Second Incompleteness Theorem has numerous applications, especially in comparing the strength of various hypotheses in Axiomatic Set Theory.

F. Hierarchies

A set Q of strings or numbers is Σ^0_2 if

$$u \in Q \Leftrightarrow (\exists x_1)(\forall x_2)R(u, x_1, x_2),$$

where the quantified variables range over natural numbers and the *matrix* R is computable, and it is Π^0_3 if, for all u

$$u \in Q \Leftrightarrow (\forall x_1)(\exists x_2)(\forall x_3)R(u, x_1, x_2, x_3)$$

with the same restrictions. The definitions extend naturally to all k, and we also set

$$\Delta^0_k = \Sigma^0_k \cap \Pi^0_k.$$

Kleene, who introduced these classes, showed that

$$\Delta^0_1 = \text{the class of recursive sets},$$

$$\Delta^0_1 \subsetneq \begin{matrix} \Sigma^0_1 \\ \cap \quad \cup \\ \Pi^0_1 \end{matrix} \subsetneq \begin{matrix} \Sigma^0_2 \\ \cap \quad \cup \\ \Pi^0_2 \end{matrix} \subsetneq \cdots$$

and that a nonempty set Q is Σ^0_1 exactly when it is *recursively* (or *computably*) *enumerable*, i.e., if

$$Q = \{f(0), f(1), \ldots\}$$

with some recursive $f : \mathbb{N} \to S^*$. Moreover, these classes increase properly and exhaust the arithmetical sets. A similar hierarchy

$$\Sigma^1_k, \Pi^1_k, \Delta^1_k$$

for the *analytical* (second-order definable) sets is constructed by allowing the quantified variables to range over the unary functions $\alpha : \mathbb{N} \to \mathbb{N}$ and the matrix to be arithmetical, so that all arithmetical sets are in Δ^1_1.

These hierarchies classify the analytical sets of natural numbers and strings by the logical complexity of their (simplest) definitions, and they are powerful tools in the theory of definability. For example, every axiomatizable theory is Σ^0_1. This rules out an axiomatization of Second-Order Logic $\mathbb{SOL}$, whose set of valid sentences (on the empty vocabulary) is not analytical. Somewhat surprisingly, it also rules out an axiomatization of the theory

$$T_f = \{A \mid \text{for all finite } (D, E), (D, E) \models A\}$$

of finite graphs, which is Π^0_1 but not Σ^0_1 (Trachtenbrot).

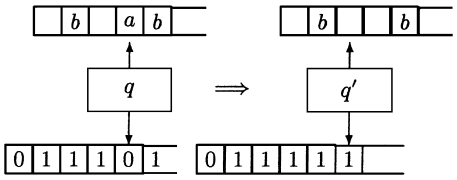


FIGURE 3 $q, a, 0 \mapsto q', \sqcup, 1, -1, +1$.

G. Turing Reducibility

Imagine a Turing machine with a second *query tape* which it handles exactly like its primary tape, implementing somewhat more complex transitions of the form

$$q, s_1, s_2 \mapsto q', s_1', s_2', m_1, m_2$$

It also has a special *query state* $q_?$, and when it goes into $q_?$, the computation stops and does not resume until some external agent (the *oracle*) replaces the contents on the query tape by some string (Fig. 3).

A string function f is *computable relative to some given* g if it can be computed by such an oracle machine, provided each time $q_?$ is reached, the string u on the query tape is replaced by the value $g(u)$. We let

$$f \leq_T g \Leftrightarrow f \text{ is computable in } g,$$

and we extend this notion of *Turing reducibility* to sets of natural numbers via their characteristic functions.

It is not hard to show that there exist Turing-incomparable sets of numbers (Kleene-Post). In fact, there exist Turing-incomparable recursively enumerable sets, but this was quite hard to prove and it was a celebrated open question for some twelve years, known as *Post's Problem*. The simultaneous, independent discovery in 1956 by Friedberg and Muchnik of the *priority method* which proved it, initiated an intense study of Turing reducibility which is still, today, one of the most active research areas of logic, the largest (and technically most sophisticated) part of *computability* or *recursion theory*.

V. RECURSION AND PROGRAMMING

In its most general form, a recursive definition of a function x is expressed by a *recursive* (or *fixed point*) *equation*

$$x(t) = f(t, x), \tag{8}$$

where the *functional* $f(t, x)$ provides a method for computing each value $x(t)$, perhaps using ("calling") other values of x in the process. It is possible to characterize the computable functions on the natural numbers using simple recursive equations of this form, generalizations of the primitive recursive definition (2) in Section III.E. Though conceptually less direct than Turing's approach through idealized machines, this modeling of computability by "recursiveness" provides a powerful tool for establishing properties of computable functions, and it is especially useful in the theory of programming languages.

A. Recursive Equations

Not every recursive equation (8) has a solution x, and some have many, e.g., the trivial $x(t) = x(t)$ which is satisfied by every function. The basic result which guarantees *canonical solutions* to a large class of recursive equations comes from the theory of *partially ordered sets*.

A *partially ordered set* or *poset* is a structure $(D, \leq_D)$, where $\leq$ is a binary relation and for all x, y, z in D,

$$x \leq_D x, \quad [x \leq_D y \,\&\, y \leq_D z] \Rightarrow x \leq_D z$$
$$[x \leq_D y \,\&\, y \leq_D x] \Rightarrow x = y;$$

a subset C of D is a *chain* if every two members of C are $\leq_D$-comparable, i.e., $x \leq_D y$ or $y \leq_D x$; and a poset D is *complete* if every chain in D has a *supremum* (least upper bound).

Every complete poset has a least element $\perp$ (the supremum of the empty chain), and every set A can be turned into a *flat poset* $A_\perp$ by adding a "bottom" below all its otherwise incomparable elements (Fig. 4). Other, basic examples include the set of all subsets of a set A (under $\subseteq$) and the set of all (finite and infinite) sequences from some set, under "extension." The Cartesian product of complete posets is complete, and, more importantly, if W is complete, then the it function spaces of all arbitrary, monotone or Scott-continuous mappings $\pi : D \to W$ are also complete, with the pointwise partial ordering

$$\pi \leq \rho \Leftrightarrow \text{for all } x, \pi(x) \leq \rho(x).$$

Here $\pi : D \to W$ is monotone if

$$x \leq_D y \Rightarrow \pi(x) \leq_W \pi(y),$$

and it is Scott-continuous if, in addition, for every chain C in D,

$$\pi(\text{supremum}(C)) = \text{supremum}(\pi[C]).$$

The Least-Fixed-Point Theorem. If $(D, \leq)$ is a complete poset and $\pi : D \to D$ is monotone, then the recursive equation

$$x = \pi(x) \tag{9}$$

has a least solution.

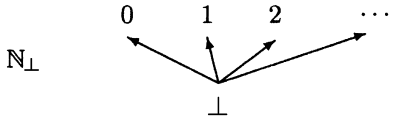


FIGURE 4 Flat poset.

The theorem is proved by setting recursively

$$x^0 = \bot, \quad x^{n+1} = \pi(x^n). \tag{10}$$

In the simplest case, which is sufficient for the applications to programming languages, the mapping π is Scott continuous, and then

$$\bar{x} = \text{supremum}\{x^0, x^1, \ldots\}$$

is the least fixed point of π. For the full result we need to extend the iteration (10) into the "transfinite," using recursion on ordinal numbers.

There is a rich theory of complete posets and various kinds of mappings on them, mostly motivated by the applications to programming, but also by earlier work in *abstract recursion*, the generalization of computability to abstract structures.

B. Programming Languages

From the mathematical point of view, a programming language $\mathbb{P}$ is very much like a logic, with a syntax, a semantics, and an *implementation*, which plays the role of an inference system.

The *syntax* is generally much more complex than that of logics, with many different categories of grammatically correct expressions. There are variables of various kinds, some of them for *functions* of specified *types*; constants which are meant to denote *acts* of interaction with the environment (input, output, interrupts); and various ways of combining grammatically correct expressions to produce new ones, using *programming constructs* like composition, "while loops," functional abstraction, and recursion. Some *closed* expressions (with no free variables) corresponding to the "sentences" of a logic are singled out, typically called *programs*. With all this complexity, the "grammar" is still specified by an induction, as it is for logics, so that it is again possible to prove properties of correct expressions and to define operations on them by *structural induction*.

In the *denotational semantics* introduced by Dana Scott, a programming language $\mathbb{P}$ is interpreted in a structure $(D, —)$ whose universe D is a complete poset, the *domain*. The points of D may include concrete *data* (words from some finite alphabet), but also functions of various sorts and complex mathematical structures which model computations, interactions, etc. For each correct expression A and each assignment α to the variables, the denotation $[\![A]\!](\alpha)$ is a point in D, determined by a structural induction of the following general form: first a (Scott-continuous) recursive equation (9) is constructed from α and the denotations of the parts of A, and then we take

$$[\![A]\!](\alpha) = \text{the least fixed point of } [x = \pi(x)].$$

The use of recursive equations is absolutely essential here, to interpret the iteration and recursive constructs which are at the heart of programming languages.

The *implementation* is a function which assigns to each program A a "machine" M_A—or, more concretely, code in the machine language of some processor—which *computes* the denotation $[\![A]\!]$ of A. In the simplest case, $[\![A]\!]$ might just be a sequence of external acts, like "printing" some file or drawing some picture on a monitor; more often $[\![A]\!]$ is a function relating input to output, or a "strategy" in some game, by which the machine responds to a sequence of external stimuli. As with inference systems, implementations come in a great variety of shapes and forms (*compilers* and *interpreters*, to name two), but they must have the basic *soundness property*, that M_A "computes" $[\![A]\!]$ in a well-understood way which relates the abstract (mathematical) denotations of programs to the behavior of machines.

Even with this grossly oversimplified description, it should be clear that the basic methodology of logic—the clean distinction between *syntax*, *semantics*, and *inference*—has had an immense influence on the development of programming languages; and that the fundamental, related notions of *symbolic computation* and *recursion* introduced by logicians in the 1930s are essential to the understanding of programming languages.

In the other direction, the study of programming languages—spurred by the need for applications—has introduced a host of interesting problems in logic, chief among them the question of *logic of programs*: What are the natural formal languages and inference systems in which the fundamental properties of programs can be expressed and rigorously proved? Much work has been done on this, but it is fair to say that the question is still open, and a formidable challenge to logicians and computer scientists.

VI. ALTERNATIVE LOGICS

From the many alternative logics which are obtained by changing the syntax, semantics or inference system of First-Order Logic, we consider, very briefly, just two.

A. Modal and Temporal Logic

Modal Logic goes back to Aristotle, the traditional founder of logic, who took *necessity* as one of the basic linguistic constructs worthy of logical study. The modern syntax is obtained by adding to $\mathbb{FOL}$ the propositional *box operator* $\Box$, so that with each formula A we have the formula $\Box A$ (with the same free variables), read *necessarily* A. The *possibility* operator is defined by the abbreviation $\Diamond A \equiv \neg\Box\neg A$.

Modal formulas are interpreted in *Kripke structures*

$$M = (W, s_0, \{M_s \mid s \in W\}, R),$$

of a specified vocabulary σ, where W is some set of *possible worlds*; s_0 is a specified "actual world"; each M_s is a σ-structure associated with the world s; and $R(s, t)$ is an *accessibility relation* on the worlds, intuitively standing for "t is a possible alternative to s." There are no fixed, general assumptions about the accessibility relation or the interpretations of the given relations on the various worlds; it could be, for example, that "Mary is John's wife" in the actual world s_0, but in alternative possible worlds John's wife might be Ellen, John may not have a wife, or he may not even exist. Assign associate objects in the possible worlds to individual variables, and the basic, semantic relation $M_s, \alpha \models A$ is defined by the Tarski conditions (for structures) with the additional clause

$$M_s, \alpha \models \Box A \\ \Leftrightarrow \text{ for all } t, \text{ if } R(s, t), \text{ then } M_t, \alpha \models A.$$

For example, if R is *transitive*,

$$R(s, t) \,\&\, R(t, t') \Longrightarrow R(s, t'),$$

then, the formula

$$\Box A \to \Box\Box A \tag{11}$$

is satisfied by all assignments, in all possible worlds, while it may fail for some A in nontransitive structures. Finally,

$$M, \alpha \models A \Leftrightarrow M_{s_0}, \alpha \models A.$$

Different conceptions of "necessity" can be modeled by placing appropriate restrictions on the accessibility relation, for example that it be transitive, linear, etc., and there is a question of constructing a suitable inference system and proving the appropriate Completeness Theorem in each case. A great deal of interesting work has been done in this area, much of it motivated by puzzles in the philosophy of language.

If we take $W = \mathbb{N}$ for the set of possible worlds, with $s_0 = 0$ and $R(s, t) \Leftrightarrow s \leq t$, and if we read $\Box A$ as "*from now on A*," we get one version of Temporal Logic, very useful for applications to computing systems. The worlds are interpreted by the *states* of some finite state machine, the propositional variables stand for properties of states, and the propositional formulas (which suffice) can express interesting properties of the system, especially if we augment the language with some additional, natural primitives like Next with the truth condition

$$M_s \models \mathsf{Next}\, A \Leftrightarrow M_{s+1} \models A.$$

For example, $\Box(p \to \mathsf{Next}\, q)$ says that "every state which has property p is followed immediately by one which has property q," and $\Diamond\Box p$ says that "p will eventually become and remain true," both interesting properties of finite state machines. This *temporal logic is decidable*, and so are various extensions of it, in which essentially all interesting *liveliness* and *fairness* properties of finite state machines can be expressed, so that one can *mechanically verify* the "correct behavior" of finite state machines. The relevant algorithms are practical, if not simple, they are used commercially, and they provide a spectacular example of the emerging field of *applied logic*.

B. Intuitionistic Logic

First-Order Intuitionistic Logic $\mathbb{FOL}_I$ has the same syntax as $\mathbb{FOL}$, and almost the same inference system: we simply replace the *Double Negation Law* $\neg\neg A \to A$, Eq. (8) in Section I.E, by the weaker

$(8)_I$ $\neg A \to (A \to B)$.

Kripke has established a Completeness Theorem for $\mathbb{FOL}_I$ using a variation of his semantics for Modal Logic, and this is useful for obtaining unprovability results for $\mathbb{FOL}_I$. The language, however, is meant to be understood constructively, and so it is not really possible to explain its semantics fully within classical mathematics. Aside from philosophical concerns, the real interest of Intuitionistic Logic comes from the proof theory of $\mathbb{FOL}_I$, which, somewhat surprisingly, also has important applications to computer science. Some sample results:

(1) For any two sentences A and B,

$$\vdash_I A \vee B \Longrightarrow \vdash_I A \text{ or } \vdash_I B,$$

and hence $\not\vdash_I p \vee \neg p$.

(2) In *Heyting Arithmetic*, i.e., the axiom system PA of Section III.A with Intuitionistic Logic, for any sentence $(\exists x)A$,

$$\mathsf{PA} \vdash_I (\exists x)A \\ \Rightarrow \text{ for some } n, \mathsf{PA} \vdash_I A\{x :\equiv \Delta_n\}.$$

(3) If $\mathsf{PA} \vdash_I (\forall x)(\exists y)A$ and $(\forall x)(\exists y)A$ is a sentence (no free variables), then there is a computable function f, such that for all n,
$\mathsf{PA} \vdash_I A\{x :\equiv \Delta_n, y :\equiv \Delta_{f(n)}\}$.

This last result is obtained with Kleene's Realizability Theory, and it illustrates the following general principle: from a constructive proof of $(\forall x)(\exists y)R(x, y)$, we can extract an algorithm which computes for each x, some y such that $R(x, y)$. There are obvious applications of this idea in computer science, and much of the current research in intuitionism is motivated by it.

VII. SET THEORY

Sets are collections into a whole of definite and separate objects of our intuition or thought, according to Georg Cantor, who initiated their mathematical study in the mid 1870s. Thus the basic relation of the theory is *membership* ($\in$),

$$x \in A \Leftrightarrow x \text{ is a member of } A,$$

and a set is completely determined by its members,

$$A = B \Leftrightarrow (\text{for all } x)[x \in A \Leftrightarrow x \in B].$$

Finite sets can be simply enumerated, e.g., $A = \{0, 5, 7\}$. Infinite sets are usually specified by means of some condition $P(x)$ which characterizes their members, and we write

$$A = \{x \mid P(x)\}$$

to indicate that A "is the set of all objects which satisfy $P(x)$."

Cantor was led to the study of arbitrary, *abstract sets* in his effort to understand the structure of some specific sets of real numbers or *pointsets*, and the theory which he created still exhibits today these two related but separate concerns. The *theory of pointsets* or *descriptive set theory* is primarily a theory of definability on the real numbers, and it is characterized by its applications to other fields of mathematics, especially analysis. *Abstract set theory* is primarily a theory of *counting*, an extension of combinatorics to the transfinite. The best set-theoretic results are about the interaction between these two poles of the subject.

At about the same time as Cantor's original contributions, Gottlob Frege initiated an effort to create a *foundation of mathematics* on the basis of set theory. Frege's approach was different (he took "function" rather than "set" as his primitive notion) and his original program was overly ambitious and failed. He had the right basic idea, however, that all objects of classical mathematics can be "defined within set theory," so that their properties can be (ultimately) derived from properties of sets. It took some time for this to take hold, but it is fair to say that since the 1930s, set theory has been the official language of mathematics, just as mathematics is the official language of science. This richness of the field makes it fertile ground for logical investigations, and it is not an accident that logicians have been involved with set theory from the beginning.

A. Cardinal Arithmetic

There are exactly as many left shoes in a (normal) shoe store as there are right shoes—we can be sure of this without counting, because of the obvious one-to-one correspondence between left and right shoes. The principle here is that *equivalent sets have the same number of members*,

$$|A| = |B| \Leftrightarrow A \sim_c B, \tag{12}$$

where $A \sim_c B$ indicates that some one-to-one correspondence exists between the members of A and the members of B, and

$$|X| = \text{the number of objects in the set } X.$$

This is a basic tool in mathematics: we count a set A by establishing a one-to-one correspondence between its members and the members of some already-counted set B. Moreover, if we set

$$A \preceq_c B \Leftrightarrow \text{for some subset } C \subseteq B,\ A \sim_c C,$$

then, obviously,

$$|A| \leq |B| \Leftrightarrow A \preceq_c B, \tag{13}$$

and we can often prove indirectly that *there are objects in B which are not in A* by showing (using arithmetic) that $|A| < |B|$, so that $B \subseteq A$ is impossible.

Cantor proposed to associate a *cardinal number* $|X|$ with every (finite or infinite) set X, so that Eqs. (12) and (13) hold, and then to use similar counting and (infinite) *cardinal arithmetic* techniques in the study of arbitrary sets. One might expect problems, because a finite set cannot be equivalent with one of its proper subsets (by the so-called *Pigeonhole Principle*), while

$$\mathbb{N} = \{0, 1, 2, \ldots\} \sim_c \{0, 2, 4, \ldots\} \tag{14}$$

via the correspondence $f(n) = 2n$. Cantor showed that, despite this "paradox," his cardinal arithmetic is a powerful tool with important applications in almost all areas of mathematics.

Cantor's first fundamental discovery was that there are (at least) two infinite sizes of sets: if

$$\aleph_0 = |\mathbb{N}|, \quad \mathfrak{c} = |\mathbb{R}| = |\text{the real numbers}|$$

are the cardinal numbers of the two most basic sets in mathematics, then

$$\aleph_0 < \mathfrak{c}. \tag{15}$$

A set A is *countable* if $|A| \leq \aleph_0$, otherwise it is, like $\mathbb{R}$, *uncountable*.

To define the arithmetical operations on (possibly infinite) cardinal numbers, choose sets K, L with no members in common so that $\kappa = |K|$, $\lambda = |L|$, and set

$$\kappa + \lambda = |K \cup L|,$$

$$\kappa \cdot \lambda = K \times L,$$

$$\kappa^\lambda = |(L \to K)|.$$

Here the *union* $K \cup L$ is the set of all objects which belong to either K or L; the *Cartesian product* $K \times L$ is the set of all ordered pairs (x, y) with $x \in K$ and $y \in L$; and the *function space* $(L \to K)$ is the set of all functions $f : L \to K$. If κ and λ are finite, we get the usual sum, product and exponential, noting, in particular, that there are κ^λ functions from a set of size λ to one of size κ. Moreover, all the familiar arithmetical identities hold, e.g., addition and multiplication are associative and commutative, multiplication distributes over addition, $\kappa^0 = 1$, and

$$\kappa^{\lambda+\mu} = \kappa^\lambda \cdot \kappa^\mu, \quad (\kappa^\lambda)^\mu = \kappa^{\lambda\cdot\mu}.$$

As examples of "proofs by counting," Cantor showed first that

$$\aleph_0 + \aleph_0 = \aleph_0 \cdot \aleph_0 = \aleph_0 \tag{16}$$

(basically because of Eq. (14)), and

$$\mathfrak{c} = 2^{\aleph_0}.$$

Both of these facts are easy, but they support the computation

$$\mathfrak{c}^2 = 2^{\aleph_0} \cdot 2^{\aleph_0} = 2^{\aleph_0+\aleph_0} = 2^{\aleph_0} = \mathfrak{c},$$

which means that there is a one-to-one correspondence between the line and the plane, and, hence, between the line and real n-space, for every n. This was new, it was surprising, and it was proved by "plain arithmetic." Eventually it motivated the development of *dimension theory*, whose basic result is that there is no continuous, one-to-one correspondence of real n-space with real m-space unless $n = m$. Moreover, the set of rational numbers is countable, and so is the set of *algebraic numbers*, the solutions of polynomial equations

$$a_0 + a_1 x + a_2 x^2 + \cdots + x^n = 0$$

with integer coefficients. Thus, since $\mathbb{R}$ is uncountable, "by simple counting" *there exist transcendental* (not algebraic) *real numbers*, a famous result of Liouville's whose original proof had rested on delicate convergence arguments for infinite series. It was a "killer application" which made set theory instantly known (and somewhat notorious) in the mathematical community.

Cardinal addition and multiplication satisfy the following *absorption laws* which basically trivializes them in the infinite case:

$$\text{if } 0 < \kappa \leq \lambda \text{ and } \lambda \text{ is infinite,}$$
$$\text{then } \kappa + \lambda = \kappa \cdot \lambda = \lambda.$$

For exponentiation, however, Cantor extended Eq. (15) to the general inequality

$$\kappa < 2^\kappa,$$

which provides infinitely many distinct "orders of infinity," perhaps what people meant when they referred to *Cantor's Paradise*. Exponentiation is the source of the deepest questions about infinite sets, chief among them Cantor's *Generalized Continuum Hypothesis* (GCH), the claim that for all infinite κ,

$$\text{(GCH)} \quad 2^\kappa = \kappa^+ = \text{the least cardinal number } > \kappa.$$

The "ordinary" case (CH) $2^{\aleph_0} = \aleph_0^+$ was No. 1 in Hilbert's list, it dominated set-theoretic research in the 20th century and, in a sense, it is still open today.

In addition to the cardinal numbers, which count the members of a set one, two, three, ... in the finite case, Cantor also introduced infinite versions of the *ordinal numbers* first, second, third, ... which assign position in a sequence. These are associated with "transfinite sequences," i.e., *well-ordered structures* $(A, \leq)$, where $\leq$ is a *linear ordering* on A (so that $x \leq y$ or $y \leq x$, for all x, y in A) and *every non-empty subset of A has a least element*. Every ordinal number α has a *successor* $\alpha + 1$ which defines "the next position," and every set of ordinal numbers A has a *least upper bound* $\sup A$. The least infinite ordinal number ω defines the first position with infinitely

$$0,\ 1,\ 2, \ldots \omega,\ \omega + 1,\ \omega + 2, \ldots \omega \cdot 2,\ \omega \cdot 2 + 1, \ldots$$

many predecessors, and it is a *limit ordinal*, without an immediate predecessor.

Ordinal arithmetic has fewer direct applications than the arithmetic of cardinal numbers, but well-ordered structures and ordinal numbers are the fundamental tools in the study of *transfinite iteration*, which is rich in applications. In a typical case, a function $f : A \to B$ is defined by *recursion* on some well ordered structure $(A, \leq)$, and then the crucial properties of f are established by *induction* along $\leq$. Moreover, the exact specification of the relation $\leq$ is often unimportant: all that matters is that *some* relation well orders A, in other words that A be *well orderable*.

$$\text{(WOP)} \quad \textit{Is every set well orderable?}$$

Specifically, is the set $\mathbb{R}$ of real numbers (where many of the applications lie) well orderable? The natural ordering of $\mathbb{R}$ won't do, since (for example) $\mathbb{R}$ has no least element, and it is hard to imagine how one could arrange all the real numbers into a transfinite sequence, with each point followed by its successor and every nonempty subset having a least element. The *Continuum Problem* (whether CH is true or not) and this *Well-Ordering Problem* were the central open problems in set theory at the turn of the 20th century.

B. The Paradoxes

Cantor developed his theory on the basis of the following *General Comprehension Principle* which flows naturally from his "definition" of sets quoted in the beginning of

this section: every definite (unambiguous) property $P(x)$ of mathematical objects, has an extension, the set $A = \{x \mid P(x)\}$ which "collects into a whole" all the objects which satisfy $P(x)$, so that

$$x \in A \Leftrightarrow P(x). \tag{17}$$

But this is not generally true: because if

$$R = \{x \mid \text{ is a set and } x \notin x\},$$

then, from Eq. (17),

$$R \in R \Leftrightarrow R \text{ is a set and } R \notin R \Leftrightarrow R \notin R$$

which is absurd. The argument was discovered in 1902 by Bertrand Russell, and it was not the first contradiction in set theory. However, earlier "paradoxes" (some of them known to Cantor) were technical, not unlike the paradoxes with infinitesimals which had been commonplace in calculus some years earlier, and it was thought that they would go away in a careful development of the subject. The *Russell Paradox* is not technical, it goes to the heart of the nature of sets, and it threw the mathematical community into a spin.

L. E. J. Brouwer initiated the *intuitionistic program* which denies that abstract sets are meaningful objects of study and also rejects some of the basic principles of logic. Mathematical objects cannot be said to "exist" in any sense independent of (mental) "mathematical activity"; and to prove that *some x has property P*, one must *construct* some specific object x which has property P. It is not enough to derive a contradiction from the assumption that *no x has property P*. Intuitionism had a strong influence in the philosophy of mathematics and remains a vibrant field of study within logic, but it never carried much favor with mathematicians: too much of classical mathematics must be thrown out to satisfy its tenets.

Hilbert proposed to "save" classical mathematics from the paradoxes and Brouwer's attack by formalizing as large a part of it as possible in some first-order, axiomatic theory T, and then establishing the consistency of T by absolutely safe, *finitistic* methods. *Formalism* is the reading of *Hilbert's Program* as a philosophical view: it alleges that once T is chosen, then T is all there is—there is nothing more to mathematics but the study of the inference relation $T \vdash A$, with no reference to meaning. Aside from the impact of Gödel's Second Incompleteness Theorem (Section IV.E) which weakens it, formalism also fails to account for the applications of mathematics: it is hard to see how the existence or not of certain patterns of meaningless symbols can have any bearing on the escape velocity of a rocket.

From those reluctant to abandon the traditional, *realist* view that mathematical objects are, well, *real*, no matter how abstract and difficult to pin down, Russell first proposed to replace set theory by his famous *theory of types*: it is claimed (roughly, and in the later *simple* version due to Ramsey), that every mathematical object is of a certain (natural number) *type n*, and that every set A is of some successor type $n+1$, such that the members of A are of the immediately preceding type n. Type theory is awkward to apply and it yields only a poor shadow of Cantor's set theory, albeit without the paradoxes. It never gained favor as a true alternative to set theory, although it has been studied extensively as a logical system, it has found its own applications (especially recently, to programming languages), and many of its fundamental ideas were eventually incorporated in the reformulation of set theory which eventually prevailed.

What has prevailed is *Axiomatic Set Theory*, first proposed in 1904 by Zermelo as a pragmatic way to avoid the paradoxes by rebuilding Cantor's set theory on the basis of a few set-theoretic principles which are basic, simple, and well understood by their uses in classical mathematics. Formalists can accept it, as nothing more but the choice of a specific set of axioms, whose "truth" is irrelevant, if, at all, meaningful. But it is the realists who, in the end, have received the greatest comfort from axiomatic set theory: because the systematic development of consequences of the axioms eventually led to a narrower, more concrete concept of *set*, which ultimately justified the axioms.

Much of modern logic was created in response to the challenge of the set-theoretic paradoxes, and that is another reason why the discipline is so intimately tied with set theory.

C. Zermelo-Fraenkel Set Theory

There are eight axioms in ZFC (Zermelo-Fraenkel Set Theory with Choice), and it is assumed that they are interpreted over some given domain of *sets* $\mathbb{V}$, which comes endowed with a binary *membership* condition, $x \in y$. The formal theory ZFC is obtained by expressing these axioms by sentences of $\mathbb{FOL}(\in)$, and it requires infinitely many sentences, because the Replacement Axiom 5 requires an *axiom scheme*. Here we will describe them briefly and informally, with a few interspersed comments.

1. *Extensionality*. Two sets are equal exactly when the have the same members.
2. *Empty set and Pairing*. There is a set $\emptyset$ with no members, and for any two sets a, b, there is a set $\{a, b\}$ whose members are exactly a and b.
3. *Unionset*. For each set A, there is a set $\cup A$ whose members are the members of the members of A,

$$t \in \cup A \Leftrightarrow (\exists x)[x \in A \,\&\, t \in x].$$

4. *Powerset*. For each set A, there is a set $\mathcal{P}(A)$ whose members are all the subsets of A.

An operation $F : \mathbb{V} \to \mathbb{V}$ is definite if it is first-order definable with parameters, i.e.,

$$F(x) = G(x, a_1, \ldots, a_k)$$

where $G(x, y_1, \ldots, y_k)$ is first-order definable, Section II.C.

5. *Replacement*. The image

$$F[A] = \{F(x) \mid x \in A\}$$

of a set A by a definite operation F is a set. (This was formulated in the 1930s, primarily by Skolem, and it is much stronger than Zermelo's original *Separation Axiom*.)

For the next two axioms, we need the notion of *function* $f : A \to B$ from one set to another, which is not among our primitives, and so we need to "reduce" the notion of function to that of set. The trick is well known: first we fix some *ordered pair operation* (x, y) which satisfies the key property

$$(x, y) = (x', y') \Leftrightarrow x = x' \,\&\, y = y', \qquad (18)$$

and then we model a function f by its graph,

$$G_f = \{(x, y) \in A \times B \mid y = f(x)\},$$

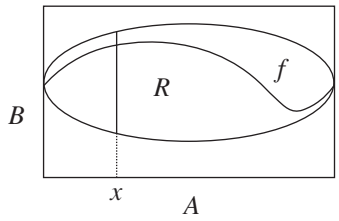
which is just a set with some special properties. It is common to use the so-called *Kuratowski pair operation*

$$(x, y) = \{x, \{x, y\}\},$$

but there are many others, and all that is needed is some operation which satisfies Eq. (18).

6. *Infinity*. There is a set I and a one-to-one function $f : I \to I$ which is not onto I, $f[I] \subsetneq I$.

Next comes Zermelo's chief contribution:



7. *Axiom of Choice* (AC). For each binary relation $R \subseteq A \times B$,

$$(\forall x \in A)(\exists y \in B)(x, y) \in R$$
$$\Rightarrow (\exists f : A \to B)(\forall x \in A)(x, f(x)) \in R.$$

In effect, AC postulates a function f which makes a choice $f(x)$ from the nonempty set $\{y \mid (x, y) \in R\}$, "simultaneously," for each $x \in A$. If B carries a well ordering $\leq$, we could take

$$f(x) = \text{the } \leq\text{-least } y \text{ such that } (x, y) \in R;$$

Zermelo showed that, conversely, AC *implies that every set is well orderable*, and identified numerous examples where the seemingly controversial AC is routinely used in mathematics. Somewhat later Hartogs showed that AC is also equivalent with the *cardinal comparability property*

$$(\forall A, B)[A \leq_c B \vee B \leq_c A]$$

without which there is no cardinal arithmetic, and this limited further opposition to AC to those who were willing to abandon completely Cantor's Paradise.

The last axiom of ZFC involves the *cummulative hierarchy of sets*, which is defined by recursion on the ordinal numbers as follows:

$$V_0 = \emptyset,$$
$$V_{\alpha+1} = \mathcal{P}(V_\alpha),$$
$$V_\lambda = \bigcup_{\alpha<\lambda} V_\alpha \quad (\text{if } \lambda \text{ is limit}).$$

8. *Foundation*. Every set is a member of some V_α.

This is a limiting axiom, not needed for the development of Cantor's set theory or its applications, but it is important because it codifies within the axiomatic theory a conception of set which replaced in the 1930s Cantor's free-wheeling notion of a "collection into a whole": each set is reached starting with "nothing" (the emptyset $\emptyset$), by "indefinite" (never ending) "iteration" of the powerset operation. Admittedly more complex than Cantor's, this notion of *grounded set* prohibits the circular constructions which lead to the paradoxes, and it can be described intuitively in sufficiently clear terms to justify the axioms.

To see how classical mathematics can be developed on the basis of these seven axioms, consider first arithmetic. A *number system* is a triple $(\mathbb{N}, 0, S)$ such that $\mathbb{N}$ is a set, $0 \in \mathbb{N}$, $S : \mathbb{N} \to \mathbb{N}$ is a one-to-one function which is never 0, and

$$[X \subseteq \mathbb{N} \,\&\, 0 \in X \,\&\, S[X] \subseteq X] \Rightarrow X = \mathbb{N};$$

we prove that there exists a number system and that every two number systems are isomorphic, and then we choose some specific number system and call its members *the natural numbers*. The real numbers are identified with some *complete ordered field*, once we prove that one such exists and any two are order isomorphic, and so forth for other structures. This process of "defining" (more accurately: *modeling faithfully*) mathematical structures in set theory has found such widespread acceptance in mathematics, that "to make a notion precise" is now viewed as synonymous with "defining it in set theory."

D. Independence Results

It is, perhaps, ironic, that the axiomatization of set theory made possible to formulate and prove its own limitations. Let ZF be the theory with axioms 1–6, i.e., without the Axiom of Choice:

Theorem. If *ZF* is consistent, then so are the theories ZFC+GCH (Gödel, 1938) and ZFC+¬CH (Paul Cohen, 1963).

In effect, ZFC can neither disprove nor prove the Continuum Hypothesis, unless a contradiction can be obtained from its "constructive" core. In addition, Cohen showed that *ZF* cannot prove the Axiom of Choice, and several additional consistency and independence results.

Gödel's proof uses an *inner model*, a sub-collection of our intended universe of sets $\mathbb{V}$: using only axioms 1–6, he defines a certain collection L of *constructible sets* and shows that if we reinterpret "set" to mean "member of L," then all the axioms of ZFC as well as GCH are true. Cohen's *forcing method* builds "virtual universes" which are "larger" than $\mathbb{V}$, and so he must describe them indirectly. This can be done with *Boolean-valued models*: a collection $M \subset \mathbb{V}$ and a binary condition E on M are defined, and then it is shown that, for a certain (complete) Boolean algebra B, the *Boolean semantics* of the "structure" (M, E) assign 1 to all the theorems of ZFC but something other than 1 to CH.

In both of these proofs, logic plays an essential role which goes much beyond providing the context in which their claims can be made precise. For example, the constructible universe L is defined by iterating the operation of taking all *first-order definable* subsets rather than $\mathcal{P}(A)$ in the cummulative hierarchy of sets, and then a strong version of the Skolem-Löwenheim Theorem is used at a crucial point to show that GCH holds in L. Through the work, initially, of Robert Solovay for forcing and Ronald Jensen for constructibility, these theories have been much generalized and continue to be very active research areas of logic, with important applications to analysis, algebra and topology.

E. Current Research in Set Theory

In one direction, set theory is more involved now with applications than ever before. Especially fruitful has been the development in the 1960s of *effective descriptive set theory*, which incorporates methods from recursion theory into the study of definability on the continuum to yield very substantial applications to analysis.

Beyond the applications, set theory has attempted to confront the fundamental problem posed by the independence results: what does one do with the Continuum Problem, now that we know that it cannot be settled in ZFC? Some have adopted a formalist view, that it is meaningless to ask "whether CH is true or not," and that "set theory is the study of all models of ZFC." This is a very active area of research.

In another direction, people have looked for *new axioms*, extending ZFC, which might provide the needed answers, and a great deal of research has been done in this direction since the 1960s. Generally speaking, two kinds of axioms have been considered. *Large cardinal axioms* are plausible generalizations of the Axiom of Infinity, which, however, have very few direct consequences for the continuum. *Determinacy hypotheses* postulate that certain (fairly simple) infinite games on the natural numbers are *determined*; somewhat technical and not especially plausible, these axioms answer most definability questions about the real numbers that are independent of ZFC, although, unfortunately, they cannot settle the Continuum Problem. In a fundamental advance made in the 1980s, Donald A. Martin, John Steel, and Hugh Woodin showed that the plausible large cardinal axioms imply the fruitful determinacy hypotheses, and so a "unified," very strong extension of ZFC has been created which is the subject of much current research. Unfortunately, it does not solve the Continuum Problem, and so the search goes on.

It may well be that set theory will continue to be dominated in the 21st century by the search for an answer to the Continuum Problem, as it certainly was during the century just ended.

SEE ALSO THE FOLLOWING ARTICLES

BOOLEAN ALGEBRA • COMPUTER ALGORITHMS • DATABASES • FUZZY SETS, FUZZY LOGIC, AND FUZZY SYSTEMS • SET THEORY

BIBLIOGRAPHY

Abramsky, S., Maibaum, T. S. E., and Gabbay, D. M., eds. (1993). "Handbook of Logic in Computer Science," Clarendon Press, Oxford.

Buss, S. R., ed. (1998). Handbook of Proof Theory. *In* "Studies in Logic and the Foundations of Mathematics," Vol. 137, Elsevier, Amsterdam/New York.

Hodges, W. (1993). Model Theory. *In* "Encyclopedia of Mathematics and Its Applications," Vol. 42, Cambridge Univ. Press, Cambridge, U.K.

Rogers, Jr., H. J. (1967). "Theory of Recursive Functions and Effective Computability," McGraw-Hill, New York.

Kunen, K. (1998). Set Theory. *In* "Studies in Logic and the Foundations of Mathematics," Vol. 102, Elsevier, Amsterdam/New York.

Moschovakis, Y. N. (1980). Descriptive Set Theory. *In* "Studies in Logic and Foundations of Mathematics," Vol. 100, North Holland, Amsterdam.



Mathematical Modeling

Xavier J. R. Avula
University of Missouri, Rolla

I. Introduction
II. Mathematical Modeling
III. Classification of Mathematical Models
IV. Formulation of Mathematical Models
V. Solution Techniques
VI. Model Validation
VII. Chaos and Complexity
VIII. Modeling with Neural Networks
IX. Mathematical Modeling and Computers
X. Concluding Remarks

GLOSSARY

Chaos Irregular and unpredictable system behavior that exhibits sensitive dependence on initial conditions in deterministic physical systems modeled by nonlinear equations.
Complexity Complexity is the collective behavior of a system containing many interacting subsystems that cannot be explained in terms of the observable entities.
Mathematical modeling Entire process of representing real-world phenomena in terms of mathematical equations and achieving their validity in an iterative manner.
Model Idealized representation of a real-world phenomenon. In the abstract sense, it is a set of instructions for generating behavioral data.
Modeling Act of relating real systems to models.
Artificial Neural Networks Networks of artificial neurons that are interconnected in different patterns to process information in a parallel distributed fashion much like the human brain.
Parameter identification Experimental determination of values of parameters that govern the system behavior.
Simulation Process of relating computers to models generating behavioral data.
System Set of elements plus a collection of rules governing the interrelationships between the elements and the behavior of elements over time.
System identification Determination of a system from its input and output time history.

MATHEMATICAL MODELING is a comprehensive process of representing real-world phenomena in terms of mathematical equations and extracting from them useful information for understanding and prediction. In recent

years, mathematical modeling has become a powerful tool to solve complex, interconnected, and interacting phenomena arising from the rapid developments taking place in science and technology. The success in physical sciences in terms of valid mathematical models has led scientists to extend the modeling methodology to other emerging fields of inquiry in which great strides have been made. The explosive growth of mathematical modeling activity has been a driving force behind the development of high-speed digital computers, which in turn aided model solutions in a symbiotic relationship.

I. INTRODUCTION

Modern civilization has its roots in human endeavor to understand the physical universe. The effort to systematically understand the universe and the various phenomena in it appears as a never-ending struggle imbibed with frustration and romance that sparked the imagination of humankind in the course of history. The birth of the scientific method and the ensuing pursuits have lead men and women of learning to the study of phenomena systematically and produced a vast edifice of knowledge in sciences and mathematics. Galileo (1564–1642), the famous Italian astronomer and physicist, firmly enunciated that the language of science is mathematics. The German philosopher Immanuel Kant (1724–1804), in the preface to his book "Metaphysical Foundations of Science," declared "that each particular discipline contains only as much science as it contains mathematics." No more words need to be wasted to say that science and mathematics are intertwined. If science and mathematics are intertwined, can technology, the child of science, exist without mathematics?

In striving to understand phenomena, one must not forget the Aristotelian adage, "The primary question is not what do we know but how do we know it." What we know and its reliability should be the consequence of the process that answers the question how do we know it.

The famous mathematician Hilbert, after a sojourn in the chemistry department at the University of Göttingen, said that chemistry is too difficult for chemists. The American mathematician Richard Bellman echoed that medicine is too difficult for physicians, politics is too difficult for politicians, and economics is too difficult for economists. So is engineering for engineers and physics for physicists. Then how can they be enlightened? The enlightenment comes from exploring the behavior of mathematical models of the processes they encounter in their respective endeavors.

The physical universe is an abundant source of wonder. Benevolently exploited, it is a life-enhancing system for all humankind. The regularity observed in the natural processes of the universe is best expressed and explained in mathematical terms. Mathematical description has provided the tools and motivation for numerous discoveries in cosmology and atomic phenomena, as well as in biology, material science, earth sciences and the social behavior of animal and human populations. In engineering and technology, mathematical concepts and analyses have contributed greatly to the understanding, design, and operation of complex systems. The cost and risk involved in testing a real physical or engineering system are usually prohibitive. The alternative is to develop a mathematical model of the system and investigate its performance. How many of us would "pay the price for reaching the sun and learning its shape, its size, and its substance?"

II. MATHEMATICAL MODELING

Derived from its Latin root modus, the word "model" is generally understood to stand for an object that represents a physical entity with a change of scale. For example, a model airplane is a scaled-down version of a "real" airplane by a few orders of magnitude. As any airplane model builder knows, the behavior of the model and the real airplane differ in more ways than one, in spite of their physical similarity, ensuring a missing ingredient—something that has fallen out during the model building process. What then is a mathematical model? A mathematical model is a set of mathematical equations representing a process or a system. It is a mathematical idealization of a real-world phenomenon. In the sense of the model just defined, it represents a change on the scale of abstraction. Also, the way one sees the world depends upon the structure of one's language; different languages give rise to different concepts, and to many "alternate realities." In the process of idealization, some simplifications will have been made in obtaining the mathematical model. Therefore, the mathematical model is less real than the system it is supposed to represent; it is a mathematical representation of the modeler's perception of some aspects of reality within the confines of a formal mathematical system. Nevertheless, it is an essential step in the construction of a theory. To quote Boltzmann, there is nothing as practical as a good theory.

In the process of mathematical modeling, the objective of the modeler is, in general, to construct a model of an observed phenomenon and use the model to predict its future course. In some cases, modeling is used to explain the known facts and lay a foundation for the theory behind the phenomenon. Thus, a model is a mathematical manifestation of a particular theory. Some modelers have an altruistic motive to apply the model characteristics to

interpret the behavior of phenomena outside their discipline, recognizing the underlying commonality of mathematical modeling. While this commonality is held in high esteem, some fundamental issues underlying the theory of modeling must be addressed. Some of the issues are:

How can a formal mathematical system represent a natural phenomenon?
Does the similarity of two natural systems imply that their models are similar?
How can multiple models of the same natural phenomenon be compared, and with what measure?
How to identify key observable parameters that are necessary for the construction of a plausible model?
Under what conditions the relationship between observables can be accepted as a "Law of Nature"?
How can multiple systems that behave similarly be considered as models of the same?
How is a given system related to its subsystems?
What features characterize "good" models?

Attention to these issues will allow mathematical modeling to play an essential part in the formation of scientific theories of phenomena. Scientific literature abounds with the view that the process of mathematical modeling is not unique. In some cases, the mathematical modeling procedure begins with the desire to construct a model from a set of observations made on the behavior of a system. According to the British physicist Maxwell, "the success of any physical investigation depends upon the judicious selection of what is to be observed as of primary importance." For many problems in science and technology, no wellsprings of observations exist to serve as a source in the modeling process. For example, the phenomena such as the behavior of a spacecraft in the environment of an unknown planet, the response of a structure to an earthquake, and the long-term ecological effects of an industrial effluent injected into the biosphere have no groundwork of observations. In such situations, the model-building process is originated in imagination modulated by intuition and experience. The schematic diagram of Fig. 1 depicts one of several paths in the process of mathematical modeling.

In Fig. 1 the step from A to B involves to a great extent the elements of intelligence and creativity of the modeler. In bringing out a well-defined problem from the maze of observations, the modeler shows the ability to correspond different entities—in this case, a real situation and a formalized statement of the problem. Mathematical training alone will not give an edge to the modeler at this stage. Experience, intuition, and other nonmathematical skills play an important role.

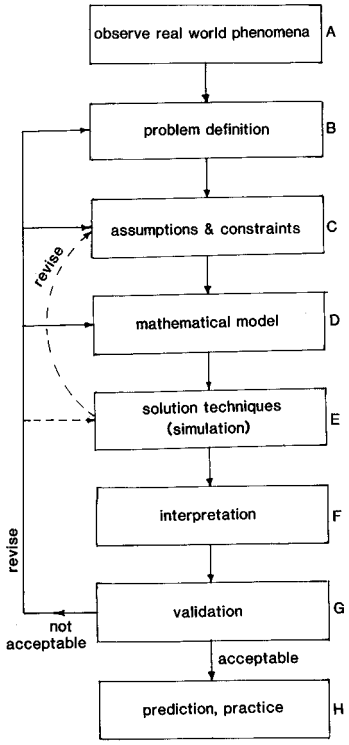


FIGURE 1 Schematic of a mathematical modeling procedure.

Once the problem is defined, a mathematical model of the real-world phenomenon is to be constructed. The real-world phenomena are exceedingly complex. It is nearly impossible to construct a model that replicates the phenomenon in its totality. The modeler must sort out the essential and significant features that need to be incorporated in the model. Once again, experience and intuition come into play. The modeler translates the features into mathematical entities and relates them under certain simplifying but realistic assumptions and constraints. Now we have a mathematical model.

A word of caution is in order. While the application of well-known equations of physics such as Newton's laws, Lagrange's, Hamilton's, and Schrödinger's equations result in mathematical models of various physical processes in the form of differential equations, there are other prescriptions such as simulation and laboratory experimentation augmented by system identification techniques for formulating models of phenomena.

A myriad of mathematical solution techniques are available to extract the system behavior from the model. By using one or more of these techniques, the solution of the model is obtained and interpreted from the viewpoint of accuracy and stability. Determination of how closely the solution approaches the original, real-world phenomenon is the next step in the modeling process. If the solution meets the imposed limits of acceptability, the model is

considered valid and then put into practice to predict a future event, or to make a decision. Otherwise, the model is invalidated, and steps B–D are repeated by revising one or more ingredients in the process. Even a scrutiny of the solution technique (step E) is worth considering. Thus the process of mathematical modeling is iterative in nature. There is no uniqueness in the model-building process.

Although the steps in the modeling process are subjective, they are somewhat similar. Figures 2 and 3, extracted from different sources, essentially have the same features presented in Fig. 1, but indicate the role of computers in mathematical modeling rather explicitly.

In recent years, the growth in mathematical modeling has been so phenomenal that it is evolving into a discipline in its own right. The offering of courses on mathematical modeling in the colleges in the United States, Europe, Asia and Australia has been on a steady increase. The international conferences and symposia on mathematical modeling and the journals that deal exclusively with mathematical modeling support this viewpoint. The proliferation of mathematical models in the scientific literature can be attributed to the explosive growth in the application of mathematical and empirical knowledge to the problems of science and technology. Even social sciences and business, which are not traditionally subjected to mathematical treatment, are not spared from the productiveness of mathematical modeling. The catalysts in the growth of mathematical modeling have been

1. The advent of the computer age with rapid developments in computer technology, specifically, the developments in speed, memory, and software
2. The developments in numerical techniques and analysis
3. The developments in systems theory and simulation
4. The progress in empirical knowledge resulting in a greater understanding of various physical processes

The role of experience and intuition in the process of mathematical modeling accord to it an air of art that makes the process subjective. However, the commonality of features in the process of mathematical modeling in all sciences and technology is so overwhelming that the subjectivity is cast as a minor essence. Let us now turn our attention to these common features that enhance the objective view of mathematical modeling.

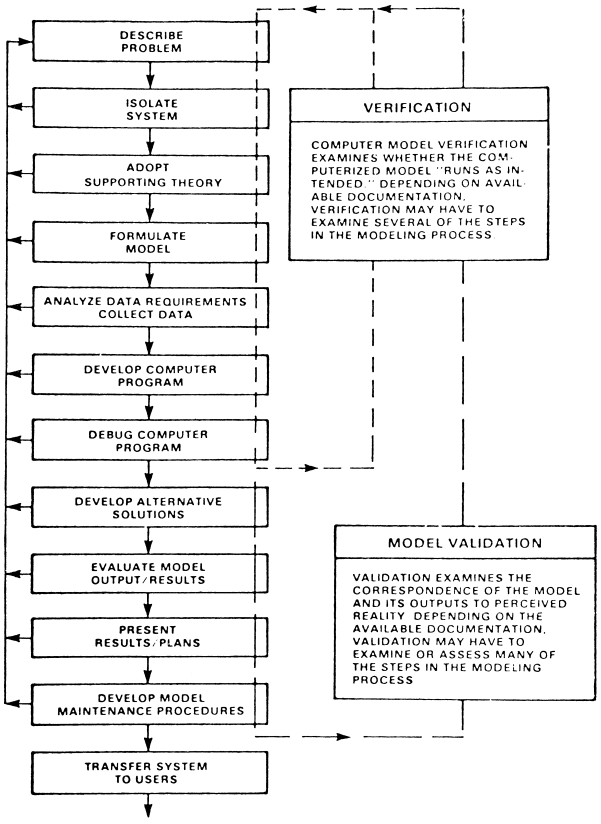


FIGURE 2 Basic steps in the modeling process. [From United States General Accounting Office (1979). PAD70-17, Guidelines for Model Evaluation, Washington, DC.]

III. CLASSIFICATION OF MATHEMATICAL MODELS

Classification of models is important to the understanding of a system's behavioral data and to interpreting or communicating the data for prediction and decision making. An understanding of the system behavior is a contribution to science, and its use for prediction and decision making constitutes technology. Mathematical models are broadly classified into dynamic and steady-state models. Obviously, the element of time is present in the dynamic models. The dynamic models are further classified into continuous-time and discrete-time models. In a continuous-time model the time advances smoothly through real numbers, while in the discrete-time model the time advances in finite jumps, each jump representing a multiple of a selected time unit.

Models constructed in terms of descriptive state variables are classified into continuous state, discrete-state, and mixed-state models. In a continuous-state model the range sets of state variables are presented by real numbers, while in a discrete-state model the variables assume a discrete set of values. In a mixed-state model both kinds of variables are present.

The broad class of continuous-time models is further subdivided into models in which the state changes continuously and those in which the state changes in discrete

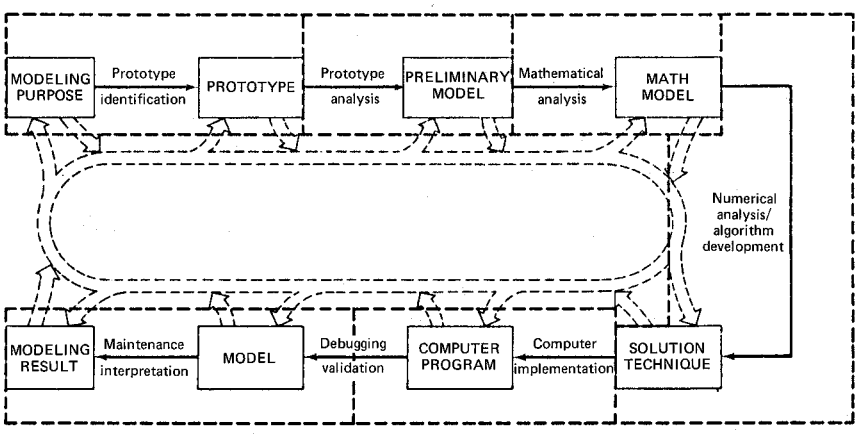


FIGURE 3 Guide to model building. [From Jacoby, L. S., and Kowilik, J. S. (1980). "Mathematical Modeling with Computers," Prentice-Hall, Englewood Cliffs, NJ.]

jumps. The former subclass of models are represented by ordinary and/or partial differential equations, for example, for the boundary-layer flow over a wing of an aircraft. On the other hand, the production line in a factory is a discrete-state system.

Yet another classification breaks down mathematical models into stochastic (or probabilistic) and deterministic models. Models with at least one random variable in their description are termed stochastic. Models other than stochastic are deterministic. In fact, deterministic is a special case of stochastic.

Another classification of models is based on how the real system interacts with the environment. If the real system is isolated from its environment, the model is called *autonomous*. If the environment exerts any kind of influence through so-called input variables from the environment without being controlled by the model, then the model is called *nonautonomous*.

When models are expressed in terms of mathematical equations, the solutions are profoundly affected by nonlinearity. Based on the type of equations, the models are classified as linear or nonlinear. These classifications can be combined with the above-mentioned model types to yield, for example, linear or nonlinear dynamic and linear or nonlinear stochastic models. Although the universe is predominantly nonlinear, some processes are linear within certain bounds. Actually, linearity can be viewed as a special case of nonlinearity. Sometimes, it is mathematically expedient to consider nonlinear processes as piece-wise linear. Systems are classified as linear if and only if they simultaneously satisfy the principle of homogeneity and the principle of superposition. The principle of homogeneity preserves the input function scale factor in transition from input to output. A system satisfies the superposition principle when the superposition of individual input functions result in a response that is the superposition of corresponding outputs. Simulations and exact solutions of linear systems of equations are well established and reported in mathematical literature.

All systems that are not linear belong to the class of nonlinear systems. As there are no general methods of solutions to solve and analyze models of nonlinear systems, simulation has become a commonly used tool for this purpose. Impacted by recent developments in numerical analysis and by advances in powerful, high-speed and high-memory digital and analog computers, simulation of nonlinear systems have blossomed into technical endeavors that are broad in scope and variety encompassing biological, ecological, social, and economic systems in addition to those in hard scientific fields such as physics, chemistry, material science, and mechanics. Modern nonlinear equations frequently encountered in simulation include, for example, the quaternion rotational equations of motion that are treated with geometric algebra which is of late applied to a range of problems in many fields of science and being promoted as a unified mathematical language in the 21st century.

Traditional quantitative techniques of modeling cannot be effectively applied to model phenomena in the so-called "soft" disciplines consisting of humanistic systems such as social, economic, ecological, and biological systems because there may arise difficulties associated with multidimensionality, subsystem interactions, inexplicable feedback mechanisms, hierarchical structures, and unpredictable behavioral dynamics. Such difficulties may also be encountered in mechanistic systems that may include some physical and engineering systems. These difficulties lead to making imprecise statements, introducing fuzziness much like human reasoning, about the characteristics and behavior of a system. An important class of models

using fuzzy sets was launched about 35 years ago. The idea of fuzzy sets is centered on the imprecision in the belonging of elements to a set. Intuitively, a fuzzy set is a set of elements that are not precise and in which there is no distinct boundary between the elements that belong to the set and those that do not. In other words, the transition from full membership to nonmembership of an element (or elements) is blurred, so that a gradation of partial membership is possible. Mathematical concepts based on this idea have been successfully applied to modeling of systems in "soft" desciplines as well as in engineering when imprecision is present. Fuzzy systems models are categorized into two types: liguistic and rule-based. An elaborate discussion of these types of models and modeling techniques is beyond the scope of this article.

IV. FORMULATION OF MATHEMATICAL MODELS

A. Conventional Modeling (Direct Modeling)

Experimental observations and measurements are generally accepted to constitute the backbone of physical sciences and engineering because of the physical insight they offer to the scientist for formulating the theory. The concepts that are developed from the observations are used as guides for the design of new experiments, which in turn are used for validation of the theory. Thus, experiments and theory have a hand-in-hand relationship. The information gathered during observation and measurement is usually presented in terms of curves, tables, block diagrams, circuit diagrams, flow diagrams, etc. for convenience of perception in the model building process. These information display techniques have been tremendously aided by the advent of high-performance computers.

Formulation of theory is equivalent to model building. Cast in mathematical terms, the theory stands as a mathematical model of reality. In the conventional sense, the first stage in model building is to propose and gather equations representing all relevant mechanisms of the phenomenon under study. Because of the diversity in the types of information available to the model builder, the equations representing the basic phenomenon may be extensive with complex interrelationships that may be difficult to untangle for easy tractability. To preserve fidelity, all equations that will have significant effect on the system behavior must be included in the model. It is always better to first produce a simple model and then refine it until it closely represents the reality, than to start with a complex model and simplify it for fear of facing mathematical difficulties.

Conventionally, the construction of mathematical models for physical and engineering phenomena begins with the application of physical laws—Newton's laws, Maxwell's laws, Kirchhoff's laws, and balance laws, which include mass balance, energy/heat balance, momentum balance, impulse balance, and entropy balance—to the phenomena being studied. In these laws, a number of relationships between the variables are expressed in terms of ordinary differential equations, partial differential equations, and difference equations. A detailed presentation of these laws is beyond the scope of this article. However, some examples are appropriate.

The first and second laws of thermodynamics in differential form are stated as

$$dU = dQ - d\mathrm{W} \tag{1}$$

$$d\mathrm{S} = dQ/T, \tag{2}$$

where U is the internal energy, Q is the heat absorbed by the system, W is the work done by the system, S is the entropy, and T is the temperature.

The fundamental law of heat conduction in one dimension is represented by the ordinary differential equation

$$dQ/dt = -k\mathrm{A}\,(dT/dx),$$

where dQ/dt is the time rate of heat transfer across the area A, k is the thermal conductivity of the medium, and dT/dx is the temperature gradient.

Newton, in his famous "Principia," expressed the second law of motion in terms of momentum, which in symbolic form becomes

$$\mathbf{F} = d\mathbf{p}/dt, \tag{3}$$

where $\mathbf{p}$ is the linear momentum (product of mass and velocity mv). This equation can be written in the familiar form as

$$\mathbf{F} = d(m\mathbf{v})/dt = m(d\mathbf{v}/dt) = \mathrm{m}\mathbf{a}, \tag{4}$$

Attempts by scientists to describe the physical world led them to the formulation of various types of partial differential equations. The mathematical model for the one-dimensional wave phenomena is the hyperbolic equation

$$\partial^2 u/\partial t^2 = c^2(\partial^2 u/\partial x^2). \tag{5}$$

The phenomena of diffusion of heat in solids is modeled by the parabolic equation derived by Fourier in the form

$$\frac{\partial u}{\partial t} = \sigma \frac{\partial^2 u}{\partial x^2}. \tag{6}$$

The Fourier heat equation in three spatial dimensions and time modeled as

$$\frac{\partial u}{\partial t} = \sigma \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) \tag{7}$$

yields for steady state (with $\partial u/\partial t = 0$)

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0, \tag{8}$$

which is an elliptic equation, also called Laplace's equation.

An example from the conservation laws is the law of conservation of mass, which can be expressed in terms mathematical symbols as

$$(\partial\rho/\partial t) + \nabla \cdot (\rho \mathbf{v}) = 0, \tag{9}$$

where ρ is the density, $\mathbf{v}$ is the velocity vector, and ∇ stands for the operation $\lfloor \hat{i}(\partial/\partial x) + \hat{j}(\partial/\partial y) + \hat{k}(\partial/\partial z) \rfloor$ rectangular Cartesian coordinates with $\hat{i}$, $\hat{j}$, $\hat{k}$ the unit vectors along the coordinate axes x, y, z, respectively.

A universally accepted mathematical model for viscous fluid flow is the famous Navier–Stokes equation,

$$\rho(D\mathbf{v}/D\mathrm{t}) = -\nabla\rho - \nabla \times [\mu\nabla \times \mathbf{v}] + \nabla[(2\mu + \lambda)\nabla \cdot \mathbf{v} \tag{10}$$

where μ, is the coefficient of viscosity, λ is the second coefficient of viscosity, and

$$D\mathbf{v}/Dt = (\rho(\partial/\partial t) + \mathbf{v} \cdot \nabla)\mathbf{v}. \tag{11}$$

in which $\mathbf{v} \cdot \nabla$ is the convective derivative.

These and other aforementioned equations play a central role in mathematical modeling of a large variety of physical systems. In addition to differential and difference equations, there are other distinct types of equations—algebraic, integral, and functional equations—that appear prominently in mathematical modeling. Not all of these equations have closed-form, analytical solutions. The modeler has to resort to numerical and other approximation methods for solutions.

In science, the purpose of modeling has been primarily for research and understanding natural phenomena. In research, even though the immediate use of the model is not clear, one pursues it for gain in comprehension, for interpreting knowledge, and for formulating clues for further investigation. In engineering, however, in addition to research, mathematical models are used for design and control. Here the knowledge of the components of the total engineering system has to be expressed in a model compatible with design criteria, which may include some or all of stability limits, error criterion, economic yield, and safety criterion. Also, modeling for design is not just for creating a new system but also for the adaptation of an existing system for a higher (or different) performance.

To drive an engineering system in a desired fashion, some control action needs to be imposed in the form of feed-back or feed-forward control, or adaptive control. There is a wide spectrum of human-made control systems in engineering, ranging from a simple on-off control to sophisticated computer control of a spacecraft. A common feature of all control systems is the so-called uncertainty factor which might be limited by using adaptive or self-optimizing principles.

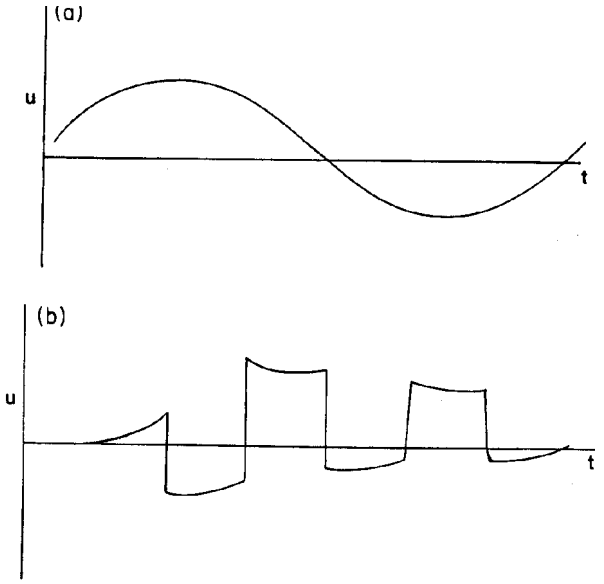


FIGURE 4 (a) Response u in Van der Pol equation for small values of the parameter 8. (b) Response ii in Van der Pol equation for large values of the parameter 8.

Even if the mathematical model representing a system is formulated in principle, it cannot be used for any practical purpose unless the range of parameters in the model is available. The importance of determining the range of parameters can be observed, for example, in the behavior of Van der Pol equation, which can be expressed as

$$(d^2u/dt^2) + \lambda(u^2 - 1)(du/dt) + u = 0. \qquad \lambda > 0 \tag{12}$$

As it can be seen in Fig. 4(a), u is almost periodic for small values of X, while for large values of λ the response of u is as shown in Fig. 4(b). The requirement to determine valid parameters leads to the identification problem. Since one cannot always have the complete knowledge of the physical aspects of phenomena for guidance to construct a model, identification of the system and its parameters becomes an important problem for which several methods have been developed in systems theory.

B. System Identification and Parameter Estimation (Inverse Modeling)

Determination of the output signal corresponding to the input time history and system characteristics is a central problem in systems theory. In contrast to this problem, given an input and output time history, can one determine

the mathematical model that describes the system behavior? This latter problem is called an inverse problem and generally is referred to as the system identification. Identification is defined as the determination, on the basis of input and output, of a system within a specified class of systems, to which the system (phenomenon) under study is equivalent. In other words, identification is the process of constructing a mathematical model of a system from prior knowledge and observations. Equivalence is often expressed in terms of an error function E, which is a function of system output y and the model output y_m, that is,

$$\mathrm{E} = \mathrm{E}(y, y_m). \tag{13}$$

Two models m_1 and m_2 , are said to be equivalent if the value of the error function is the same for both models, that is,

$$\mathrm{E}\big(y, y_{m_1}\big) = \mathrm{E}\big(y, y_{m_2}\big). \tag{14}$$

A model can be identified with reality if the value of the error function is minimum, and in recent years numerous techniques for minimizing error functions have been developed.

Although in identification problems no prior knowledge of phenomena is assumed, in reality the modeler has at least a partial understanding of the process to be modeled, through experience and intuition. In such cases the identification problem is reduced to finding the numerical values of a number of parameters (coefficients in the differential equations representing the phenomenon) and state variables. The identification problem is thus reduced to a parameter estimation problem. Parameter estimation is defined as the experimental determination of values of parameters that govern the system behavior, assuming that the structure of the process is known.

C. Methods of System Identification

1. Frequency Response Method

Certain linear stationary processes can be identified by the so-called frequency response method. In this method, sine-wave inputs are applied to the system, and the steady-state outputs in terms of the magnitude ratio and the phase shift are measured over the entire range of frequencies of interest. Knowing the output $Y(s)$ corresponding to the input $X(s)$, the system behavior (transfer function) $G(s)$ is determined (see Fig. 5) by the operation

$$G(s) = Y(s)/X(s). \tag{15}$$

This method is applicable only to stable systems, because the output $Y(s)$ for unstable systems cannot be measured.

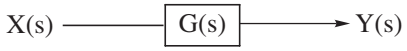


FIGURE 5 Relationship of the transfer function $G(s)$ to the input $X(s)$ and the output $Y(s)$.

2. Step-Response Method

Step input is the simplest input that can be applied to a system, like sudden closing or opening of a valve in a hydraulic network. In reality, strict step input is physically impossible, but as long as the rise time of the step input is much shorter than the period of the highest frequency in the system, the input can be considered as a step input. Step-response problems are considered as impulse-response problems, and the determination of transfer function falls into the category of time-domain problems, which can be solved by one of the gradient methods.

3. Deconvolution Method

Sometimes the input and output of a system can be related by the convolution integral

$$y(t) = \int_0^t x(\tau)w(t-\tau)\,d\tau, \tag{16}$$

where $w(\tau)$ is the impulse response of the system. The determination of the impulse response from the input and the output is called deconvolution. Once $w(\tau)$ is determined, the system transfer function can be found by using the step response method already described.

4. Cost Function Method

In identification problems, cost means penalty for not achieving correct identification. The cost function is usually expressed as a squared error (difference between the assumed value and the actual value of a parameter) function that has to be minimized for correct, or nearly correct, parameter estimation. Cost functions are mainly of two types with some variations depending on the choice of the parameter: (1) the *maximum likelihood* (ML) cost function in which the ML estimator does not assume any prior knowledge of the parameter, and (2) the *maximum a posteriori* (MAP) cost function which considers prior knowledge of the parameter obtained by maximizing the probability density of the parameter.

5. Gradient Techniques

Gradient techniques are associated with computational methods for system identification. Here, one attempts to minimize the cost function with each iteration in a direct

fashion after expanding the cost function in Taylor series about an assumed parameter to a desired order. In the computational algorithm, one picks the variation in the parameter to make the steepest descent toward the minimum cost function. The computation is repeated until there is no significant change in the parameter value from iteration to iteration. Widely applied among these methods are the first-order, second-order, and conjugate gradient methods.

6. Quasilinearization Method

As opposed to gradient technique, this is an indirect method in which a sequence of functions is iteratively computed until the functions converge to the solution. In this method, recurrence relations are obtained in the form of linear differential equations, even if the model equation has nonlinearity in its structure—hence the name quasilinearization. This method is also called the generalized Newton–Raphson method.

Consider, for example, the nonlinear differential equation

$$dx/dt = f(x), \qquad x(0) = c, \qquad t > 0 \tag{17}$$

in which the function f is continuous in x and time t and has continuous bounded second partial derivatives with respect to x for all x and t. The function can be expanded around $x^{(0)}(t)$ in Taylor series as

$$f(x) = f\left(x^{(0)}\right) + \frac{\partial f\left(x^{(0)}\right)}{\partial x}\left(x - x^{(0)}\right) + . \tag{18}$$

Neglecting the higher-order terms and combining Eqs. (17) and (18) yields the linear differential equation

$$\frac{dx}{dt} = f\left(x^{(0)}\right) + \frac{\partial f\left(x^{(0)}\right)}{\partial x}\left(x - x^{(0)}\right) \tag{19}$$

$$x(0) = c. \tag{20}$$

Proceeding likewise, one obtains

$$\frac{dx^{(n+1)}}{dt} = f\left(x^{(n)}\right) + \frac{\partial f\left(x^{(n)}\right)}{\partial x}\left(x^{(n+1)} - x^{(n)}\right) \tag{21}$$

$$x^{(n+1)}(0) = c. \tag{22}$$

The iterative process begins with an initial approximation $x^{(0)}(t)$ and proceeds until the sequence of functions converges.

7. Invariant Imbedding

The basic idea of invariant imbedding is to change a specific problem into a general problem. The solution to a specific problem is imbedded in the solution of a more general problem for which a sequential estimate of system parameters is made in the solution procedure. Invariant imbedding converts many boundary-value problems into initial-value problems, which makes both the analysis and the computational solution easier. This form of imbedding is also useful in engineering problems where sensitivity analysis has to be performed.

V. SOLUTION TECHNIQUES

For a mathematical modeler, the entire world of applied mathematics and new mathematical concepts being developed from time to time are open for use. The advent of high-speed digital computers has aided the solution techniques and opened the doors for modeling more complex systems. Some solution techniques practiced in modeling include:

1. Rigorous analysis
2. Symmetry methods
3. Obtaining *a priori* bounds on the solutions
4. Method of isoclines
5. Perturbation theory
6. Asymptotic analysis
7. Group theory
8. Inequalities
9. Integration by parts
10. Extremum principles
11. Numerical approximations (including finite difference, finite element, boundary element methods, invariant imbedding, etc.)
12. Variational principles
13. Linearization and quasi-linearization
14. Integral methods
15. Complex-variable methods
16. Graph theory
17. Mathematical programming.

This list is by no means exhaustive. For detailed analyses of these topics, the reader must refer to textbooks on applied mathematics. There are separate textbooks for several of these topics.

VI. MODEL VALIDATION

Developing a mathematical model is a long and arduous task. In mathematical modeling, the goal of the modeler is to ensure that the model replicates the phenomena being modeled to an acceptable degree. The procedures followed to test the fidelity of the model in reproducing the real

system it represents constitute model validation. On the surface, it appears that validation is something that should be done at the end of the model construction. But, in fact, validation should be carried out throughout the modeling process. A valid model can be expected by being logically consistent at each step of the modeling process, by reexamining the assumptions and constraints without sacrificing the mathematical rigor, and by turning every stone of applicable mathematical knowledge. Is it not logical that to obtain valid model behavior one must use valid assumptions and constraints? By doing so one would obtain the most accurate model, but perhaps one difficult to solve. Suppose, by a judicious choice of a solution technique from the myriad of techniques available, the modeler determines the system behavior. This step invariably (at least in physical sciences and technology) involves a numerical approximation associated with computer programming, which should be carefully carried out and correctly implemented. To generate confidence in the model, the system behavior must be compared with the real system data. It is nearly impossible, except for some simple situations, to obtain global agreement (agreement over the entire range of parameters). Then the modeler must choose the range of validity for various model parameters and establish other criteria, such as comparison of results by different solution techniques, by which to judge the model validity.

The validity of a model should not be judged by mathematical rationality alone; nor it should be judged purely by empirical validation at the cost of mathematical and scientific principles. A combination of rationality and empiricism (logic and pragmatism) should be used in the validation. If necessary, all or some of the steps in the modeling process should be repeated several times until the model is acceptable for use.

VII. CHAOS AND COMPLEXITY

In deterministic physical and mathematical systems, when the model equations are nonlinear, the evolution of the system behavior becomes irregular and unpredictable and exhibits sensitive dependence on initial conditions. Such behavior is called *chaos*. It occurs in vibrating objects, in rotating or heated fluids and in some chemical reactions. Understanding chaotic behavior involves tracking the time evolution of nonlinear dynamical system or that of natural phenomena modeled by a set of nonlinear differential equations that arise from classical equations of physics. With the exception of some first-order equations, analytical solutions of nonlinear differential equations are either difficult or impossible. The desire to overcome this difficulty has presented a significant motivation for the advancement of numerical methods and innovations of computational algorithms that lead to novel computer architectures. New discoveries in nonlinear dynamics have created new concepts and tools such as fractal dimensions and Lyapunov exponents for detecting and quantifying chaos in physical systems. There are no formal, theoretical criteria to determine under what conditions a dynamical system in general would become chaotic. Significant effort is needed to determine how and when more general and complex physical systems will become chaotic.

Complexity is one of the most perplexing problems of systems theory. Many relatively independent subsystems that are highly interconnected and interactive manifest complex systems. As a result, the collective behavior of complex systems is reflected in reproducing the functions of truly complex, self-organizing, replicating, learning and adaptive systems. Systems consisting of a large number of interacting elements lead to a perception of complex and disorderly behavior. It is generally believed that biological systems involve more interacting elements than physical systems, and therefore the latter are simple with orderly behavior. However, recent developments in irreversible thermodynamics, in the theory of dynamical systems, and in classical mechanics have narrowed the gap between the simple and complex, and between order and disorder; under certain conditions, simple systems are also known to exhibit complex behavior. Complexity in system behavior is characterized by instabilities and bifurcations. In modeling complex behavior of a system, one must first assess the nonlinear character of the underlying dynamics and identify a set of variables that control these instabilities and bifurcations.

VIII. MODELING WITH NEURAL NETWORKS

As processes increase in complexity, they become less amenable to direct mathematical modeling based on physical laws. In the later half of the 20th century, *artificial neural networks* have made inroads into several disciplines with a wide range of applications. An artificial neural network is a network of interconnected units called artificial neurons that are connected in different patterns to process information in a parallel distributed fashion much like the human brain. A significant aspect of neural networks is their ability to learn how to process information in *supervised* and *unsupervised* modes. They are used for simulation of physical systems that are modeled by massively parallel networks. In recent years, applications have been developed for modeling simple biological structures with known functions, for the modeling of higher functions of the central nervous system, for solving complicated problems in artificial intelligence and cognitive sciences,

for pattern recognition, and for solving combinatorial optimization problems. Further applications have been extended to medical diagnosis, financial services including stock price prediction, intelligent control of engineering plants, and manufacturing. Neural networks are also well adapted to ill-posed problems, those with damaged or incomplete data.

IX. MATHEMATICAL MODELING AND COMPUTERS

In all endeavors involving mathematical modeling, one cannot fail to perceive that computers act as an interface between phenomena and the various stages of mathematical modeling, including validation and use. Spectacular advances have been achieved in modeling complex and large systems using computers by removing the need for analytical solutions to differential equations representing the systems. The advances in high-speed computers and efficient numerical algorithms have generated acceptable numerical solutions to systems of the equations hitherto intractable and thus made modeling of complicated, interconnected, and interacting systems possible. As a matter of fact, complicated mathematical models expressed in terms of nonlinear partial differential equations and new applications have been the driving force behind the development of large computing machines with huge amounts of shared memory and a processing rate of up to three trillion floating-point operations per second. The need for more efficient computers to simulate complicated models has driven the computer scientists to think innovatively in the direction of new architectures and procedures. The greatest potential for achieving higher speed lies in adding parallel processors. In recent years, neurocomputers based on artificial neural networks have been built to process information efficiently and cost-effectively, but complementary to algorithmic computing. With more developments in the computer technology on the horizon, mathematical modelers can expect to achieve more success in modeling and simulation of very complicated systems, and also in revising and reworking the old models, in which drastic simplifications had to be made in the past. Despite the advances in computers and computational technology, the role of analytical methods in deciphering mathematical models should not be overlooked because they offer a valuable insight to a wide variety of problems.

X. CONCLUDING REMARKS

In this article, mathematical modeling concepts in relation to physical sciences, engineering, and technology have been surveyed. In recent years, mathematical modeling has pervaded all branches of knowledge, bringing forth greater understanding of processes under investigation. In engineering and technology it provides the analytical basis for design and control in which predictions can be confidently made without spending valuable resources of money and effort.

Successful applications of mathematical modeling techniques in engineering sciences have led the way to extend the techniques to more exotic areas of inquiry, like nanotechnology, nuclear-reactor engineering, material science, environment, weather prediction, biological processes, space sciences, cosmology, and also social sciences. Although the general philosophy of modeling in these new areas remains the same as discussed in this article, the simulation procedures and validation criteria are different and dependent on the types of models and the disciplines they belong to.

Mathematical modeling is a vast, multidisciplinary field that pleads to engage the interest and dediation of engineers, scientists and mathematicians to solve the problems facing the humankind. A significant development in the mathematical modeling activity is the availability of very-high-speed computers, which can solve a variety of complex models. In spite of all the advances in empirical knowledge, solution techniques, and computer assistance, it must be noted that human intelligence, experience, and intuition still play a significant role in mathematical modeling.

SEE ALSO THE FOLLOWING ARTICLES

CHAOS • CONTROLS, ADAPTIVE SYSTEMS • DIFFERENTIAL EQUATIONS • DISCRETE SYSTEMS MODELING • GROUP THEORY • LINEAR SYSTEMS OF EQUATIONS • NUMERICAL ANALYSIS • PERTURBATION THEORY • PROBABILITY • STOCHASTIC PROCESSES • SYSTEM THEORY

BIBLIOGRAPHY

Andrews, J. G., and McLone, R. R. (1976). "Mathematical Modelling," Butterworths, London.

Aris, R. (1995). "Mathematical Modelling Techniques," Dover, Mineola, NY.

Atherton, D., and Borne, P. (eds.) "Concise Encyclopedia of Modelling and Simulation," Elsevier, New York.

Avula, X. J. R. (ed.) (1977). "Proceedings of the First International Conference on Mathematical Modelling," 5 vols., University of Missouri-Rolla, Rolla, MO.

Avula, X. J. R., Bellman, R. E., Luke, Y., and Rigler, A. K. (eds.) (1979). "Proceedings of the Second International Conference on Mathematical Modelling," University of Missouri-Rolla, MO.

Avula, X. J. R., Kalman, R. E., Liapis, A. I., and Rodin, E. Y. (eds.) (1983). "Mathematical Modelling in Science and Technology," Proceedings of the Fourth International Conference, Zurich, Switzerland, 1983, Pergamon, New York.

Avula, X. J. R., Leitmann, G., Mote, Jr., C. D., and Rodin, E. Y. (eds.) (1987). "Mathematical Modelling in Science and Technology," Proceedings of the Fifth International Conference, University of California, Berkeley (July 1985), Mathematical Modelling, Vol. 8, Pergamon Press, New York.

Avula, X. J. R. (ed.) (1993). "Mathematical Modelling in Science and Technology," Proceedings of the Eighth International Conference, University of Maryland, College Park (April 1991). Principia Scientia, St. Louis, MO.

Avula, X. J. R., and Mote Jr., C. D. (eds.) (1994). "Mathematical Modelling in Science and Technology," Proceedings of the Ninth International Conference, University of California, Berkeley (July 1993), Principia Scientia, St. Louis, MO.

Avula, X. J. R., and Nerode, A. (1996). "Mathematical Modelling in Science and Technology," Proceedings of the Tenth International Conference, Boston, MA (July 1995). PrincipiaI Scientia, St. Louis, MO.

Avula, X. J. R., and Nerode, A. (1998). "Mathematical Modelling in Science and Technology," Proceedings of the Eleventh International Conference, Georgetown University, Washington, DC (July 1997), Principia Scientia, St. Louis, MO.

Avula, X. J. R. (2000). "Mathematical Modelling in Science and Technology," Proceedings of the Twelfth International Conference, Chicago, IL (July 1999). Principia Scientia, St. Louis, MO.

Bandemer, H. (1993). "Modelling Uncertain Data," Wiley, New York.

Bellman, R., and Roth, R. (1986). "Methods in Approximation: Techniques for Mathematical Modelling," Kluwer Academic, Norwell, MA.

Bellman, R., and Wing, G. M. (1975). "An Introduction to Invariant Imbedding," Wiley, New York.

Bellman, R., and Roth, R. (1983). "Quasilinearization and the Identification Problem," World Scientific, Singapore.

Caldwell, J., and Ram, Y. M. (1999). "Mathematical Modelling Concepts and Case Studies," Kluwer Academic, Norwell, MA.

Casti, J. L. (1989). "Alternate Realities," Wiley, New York.

Cowan, G. A., Pines, D., and Meltzer, D. (eds.) (1994). "Complexity: Metaphors, Models and Reality," Addison Wesley, Reading, MA.

Cross, M., and Moscardini, A. O. (1985). "Learning the Art of Mathematical Modelling," Wiley, New York.

Dym, C. L., and Ivey, E. S. (1980). "Principles of Mathematical Modeling," Academic, New York.

Eykhoff, P. (1974). "System Identification: Parameter and State Estimation," Wiley, London.

Fulford, G., Forrester, P., and Jones, A. (1999). "Modelling with Differential and Difference Equations," Cambridge University Press, New York.

Gibbons, M. M. (1995). "A Concrete Approach to Mathematical Modelling," Wiley, New York.

Haber, R., and Keviczky, L. (2000). "Nonlinear System Identification: Input-Output Modeling Approach," Kluwer Academic, Norwell, MA.

Jacoby, S. L. S., and Kowalik, J. S. (1980). "Mathematical Modeling with Computers," Prentice-Hall, Englewood Cliffs, NJ.

Jerome, J. W. (ed.) (1998). "Modelling and Computation for Applications in Mathematics, Science, and Engineering," Oxford University Press, New York.

Kagawa, Y. (1994). "Modelling and Simulation and Identification," Acta Press, Anaheim, CA.

King, J. R. (2000). "Emerging areas of mathematical modelling," *Phil. Trans. Roy. Soc. Lond. A* **358,** pp. 3–19.

Lasenby, J., Lasenby, A. N., and Doran, C. J. L. "A unified mathematical language for physics and engineering in the 21st Century," *Phil. Trans. Roy. Soc. Lond. A* **358,** pp. 21–39.

May, R. M. (1976). "Simple mathematical models with very complicated dynamics," *Nature* **261,** 459–467.

Meskens, N., and Roubens, M. (eds.) (1999). Kluwer Academic, Norwell, MA.

Nicholson, H. (ed.) (1980). "Modelling of Dynamical Systems," Vols. 1 and 2, Peter Peregrinus Ltd. (for The Institution of Electrical Engineers), Stevenage, U.K.

Nicolis, G., and Prigogine, I. (1989). "Exploring Complexity," W. H. Freeman & Co, New York.

Rodin, E. Y., and Avula, X. J. R. (eds.) (1989). "Mathematical Modelling in Science and Technology," Proceedings of the Sixth International Conference, St. Louis, MO (August 1987), Elsevier Science, New York.

Rodin, E. Y., and Avula, X. J. R. (eds.) (1990). "Mathematical Modelling in Science and Technology," Proceedings of the Seventh International Conference, Chicago, II (August 1989) Elsevier Science, New York.

Saaty, T. L., and Alexander, J. M. (1981). "Thinking with Models: Mathematical Models in the Physical, Biological, and Social Sciences," Pergamon, New York.

Sinha, N. K., and Kuszta, B. (1983). "Modeling and Identification of Dynamic Systems," Van Nostrand Reinhold, New York.

Stark, J. (2000). "Observing complexity, seeing simplicity," *Phil. Trans. Roy. Soc. Lond. A* **358,** 41–61.

Whorf, W. (1956). "Language and Thought," MIT Press, Cambridge, MA.

Yager, R. R., and Filev, D. P. (1994). "Essentials of Fuzzy Modeling and Control," Wiley, New York.

Zadeh, L. A. (1965). "Fuzzy sets," *Information and Control* **8,** 338–353.



Measure and Integration

G. de Barra
University of London

I. Introduction
II. Lebesgue Measure
III. The Lebesgue Integral
IV. The L^p Spaces and Inequalities for Integrals
V. Differentiation and Integration
VI. Product Spaces and Product Measures
VII. General Measures
VIII. The Radon–Nikodým Theorem and Signed Measures
IX. Extensions of the Definition of Measure
X. Extensions of Measures and Lebesgue–Stieltjes Integrals
XI. The Radon–Nikodým Property for Banach Spaces
XII. Measure and Fractals

GLOSSARY

Almost everywhere (a.e.) Except on a set of measure zero.
Complement Set of points of the space not in A, denoted $\complement A$.
Countable set Set which may be enumerated or put in one-to-one correspondence with the integers.
Empty set Set with no elements. (Notation $\varnothing$.)
Measurable function Function whose values above any given value are taken in a measurable set.
Measurable set Set whose measure is defined.
Measure Quantity defined on sets usually taking a positive value, denoted $m(A)$ for A a set of real numbers or $\mu(A)$ more generally.
Set Class of objects with property P (say). (Notation $\{x : x$ satisfies $P\}$.)
Set difference Set of points of A not in B, denoted $A \setminus B$.

THE LEBESGUE THEORY of measure and integration provides the framework for considering the limiting operations of analysis as used in pure mathematics and theoretical physics. In particular it shows when sums and integrals or limits and integrals may be interchanged. It includes the study of the theory of sets in the real line or higher dimensions with special emphasis on countable operations on sets.

I. INTRODUCTION

The theory of integration arose informally. It was formalized by Riemann but in a way which restricted its application severely. These restrictions are considered in Section III. Meanwhile in applications the subject was expanding as applied mathematicians and students of probability needed results on the limits of integrals, not provided by Riemann's theory. The Lebesgue theory supplied

the main deficiencies and the limit theorems described in Section III together with the results on differentiation of Section V rounded off the subject. The needs of the applied mathematicians were provided for also by the results on the L^p spaces of functions described in Section IV and the results using product spaces of Section VI. The Lebesgue integral turns out to be the appropriate definition to deal with orthogonal expansions and with the Fourier and Laplace transforms. The needs of the probabilists were supplied by the general theory of measure in Section VII, the decomposition theorems of Section VIII, and the results on products of measure spaces in Section VI. The differentiation of measures considered in Section VIII is of central importance in functional analysis, as is seen in Section XI.

II. LEBESGUE MEASURE

For any set A contained in the real line $\mathbb{R}$ we define the *Lebesgue outer measure* (or *outer measure*) to be the quantity $m^*(A)$ given by $m^*(A) = \inf \sum_n l(l_n)$ where we are taking the infimum or greatest lower bound over all collections $\{I_n\}$ of intervals such that $A \subseteq \cup I_n$ and where $l(I)$ denotes the length of the interval I. From this definition we have immediately that $m^*(A)$ is nonnegative; $m^*(A) = 0$ if A is a one point set or empty; $m^*(A) \leqq m^*(B)$ whenever $A \subseteq B$. It is fairly easy to show that the outer measure of any interval equals the length of the interval, so outer measure has for some sets at least the properties we would desire. Also, for any sequence of sets $\{E_i\}$ it is easy to see that

$$m^*\left(\bigcup_{i=1}^{\infty} E_i\right) \leqq \sum_{i=1}^{\infty} m^*(E_i)$$

(i.e., m^* is subadditive). We, however, cannot in general assume that equality will occur here even if the sets are pairwise disjoint. The outer measure has another desirable property: the outer measure of a set is unchanged if the set is shifted to left or right (i.e., m^* is translation invariant). It also has a regularity property: for any set A and any $\varepsilon > 0$ there is an open set O containing A and such that $m^*(O) \leqq m^*(A) + \varepsilon$.

In order to achieve the desired additivity we restrict m^* to the class $\mathcal{M}$ of Lebesgue measurable sets. We define a set E to be *Lebesgue measurable* (or just *measurable*) if for each set A

$$m^*(A) = m^*(A \cap E) + m^*(A \cap \complement E)$$

So a measurable set divides an arbitrary set in an additive way as regards outer measure. It follows from this definition that intervals are measurable. Also, all sets of zero outer measure are measurable (and so therefore are all subsets of such sets). The class of measurable sets has the important property of being a σ algebra: that is, the class contains the whole space ($\mathbb{R}$ in this case) and is closed under the formation of complements and countable unions. Only the last part here presents any difficulty. Denote m^* restricted to the σ-algebra $\mathcal{M}$ by m, called *Lebesgue measure*. Then we easily have that for any sequence $\{E_i\}$ of disjoint measurable sets

$$m\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} m(E_i)$$

(i.e., m is additive on disjoint measurable sets). For some purposes it is convenient to restrict the measure m still further to the Borel sets which may be defined as the smallest σ algebra containing the intervals. These sets are more natural in some contexts, but difficulties can arise from the fact that whereas for the class $\mathcal{M}$ every subset of a set of zero measure is again measurable, this is no longer true of the smaller class of Borel sets. So we confine our attention to the Lebesgue measurable sets.

It is easily seen that every nonempty open set has positive measure. From the countable additivity of m it is obvious that every countable set has zero measure. It is less obvious that there exist uncountable sets of zero measure. A standard example is the Cantor ternary set. It is formed by removing from the interval [0, 1] the middle third $(\frac{1}{3}, \frac{2}{3})$, then the middle thirds of the two remaining intervals, namely, $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$, then the middle thirds of the remaining intervals, and so on. The total set removed has measure adding to 1, so the residual set (the Cantor set) has measure 0. But the Cantor set contains all the numbers between 0 and 1 with expansions to base 3 consisting of 0's and 2's. So it has the same cardinality as the set of all binary expansions, and so is uncountable. It is also true that there exist sets of the real line which are not measurable. But these sets cannot be constructed and indeed can only be shown to exist using the axiom of choice or an equivalent tool of mathematical logic. The existence of these nonmeasurable sets is not crucial for the theory, but if they did not exist the theory would lose some of its content, for all sets would be measurable.

We consider now a "continuity" property of measure. Suppose $\{E_i\}$ is a sequence of measurable sets, $E_0 \subseteq E_1 \subseteq E_2 \subseteq \cdots$. Then $m(\cup_{i-1}^{\infty} E_i) = \lim m(E_i)$. Also, if $F_0 \supseteq F_1 \supseteq F_2 \supseteq \cdots$, and $m(F_0) < \infty$, then $m(\cap_{i=1}^{\infty} F_i) = \lim m(F_i)$ where again the sets $\{F_i\}$ are supposed measurable. If, as is conventional, we write here $\lim E_i = \cup_{i-1}^{\infty} E_i$ and $\lim F_i = \cap_{i=1}^{\infty} F_i$, then this result reads $m(\lim E_i) = \lim m(E_i)$ for any increasing sequence of measurable sets and also for any decreasing sequence of sets of finite measure. Lebesgue measure also has a regularity

property stronger than that of outer measure. If E is any measurable set and ε any positive number then there exists an open set $O \supseteq E$ with $m(O \setminus E) \leqq \varepsilon$ and a closed set $F \subseteq E$ with $m(E \setminus F) \leqq \varepsilon$.

Although the sets of zero measure can be uncountable, they turn out to be negligible for the purposes of integration and we say that a property holds "almost everywhere" (a.e.) if it holds except possibly on a set of zero measure. Note that sets of infinite measure frequently occur (e.g., $[1, \infty)$ is of infinite measure). So identities or inequalities involving the measures of sets must be written carefully with this in mind. In particular, indeterminate expressions of the form $\infty - \infty$ must be avoided.

The importance of Lebesgue measurable sets is that they allow us to define measurable functions. Recall that we say a function f is continuous if the set of points x with $f(x) > \alpha$ is open, for each α (i.e., $f^{-1}(\alpha, \infty)$ is open). We define f to be *Lebesgue measurable* (or just *measurable*) if $f^{-1}(\alpha, \infty)$ is a measurable set for each α. Since open sets are measurable it follows that continuous functions are measurable. It follows easily from the definition that, if f is measurable, $\{x: f(x) = \alpha\}$ is measurable for each α. Also, the constant functions are measurable (since they are continuous). Of special importance are the characteristic functions. We denote by χ_A the characteristic function (or indicator function) of the set A: $\chi_A = 1$ on A, $\chi_A = 0$ on $\mathcal{C}A$. Then from the definition of measurable functions we have that a characteristic function χ_A is measurable if, and only if, the set A is measurable. It is also easy to see that a constant multiple cf of a measurable function f is measurable. With a little more care we can see that the sum of measurable functions is measurable. So any finite linear combination of measurable functions is measurable and the product of measurable functions is measurable.

As indicated earlier it is important in applications to deal with limiting operations. So we note that the sup and inf of any finite or countable family of measurable functions is again measurable; this uses the fact that $\mathcal{M}$ is a σ algebra allowing countable set operations. It follows that for any sequence $\{f_n\}$ of measurable functions, lim sup f_n and lim inf f_n are again measurable. Here lim sup f_n is the inf$(N \geqq 1)$ of the sup$\{f_n: n \geqq N\}$, and lim inf $f_n = -\limsup(-f_n)$. If $\lim f_n$ exists it is the common value of lim sup f_n and lim inf f_n, so we have that the limit, when it exists, of a sequence of measurable functions is again measurable. The corresponding result is not true of continuous functions. As important special cases we have that $f^+ = \max(f, 0)$, and $f^- = \max(-f, 0)$ and $|f| = f^+ + f^-$ are measurable whenever f is.

For continuous functions the sup and inf are important. For a measurable function we are more interested in ess sup f and ess inf f (the essential supremum and essential infimum of f) given by ess sup $f = \inf\{\alpha: f \leqq \alpha \text{ a.e.}\}$, ess inf $f = \sup\{\alpha: f \geqq \alpha \text{ a.e.}\}$. So if we disregard sets of measure zero we replace the sup f by the possibly smaller ess sup f and the inf f by the possibly larger ess inf f. Functions for which both of the numbers are bounded are called essentially bounded. A simple property of the essential supremum to which we refer later is that for any two measurable functions

$$\text{ess sup}\,(f + g) \leqq \text{ess sup}\, f + \text{ess sup}\, g$$

and

$$\text{ess inf}\,(f + g) \geqq \text{ess inf}\, f + \text{ess inf}\, g$$

So finite linear combinations of essentially bounded functions are again essentially bounded. As an example suppose $f(x) = 1$ for x rational, $f(x) = 0$ otherwise. The f is measurable (it is the characteristic function of a countable set which of course has measure zero), and sup $f = 1$, but ess sup $f = 0$.

All the functions considered here have been real valued. For some purposes, however, it is convenient to extend the definitions to complex-valued functions. We describe the function f as measurable if Re(f) and Im(f), the real and imaginary parts of f, are measurable in the previous sense. To avoid difficulties we shall assume that any complex-valued functions considered take only finite values. It is easy to check that $|f|$ is measurable whenever f is a complex-valued measurable function.

III. THE LEBESGUE INTEGRAL

We now show how the Lebesgue integral is defined, how its value may be obtained in practice, what the principal theorems are which make its use attractive, and how these theorems may be used in examples. We revert for the moment to real-valued functions.

As with most definitions of the integral we specify the integral for a certain basic class of functions and show how the definition can be extended to a more general class. Not all functions or even all measurable functions have integrals; the restrictions arise in a natural way from the definition. So we first consider nonnegative simple functions. These are measurable functions taking only a finite number of values. Equivalently we write the real line $\mathbb{R}$ as the union of a finite number of measurable sets

$$\mathbb{R} = \bigcup_{i=1}^{n} A_i$$

and consider the function

$$\phi(x) = \sum_{i=1}^{n} a_i \chi_{A_i}$$

where the a_i are finite nonnegative numbers and χ_{A_i}, as before, stands for the characteristic function of the set A_i. Typically, one of the a_i will be zero. As special cases of simple functions we have characteristic functions of measurable sets (or of intervals, in particular). For such a characteristic function χ_A we define the integral $\int \chi_A\,dx$ to be $m(A)$. Since we want our integral to be linear we define

$$\int \phi\,dx = \sum_{i=1}^{n} a_i m(A_i)$$

where ϕ is given as above. We use the convention that in such sums $0 \cdot \infty = 0$. This definition is extended immediately to all nonnegative measurable functions f on $\mathbb{R}$ by

$$\int f\,dx = \sup \int \phi\,dx$$

the supremum being taken over all nonnegative simple functions ϕ, $\phi \leqq f$.

This would seem to give two ways of arriving at the integral of a nonnegative simple function, but they are easily shown to give the same value. Restricting to a measurable subset E of $\mathbb{R}$ we may write $\int f\chi_E\,dx$ as $\int_E f\,dx$ for any nonnegative measurable function f; in particular, if $E = [a, b]$ this is written as just $\int_a^b f\,dx$.

From the definition we have immediately some properties of the integral for nonnegative simple functions: if

$$\phi = \sum_{i=1}^{n} a_i \chi_{A_i}$$

then

$$\int_E \phi\,dx = \sum_{i=1}^{n} m(E \cap A_i)$$

for any measurable set E; $\int_{A\cup B} \phi\,dx = \int_A \phi\,dx + \int_B \phi\,dx$ for any disjoint measurable sets A and B, so the integral is additive: $\int a\phi\,dx = a\int \phi\,dx$ for any positive number a.

From these properties and from the definition we have some properties of the integral for nonnegative measurable functions: $\int f\,dx = 0$ if, and only if, $f = 0$ a.e.; if $f \leqq g$ then $\int f\,dx \leqq \int g\,dx$ (f and g nonnegative measurable functions); if $a \geqq 0$ then $\int af\,dx = a\int f\,dx$; if A and B are measurable and $A \supseteq B$ then $\int_A f\,dx \geqq \int_B f\,dx$.

The important theorems regarding the Lebesgue integral are concerned with sequences or limits, a basic one being Fatou's lemma: let $\{f_n\}$ be a sequence of nonnegative measurable functions; then

$$\int \liminf f_n\,dx \leqq \liminf \int f_n\,dx$$

So if the sequence $\{f_n\}$ converges at each point to the function f (necessarily measurable) and if for some infinite subsequence we have

$$\int f_{n_i}\,dx \leqq A$$

then

$$\int f\,dx \leqq A$$

This result has far-reaching applications and its proof is based directly on the result of the last section that $\lim m(A_i) = m(\lim A_i)$. We must allow integrals and limits of integrals to take infinite values; since all the numbers considered here are nonnegative no difficulties can arise.

As a companion result to Fatou's lemma we have the Lebesgue monotone convergence theorem. Let $\{f_n\}$ be a sequence of nonnegative measurable functions such that for each x, $\{f_n(x)\}$ is monotone increasing. So $\{f_n(x)\}$ has a limit $f(x)$ at each point (this limit may be infinite). Then the function f is a nonnegative measurable function and $\int f\,dx = \lim \int f_n\,dx$. This follows from Fatou's lemma since $\int f\,dx \leqq \liminf \int f_n\,dx \leqq \limsup \int f_n\,dx \leqq \int f\,dx$ as $f_n \leqq f$ for all n. So equality holds.

By partitioning the range of a nonnegative measurable function f into intervals and taking the simple function which, on the measurable set for which f has its values in one such interval, takes the minimum of these values we obtain a simple function approximating to f. Choosing a sequence of finer partitions gives a sequence of simple functions tending monotonically to f. Their integrals converge monotonically to that of f and provide an alternative and more manageable definition of $\int f\,dx$. Indeed this approach shows immediately that for any finite sequence $f_i, \ldots, f_n$ of nonnegative measurable functions we have

$$\int \sum_{i=1}^{n} f_i\,dx = \sum_{i=1}^{n} \int f_i\,dx$$

This extends by the Lebesgue monotone convergence theorem to infinite sums: for any sequence $\{f_n\}$ of nonnegative measurable functions,

$$\int \sum_{n=1}^{\infty} f_n\,dx = \sum_{n=1}^{\infty} \int f_n\,dx$$

This result is very useful in applications; note that it does not require the sum on the right-hand side to be finite.

We now wish to extend the definition of the integral to functions not necessarily nonnegative. If a function f is nonnegative on the set A and negative on the complementary set $B = \complement A$ we want $\int f\,dx$ to be $\int_A f\,dx - \int_B (-f)\,dx$. Equivalently, write $f^+(x) = \max(f(x), 0)$, $f^-(x) = \max(-f(x), 0)$. Then f^+ and f^- are measurable provided f is; f^+ and f^- are nonnegative; $f = f^+ - f^-$ and $|f| = f^+ + f^-$. We define, for any measurable function f, $\int f\,dx$ as $\int f^+\,dx - \int f^-\,dx$ provided at

least one of these integrals is finite; if both are finite we say f is *Lebesgue integrable*, or *integrable* in short. So the measurable function f is integrable provided the nonnegative measurable function $|f|$ has a finite integral. Then the integrals of integrable functions inherit the properties of the simpler integral, namely,

$$\int (af + bg)\,dx = a\int f\,dx + b\int g\,dx$$

for any integrable functions f and g; the integral is monotone: $f \leqq g$ implies $\int f\,dx \leqq \int g\,dx$; the integral is additive on sets: $\int_A f\,dx + \int_B f\,dx = \int_{A\cup B} f\,dx$ for disjoint measurable sets A and B. From the definition we have easily that $|\int f\,dx| \leq \int |f|\,dx$ for any integrable function f. Also a mean value theorem applies: if f is measurable and if g is integrable and α, β are real with $\alpha \leqq f \leqq \beta$ a.e. then $\int f|g|\,dx = \gamma \int |g|\,dx$ for some γ, $\alpha \leqq \gamma \leqq \beta$. In particular, if E is measurable with $m(E) < \infty$ and $\alpha \leqq f \leqq \beta$ on E then $\alpha m(E) \leqq \int_E f\,dx \leqq \beta m(E)$.

We come now to the main theorem of this section: Lebesgue's dominated convergence theorem. It states that if $\{f_n\}$ is a sequence of measurable functions such that $|f_n| \leqq g$ where g is integrable and if $\lim f_n = f$ a.e. then f is integrable and

$$\lim \int f_n\,dx = \int f\,dx$$

This follows immediately on applying Fatou's lemma to the nonnegative sequences $\{g + f_n\}$ and $\{g - f_n\}$ in turn. As a corollary we have that $\int |f_n - f|\,dx \to 0$ (i.e., f_n tends to f "in the mean"). It is quite easy to extend this result to a family of measurable functions indexed by a parameter. So if $\{f_\alpha\}$ is such a family and $f_\alpha \to f$ at each point as $\alpha \to \alpha_0$, and $|f_\alpha| \leqq g$, an integrable function, then f is integrable and $\lim_{\alpha\to\alpha 0} \int f_\alpha\,dx = \int f\,dx$. This version is useful in some applications; in particular it allows us to consider an integral $\int_{-\infty}^{\infty} f\,dx$ as limit of the integrals $\int_a^b f\,dx$ as $a \to -\infty$, $b \to \infty$, $|f|$ being the "dominating integrable function."

Another useful version of the dominated convergence theorem is as follows. Let $\{f_n\}$ be a sequence of integrable functions such that

$$\sum_{n=1}^{\infty} \int |f_n|\,dx < \infty$$

Then the series $\sum f_n(x)$ converges a.e., its sum $f(x)$ is integrable and $\int f\,dx = \sum_{n=1}^{\infty} \int f_n\,dx$. This follows on using the dominating function $\sum_{n=1}^{\infty} |f_n|$.

In order to apply these limiting theorems to specific examples we need to be able to obtain the integrals of elementary functions. It is easily seen, from the mean value theorem referred to earlier, that if f is a continuous function on a finite interval $[a, b]$ (so that f is certainly measurable) then f is integrable and the function $F(x) = \int_a^x f(t)\,dt (a < x < b)$ is differentiable with $F' = f$. So for continuous functions, integrals can be obtained using indefinite integrals, in the usual elementary manner. In particular the usual devices of integration by parts or by substitution, which follow from the corresponding rules for derivatives, can be applied in the integration of continuous functions.

All the theory of this section has dealt with real-valued functions. It may be extended to complex-valued functions just as in the last section. Define the complex-valued measurable function f to be integrable provided its real and imaginary parts Re f and Im f are integrable. The elementary theory will still be true as will the modulus inequality $|\int f\,dx| \leqq \int |f|\,dx$ though the proof now needs a little more care. Lebesgue's monotone convergence theorem and Fatou's lemma have no direct application but Lebesgue's dominated convergence theorem and its counterpart for series apply unchanged to a complex-valued sequence $\{f_n\}$. Indeed, to prove this one need consider only the sequences $\{\mathrm{Re}\, f_n\}$ and $\{\mathrm{Im}\, f_n\}$ separately.

We now have several theorems allowing us to take limits "under the integral sign" or to interchange summation and integration. In the case of nonnegative functions this interchange is always possible; in the general case a finiteness condition needs to be imposed. We now give a few examples showing how the theorems may be applied.

EXAMPLE 1. Show that

$$\int_0^1 \frac{x \log x}{1 - x}\,dx = -\sum_1^{\infty} \frac{1}{(n+1)^2}$$

The integrand on the left-hand side may be considered on (0, 1), as one point does not affect the integral, and there it equals $\sum_0^{\infty} x^{n+1} \log x$. Since $\sum_0^{\infty} x^{n+1} \log 1/x$ is a series of nonnegative functions we may integrate it term by term to get $\sum_1^{\infty} 1/(n+1)^2$ as required.

EXAMPLE 2. Show that

$$\lim \int_0^{\infty} \frac{dx}{(1 + x/n)^n x^{1/n}} = 1$$

We may suppose $x > 0$ in the integral and write $f_n(x)$ for the integrand. Then $\lim f_n(x) = e^{-x}$ and $\int_0^1 e^{-x}\,dx = 1$. So we need to interchange the integral and the limit. We may construct a dominating function $g(x)$ as follows. For $0 < x < 1$,

$$(1 + x/n)^n x^{1/n} > x^{1/2} \quad (n > 1)$$

For $1 \leqq x$,

$$(1+x/n)^n x^{1/n} \geqq (1+x/n)^n$$
$$> 1 + x + \frac{1}{2}(n(n-1))x^2$$
$$> \frac{1}{4}x^2 \quad (n > 1)$$

So let $g(x) = x^{-1/2}\,(0 < x < 1)$, $g(x) = 4/x^2$ for $1 \leqq x$. Then g is integrable and the dominated convergence theorem justifies the interchange.

EXAMPLE 3. Show that

$$\int_0^1 \sin x \log x \, dx = \sum_{n=1}^{\infty} \frac{(-1)^n}{(2n)(2n)!}$$

We have

$$\sin x \log x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} \log x = \sum_{n=0}^{\infty} f_n(x)$$

say. But

$$\int_0^1 |f_n(x)|\, dx = (-1)^{n+1} \int_0^1 f_n(x)\, dx$$
$$= \frac{1}{(2n+2)(2n+2)!}$$

Since the sum of these terms is finite the required interchange $\sum \int f_n\, dx = \int \sum f_n\, dx$ may be carried out and yields the desired result. It is often the case, as here, that the same calculation yields the finiteness condition and provides the answer.

EXAMPLE 4. Show that

$$\int_0^{\infty} \frac{\sin t}{e^t - x}\, dt = \sum_{n=1}^{\infty} \frac{x^{n-1}}{n^2+1}, \qquad -1 \leqq x \leqq 1$$

The integrand here is

$$\lim_{N\to\infty} \sum_{n=0}^{N} e^{-t} \sin t (x e^{-t})^n$$

which is in modulus $\leq 2t/(e^t - x)$ on summing this finite series. But $2t/(e^t - x)$ is integrable and nonnegative on $(0, \infty)$ so the dominated convergence theorem may be applied to the sequence of partial sums, and we get

$$\int_0^{\infty} \frac{\sin t}{e^t - x}\, dt = \sum_{n=0}^{\infty} x^n \int_0^{\infty} e^{-(n+1)t} \sin t \, dt$$
$$= \sum_{n=0}^{\infty} \frac{x^n}{1 + (n+1)^2}$$

by the usual elementary integration by parts.

Integrals are often encountered first by the student as Riemann integrals. We recall one form of the definition. Suppose f is bounded on the finite interval $[a, b]$. Write

$$S_D = \sum_{i=1}^{n} M_i(\xi_i - \xi_{i-1})$$

where $a = \xi_0 < \xi_1 < \cdots < \xi_n = b$ is a partition D of $[a, b]$ and $M_i = \sup f$ in $[\xi_{i-1}, \xi_i]$. Similarly

$$s_D = \sum_{i=1}^{n} m_i(\xi_i - \xi_{i-1})$$

where $m_i = \inf f$ in $[\xi_{i-1}, \xi_i]$. Then f is Riemann integrable on $[a, b]$ if given $\varepsilon > 0$ there exists D such that $S_D - s_D < \varepsilon$. Over all such partitions D we have $\inf S_D = \sup s_D$ with the common value as the Riemann integral of f.

Now the sums S_D, s_D may be regarded as the integrals (Lebesgue or Riemann) of step functions. When a sequence of partitions is chosen, each being a refinement of the previous one, then a monotone sequence of step functions is obtained and the limit theorems already obtained show that: if a function is bounded on a bounded interval and is Riemann integrable then it is Lebesgue integrable and the values of the integrals are the same. Of course many more functions are Lebesgue integrable than are Riemann integrable. Indeed if we examine the definition of the Riemann integral carefully we find that a function f bounded on a finite interval is Riemann integral if, and only if, there is a continuous function g which equals f a.e. This shows how restricted the class of Riemann integrable functions is and also shows that even to consider in detail the "elementary" Riemann integral we need to introduce measurable sets. The relationship is not so clear for "improper" Riemann integrals. For example, $\int_0^{\infty} x^{-1} \sin x\, dx$ exists conventionally as a Riemann integral, with a finite value, but when we try to find it as a Lebesgue integral we obtain a conditionally convergent series with $\int_0^{\infty} |x^{-1} \sin x|\, dx = \infty$, so $x^{-1} \sin x$ is not Lebesgue integrable.

Finally we note that this possibility of approximation by simpler functions which we have seen for Riemann integrable functions holds also for Lebesgue integrable functions. Indeed let f be bounded and measurable on the finite interval $[a, b]$, so f is integrable, and let $\varepsilon > 0$. Then there exists a step function h such that $\int_a^b |f - h|\, dx < \varepsilon$ and a continuous function g vanishing outside a finite interval (not necessarily identical with $[a, b]$) such that $\int_a^b |f - g|\, dx < \varepsilon$. Since, as observed after the dominated convergence theorem, $\int f\, dx$ may be regarded as the limit of integrals over finite intervals, and since in the same way the integral of an unbounded integrable function may be regarded as the limit of the integrals of bounded "truncated" functions tending to f, this result applies quite generally.

IV. THE L^p SPACES AND INEQUALITIES FOR INTEGRALS

We consider the class of real-valued measurable functions such that $|f|^p$ is integrable where p is a fixed positive number. We also introduce the convention that functions equal almost everywhere are to be identified. With this convention the class described is the space $L^p(\mathbb{R})$, or $L^p(a, b)$ if we are restricting ourselves to functions with $|f|^p$ integrable on the interval (a, b). The elements of $L^p(a, b)$ are *classes* of functions such that in each class any two functions are equal a.e. so that their integrals over all subsets of (a, b) are identical. This distinction between functions and classes of functions need not give rise to difficulty. For $p = 1$ we recover the integrable functions, with the convention stated. In addition we define $L^\infty(\mathbb{R})$ to be the essentially bounded functions of Section II with the same convention. We indicate later how these definitions extend to more general measures than Lebesgue measure.

It is easily seen that if $|f|^p$, $|g|^p$ are integrable so is $|af + bg|^p$ for a, b constant. So $L^p(\mathbb{R})$ is a vector space for $0 < p < \infty$. Also from the inequality for ess sup $|f + g|$ of Section II we have that $L^\infty(\mathbb{R})$ is a vector space. The spaces $L^p(a, b)$ are related in a simple way for a, b finite: if $0 < p < q \leqq \infty$, then $L^q(a, b) \subset L^p(a, b)$.

The spaces $L^p(\mathbb{R})$ (or $L^p(a, b)$) for $1 \leqq p \leqq \infty$ have the special property of being "normed spaces." A space X is said to be a real *normed* space if X is a real vector space and we can define a nonnegative number $\|x\|$ for each $x \in X$ in such a way that $\|x\| = 0$ if, and only if, $x = 0$ (the zero of the vector space), and $\|ax + by\| \leqq |a|\|x\| + |b|\|y\|$ for any real numbers a, b. In the case of L^p $(1 \leqq p < \infty)$ the norm of f may be defined by $\|f\|_p = (\int |f|^p\, dx)^{1/p}$; in the case of L^∞ the norm is defined by $\|f\|_\infty = \text{ess sup}|f|$. In each case this norm is unaffected if the function changes value on a set of measure zero, so it is genuinely a number associated with an element of the vector space L^p. It may be easily seen that for $1 \leqq p \leqq \infty$, $\|f\|_p \geqq 0$ and equals zero only if f is the zero of L^p (i.e., the class of functions vanishing a.e.). Also clearly $\|af\|_p = |a|\|f\|_p$. The subadditive property of the norm on L^p (i.e., $\|f + g\|_p \leqq \|f\|_p + \|g\|_p$) is Minkowski's inequality to which we refer again later). The same property for L^∞ follows from the inequality: ess sup$|f + g| \leqq$ ess sup$|f|$ + ess sup$|g|$ referred to previously. The norm on $L^\infty(a, b)$ seems different from those on $L^p(a, b)$ for $1 \leqq p \leqq \infty$ but a careful limiting argument shows that if $f \in L^\infty(a, b)$ $(a, b$ finite), so that $f \in L^p(a, b)$ for each p in $(1, \infty)$, then $\lim_{p\to\infty} \|f\|_p = \|f\|_\infty$.

We now turn to three inequalities which use the idea of convex functions. The real-valued function f is *convex* if for any two numbers x, y and any number t, $0 \leqq t \leqq 1$, we have

$$f(tx + (1 - t)y) \leqq tf(x) - (1 - t)f(y)$$

Geometrically this states that the segment jointing the points $(x, f(x))$ and $(y, f(y))$ lies above the graph of f between x and y (or more precisely, never lies before it). For example, $f(x) = e^x$ and $f(x) = x^2$ are convex functions. It is easily shown that if a function f has a second derivative f'' which is positive then f is convex. If $-f$ is convex then f is said to be *concave*. Clearly $\log x (x > 0)$ is concave.

We use the inequality

$$a^{1/p}b^{1/q} \leqq (a/p) + (b/q)$$

where $a, b, p, q, > 0$ and $1/p + 1/q = 1$. On taking logs this is equivalent to

$$1/p \log a + 1/q \log b \leqq \log (a/p + b/q)$$

which follows from the concavity of $\log x$. The inequality obtained is a generalized arithmetic-geometric mean inequality and substituting

$$a = \frac{|f|^p}{(\|f\|_p)^p}, \qquad b = \frac{|g|^q}{(\|g\|_q)^q}$$

and integrating we get Hölder's inequality. This states that if

$$1 < p < \infty, \qquad 1 < q < \infty, \qquad 1/p + 1/q = 1,$$
$$f \in L^p(\mathbb{R}) \qquad \text{and} \qquad g \in L^q(\mathbb{R})$$

then $fg \in L^1(\mathbb{R})$ and

$$\|fg\|_1 = \int |fg|\, dx$$
$$\leqq \left(\int |f|^p\, dx\right)^{1/p} \left(\int |g|^q\, dx\right)^{1/q}$$

Here p, q are described as conjugate indices; the special case $p = q = 2$ is called the Cauchy-Schwarz inequality. Of course instead of $\mathbb{R}$ we may consider integrals over any interval or any measurable set. Equality occurs in Hölder's inequality if, and only if, $af^p + bg^q = 0$ a.e. for some constants a, b not both zero.

If p, q are conjugate and $p \to 1$ then $q \to \infty$ so we may regard $1, \infty$ as conjugate. A special case of Hölder's inequality holds here: if $f \in L^1(\mathbb{R})$ and $g \in L^\infty(\mathbb{R})$ then $fg \in L^1(\mathbb{R})$ and

$$\|fg\|_1 \leqq \|f\|_1 \|g\|_\infty$$

This is quite easily proved directly.

With the help of Hölder's inequality we can easily prove Minkowski's inequality: if $p \geqq 1$ and $f, g \in L^p(\mathbb{R})$ then

$$\|f + g\|_p \leqq \|f\|_p + \|g\|_p$$

Equality holds in the case $p > 1$ if, and only if, $af = bg$ a.e. where a, b are nonnegative constants, not both zero.

The special case of Minkowski's inequality for $p = \infty$ has been referred to already.

Our third inequality is Jensen's inequality. This states that if f is a measurable function defined on the interval [0, 1] and with values in the finite range $[a, b]$ and if ϕ is a convex function on $[a, b]$ then

$$\phi\left(\int_0^1 f\,dx\right) \leq \int_0^1 (\phi \circ f)\,dx$$

The proof requires a careful examination of convex functions. In the case where ϕ is *strictly convex* (i.e., the segment referred to in the definition is strictly above the graph), we have that equality occurs in Jensen's inequality if, and only if, f is constant a.e., having the value $\int_0^1 f\,dx$. An important special case is obtained by assuming h to be a nonnegative measurable function such that $\log h$ is integrable over [0, 1] and taking f to be $\log h$ and $\phi(x)$ to be e^x: then we have

$$\exp\left(\int_0^1 \log h\,dx\right) \leqq \int_0^1 h\,dx$$

By induction proofs we can show that Hölder's and Minkowski's inequalities extend to n functions. For example, for the case $n = 3$, Hölder's inequality reads: if $1 < p < \infty$, $1 < q < \infty$, $1 < r < \infty$, $1/p + 1/q + 1/r = 1$, $f \in L^p$, $g \in L^q$, and $h \in L^r$ then $fgh \in L^1$ and $\|fgh\|_1 \leqq \|f\|_p \|g\|_q \|h\|_r$.

Hölder's inequality may be used to obtain inequalities for any integral of the product of functions. For example, to obtain an upper bound for $\int_0^{\pi/2} x^{-1/4} \cos x\,dx$ take $f(x) = x^{-1/4}$, $g(x) = \cos x$, $p = q = 2$ to get the bound $(\pi/2)^{3/4}$. As a second example,

$$\int_a^b f\,dx \leqq \left(\int |f|^p\,dx\right)^{1/p} (b-a)^{1/q}$$

for conjugate p and q on taking $g(x) \equiv 1$. This shows that $L^p(a, b) \subseteq L^1(a, b)$ for $p > 1$ and in the special case $a = 0, b = 1$ is just Jensen's inequality with $\phi(x) = x^p$.

As for any normed space the norm on L^p may be used to define a distance function with the "distance" between the functions f and g in L^p defined as $d(f, g) = \|f - g\|_p$. Then Minkowski's inequality states that the triangle inequality does indeed hold if $p \geq 1$, in which case L^p is a metric space. Convergence in terms of this metric or distance function is called convergence in the mean of order p ($p \geqq 1$). So the sequence $\{f_n\}$ converges to f in the mean of order p if each $f_n \in L^p$ and $\|f_n - f\|_p \to 0$ as $n \to \infty$. The special case of $p = 1$ was referred to in Section III, in connection with Lebesgue's dominated convergence theorem. The most important property of the L^p spaces is that of completeness: every sequence which is a Cauchy sequence, in terms of the distance function d, converges. In the usual L^p notation this may be restated as follows: if $1 \leqq p < \infty$ and $\|f_n - f_m\|_p \to 0$ as $n, m \to \infty$ then there exists a function $f \in L^p(\mathbb{R})$ such that $\|f_n - f\|_p \to 0$ as $n \to \infty$. We have in addition the existence of a subsequence $\{f_{n_i}\}$ such that pointwise convergence holds: $f_{n_i} \to f$ a.e. This result is proved by Fatou's lemma and Minkowski's inequality. The corresponding result in $L^\infty(\mathbb{R})$ is also true, with a more direct proof: if each $f_n \in L^\infty(\mathbb{R})$ and $\|f_n - f_m\|_\infty \to 0$ as $n, m \to \infty$ then there exists a function $f \in L^\infty$ such that $\lim f_n = f$ a.e. and $\|f - f_n\|_\infty \to 0$ as $n \to \infty$. These results need to be used with care: even if the functions f_n are specified functions, the limit f in L^p is defined only as an element of L^p and its value at any specific point cannot be obtained. The completeness of the space L^2 is of special importance in physical applications.

V. DIFFERENTIATION AND INTEGRATION

Differentiation and integration are closely connected; since we have extended the elementary notion of integral we must deal carefully with differentiation so that this relation continues to hold. The first point to note is that continuous functions are not very relevant here; indeed continuous functions which are nowhere differentiable are easily constructed. For example, on the interval (0, 1) let $f_n(x)$ denote the distance from x to the nearest number of the form $m/10^n$ where m and n are nonnegative integers. Then f_n has a "sawtooth" graph with 10^n "teeth" and $\max f_n = \frac{1}{2} \cdot 10^{-n}$. So f_n certainly continuous, and $\sum f_n(x)$ is uniformly convergent with sum $f(x)$, say. Then f is continuous but for each $x \in (0, 1)$ by considering its decimal expansion we can show that the graph of $y = f(x)$ has no tangent at this point.

We now consider an important class of functions which although not necessarily continuous are well behaved as regards differentiability, namely, the functions of bounded variation. We suppose f is defined and finite-valued on the finite interval $[a, b]$ and take a partition $a = x_0 < x_1 < \cdots < x_k = b.$ of $[a, b]$. Then we form the sums

$$p = \sum_{i=1}^{k} (f(x_i) - f(x_{i-1}))^+$$

$$n = \sum_{i=1}^{k} (f(x_i) - f(x_{i-1}))^-$$

$$t = p + n = \sum_{i=1}^{k} |f(x_i) - f(x_{i-1})|$$

where we are using the notation $a^+ = \max(a, 0)$ and $a^- = \max(-a, 0)$. So $t, p, n, \geqq 0$ and $f(b) - f(a) = p - n$.

Taking upper bounds over all partitions of $[a, b]$, and keeping to the same function f we let $P = \sup p$, $N = \sup n$, $T = \sup t$ and call these quantities the *positive, negative*, and *total* variations of f over $[a, b]$. If T is finite, f is said to be of bounded variation over $[a, b]$ or to belong to the class $\mathrm{BV}[a, b]$. Taking suprema over partitions the relations for t, p, n give $T = p + N$ and $f(b) - f(a) = P - N$. Also, each of these variations is additive over intervals, so if $a < c < b$ then $T[a, b] = T[a, c] + T[c, b]$ and similarly for P, N. This follows immediately from the corresponding identities for t, p, n, and leads to the important result that a function f is of bounded variation over $[a, b]$ if, and only if, it may be written as the difference of two finite-valued monotone increasing functions g and h, say. These are obtained by defining $g(x) = P[a, x] + f(a)$ and $h(x) = N[a, x]$. The converse follows from the fact that any finite-valued monotone function is of bounded variation, and so therefore is the difference of two such functions. Indeed, the class of functions $\mathrm{BV}[a, b]$ forms a vector space; linear combinations of functions of bounded variations are again functions of bounded variation.

Functions of bounded variation share the "good" properties of monotone functions. Since any finite-valued monotone increasing function is continuous except possibly on a countable set the same is true for functions $f \in \mathrm{BV}[a, b]$, so in particular these functions are measurable. An example of a monotone-increasing function with a countable number of discontinuities is provided by letting $\{r_i\}$ be an enumeration of the rational numbers in $[0, 1]$ and defining $f(x) = \sum_{r_n \leqq x} 2^{-n}$. Then f is discontinuous at each rational in the interval. For an example of a function f not of bounded variation on $[0, 1]$ define f arbitrarily at $x = 0$ and let $f(x) = \sin(1/x)$ otherwise; another example is given by $f(x) = x \sin(1/x)$, $x \neq 0$, $f(0) = 0$. This example shows directly that f may be continuous but not of bounded variation.

Lebesgue proved that if $f \in \mathrm{BV}[a, b]$, then f is differentiable a.e. and its derivative is finite a.e. This is a significantly more difficult result to prove than the earlier results of this section. The two examples of the previous paragraph show that the converse of this theorem is not true.

We consider now indefinite integrals and write for any integrable function f, $F(x) = \int_a^x f\,dt$, so that F is the indefinite integral of f over the interval $[a, b]$, say, on which f is integrable. Then Lebesgue's dominated convergence theorem, applied to the family of functions $\chi_{[a,x]} f$ where $x \to x_0$ shows that F is continuous. It also follows easily from the definitions that $F \in \mathrm{BV}[a, b]$, its total variation being bounded by $\int_a^b |f|\,dt$. So F' exists a.e. and it can easily be shown that $F' = f$ a.e. in $[a, b]$. We cannot expect F to be everywhere differentiable (e.g., let f be a step function, then F' does not exist at the points of discontinuity). Nor can we assume that whenever F' exists it equals f, for we may take a continuous function as f (so that F' exists at all points) and then change f at a single point x_0. Then F is unchanged but $F'(x_0)$ cannot equal $f(x_0)$.

Indefinite integrals, however, have the important property of being absolutely continuous. A function f is said to be *absolutely continuous* on $[a, b]$ if given $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\sum_{i=1}^{n} |f(x_i) - f(y_i)| < \varepsilon$$

whenever $\sum_{i=1}^{n} |x_i - y_i| < \delta$ for any finite set of disjoint intervals (x_i, y_i) in $[a, b]$. Taking the special case $n = 1$ we see that absolutely continuous functions are continuous. Considering any partition of $[a, b]$ and introducing new partition points at a distance at most δ apart we can show that every absolutely continuous function is of bounded variation.

Now for any integrable function f we have that $\int_E |f|\,dt$ tends to zero as $m(E) \to 0$. This is obvious for a bounded function f, and follows for an unbounded function since the integral of f over any set is the limit of integrals of the functions f_n where $f_n = f$ provided $|f| \leqq n$, $f_n = \pm n$ otherwise. From this it follows (with $E = \cup_{i=1}^{n} (x_i, y_i)$) that if f is integrable over $[a, b]$ its indefinite integral is absolutely continuous there. It is a little more difficult to prove the converse: every absolutely continuous function is an indefinite integral, indeed it is the indefinite integral of its derivative, a function which we know to exist a.e. That the derivative of any function of bounded variation f is measurable can be seen from the fact that it may be obtained as the limit of a sequence of ratios $g_n(x) = n(f(x + 1/n) - f(x))$ which are themselves measurable. So, on finite intervals, a function is an indefinite integral if, and only if, it is absolutely continuous.

VI. PRODUCT SPACES AND PRODUCT MEASURES

For any two spaces X and Y the *Cartesian product* $X \times Y$ is the set of ordered pairs $\{(x, y): x \in X, y \in Y\}$. We will concern ourselves with the case $X = Y = \mathbb{R}$ so that $X \times Y$ is just the plane $\mathbb{R}^2$. However, the product notation is useful for subsets of $\mathbb{R}^2$. We call a set E in $\mathbb{R}^2$ a *rectangle* if $E = A \times B$, $A \subseteq \mathbb{R}$, $B \subseteq \mathbb{R}$. For the purpose of measure and integration the important sets are the *measurable rectangles*: sets of the form $A \times B$ where A and B are measurable sets. These include as special cases the "genuine" rectangles where A and B are intervals.

From the basic measurable rectangles we can form the σ-algebra $\mathcal{M} \times \mathcal{M}$ which is the least σ algebra containing the measurable rectangles. So within $\mathcal{M} \times \mathcal{M}$ we may

form complements and countable unions. It can be shown that if we take all finite unions of measurable rectangles and then take the smallest class of sets which contains these sets and is closed under the formation of countable increasing unions and countable decreasing intersections then we get just the σ-algebra $\mathcal{M} \times \mathcal{M}$. As we shall see this is crucial for the theory. We shall call the sets of $\mathcal{M} \times \mathcal{M}$ the measurable sets of the plane. An alternative definition extends $\mathcal{M} \times \mathcal{M}$ by also including all subsets of sets of measure zero, so as to get a complete measure space (see Section VII).

For any set E in $\mathbb{R}^2$ define the *x section* of E to be the set $E_x = \{y: (x, y) \in E\}$ and the *y section* of E to be the set $E^y = \{x: (x, y) \in E\}$. These are sets in $\mathbb{R}$, not $\mathbb{R}^2$. It can be shown quite easily that measurable sets in the plane have measurable sections. We wish now to define the product measure on the sets of $\mathcal{M} \times \mathcal{M}$. This definition should provide, for the measurable rectangle $A \times B$, the measure $m(A)m(B)$. We use the result that if E is a measurable set in the plane then $\phi(x) = m(E_x)$ and $\psi(y) = m(E^y)$ are measurable functions of x and y, respectively, and

$$\int \phi \, dx = \int \psi \, dy \tag{1}$$

The common value of these integrals is then taken as the measure of E. It is clear that for the measurable rectangle $E = A \times B$, $\phi(x) = \chi_A(x)m(B)$ so $\int \phi \, dx = m(A)m(B)$. Also, $\psi(y) = \chi_B(y)m(A)$ so $\int \psi \, dy = m(A)m(B)$ also. So identity (1) holds for measurable rectangles and the measure obtained is the desired one. It holds similarly for finite unions of measurable rectangles. Using the Lebesgue monotone convergence theorem we obtain (1) for the union of monotone increasing sequences of sets $\{E_n\}$, as the sequences $\{\phi_n\}$ and $\{\psi_n\}$ are similarly monotone. Similarly for a decreasing sequence of sets $\{F_n\}$, contained in a bounded rectangle, the result for the intersection follows from the Lebesgue dominated convergence theorem. Since the plane may be written as the union of a sequence of bounded rectangles

$$\mathbb{R}^2 = \bigcup_{n,m=-\infty}^{\infty} \{[n, n+1) \times [m, m+1)\}$$

the result follows for all the sets of $\mathcal{M} \times \mathcal{M}$ on adding together the results for the component pieces.

These definitions and the sketch proof just given work equally well with general measures μ, ν say (see Section VII), with a measurable rectangle $A \times B$ having the measure $\mu(A)\nu(B)$. The only condition on the measures is that each coordinate space can be written as a countable union of sets of finite measure for μ or ν, as the case may be, so that the product space may be decomposed into a sequence of sets of finite measure as for $\mathbb{R}^2$ in the preceding paragraph.

If f is a function of x and y we may similarly define the x section of f as the function $f_x(y) = f(x, y)$ for each fixed x, and the y section of f as $f^y(x) = f(x, y)$ for each fixed y. Since measurable sets have measurable sections it follows easily that if f is measurable with respect to $\mathcal{M} \times \mathcal{M}$ then f_x and f^y are measurable with respect to m.

For any nonnegative measurable function f we have the result that the integral of f may be expressed in terms of repeated integrals in either order and both of these integrals are defined. More precisely, let f be nonnegative and measurable with respect to $\mathcal{M} \times \mathcal{M}$; write $\phi(x) = \int f_x \, dy$, $\psi(y) = \int f^y \, dx$. Then ϕ and ψ are measurable and

$$\int \phi \, dx = \int f \, dx \, dy = \int \psi \, dy \tag{2}$$

where the middle integral is the Lebesgue integral of f with respect to the product measure defined earlier in this section. Identity (2) has already been proved for the case of a function of the form χ_E by identity (1), and so for any nonnegative measurable simple function. Taking a sequence of such functions f_n increasing monotonically to f, the sections $(f_n)_x \uparrow f_x$ and $(f_n)^y \uparrow f^y$ and the Lebesgue monotone convergence theorem gives identity (2).

We need now to extend (2) to functions not necessarily nonnegative and expect now to find some finiteness condition coming in. Now, identity (2) applied to $|f|$ states that $|f|$ is integrable if, and only if, each of the iterated integrals of $|f|$ (or, more precisely, of the sections of $|f|$) is finite, and then all three integrals are equal. In this case we write f as the difference $f = f^+ - f^-$ of nonnegative measurable functions and we apply identity (2) to f^+, f^-, and their sections. On subtracting the results we get

$$\int dx \int f_x \, dy = \int f \, dx \, dy = \int dy \int f^y \, dx \tag{3}$$

So from the remarks concerning $|f|$ we deduce Fubini's theorem which states that if f is a measurable function of x and y and *either* of the iterated integrals of $|f|$ is finite then so is the other, $|f|$ is integrable and identity (3) holds for f.

An important application of this theory is in connection with the Laplace and Fourier transforms of an integrable function. We will describe the application in the Fourier transform case; the other is similar. The following result allows us to define the *convolution* of two functions. Let f and g be integrable functions; then $f(y-x)g(x)$ is an integrable function of x for almost all y and if $h(y)$ is defined for these y by $h(y) = \int f(y-x)g(x) \, dx$ then h is integrable and $\|h\|_1 \leqq \|f\|_1 \|g\|_1$. To prove this we need to show that $f(y-x)g(x)$ is measurable with respect to $\mathcal{M} \times \mathcal{M}$. This is not very difficult and assuming

it we may write $H(y)=\int |f(y-x)g(x)|\,dx$. Then Fubini's theorem gives

$$\int H(y)\,dy = \int dy \int |f(y-x)g(x)|\,dx$$
$$= \int |g(x)|\,dx \int |f(y-x)|\,dy$$

Now, as noted in Section II, Lebesgue measure is invariant under translation, so we may make a translation of variables without changing the integral. Thus

$$\int |f(y-x)|\,dy = \int |f(y)|\,dy$$

and

$$\|H\|_1 \leqq \|g\|_1 \|f\|_1$$

So H and hence h is finite-valued a.e., h is integrable, and $\|h\|_1 \leqq \|f\|_1\|g\|_1$. Extending $(2\pi)^{-1/2}h$ by giving it any fixed value on the exceptional set we obtain the convolution

$$(f*g)(y) = (2\pi)^{-1/2}\int f(y-x)g(x)\,dx$$

of the integrable functions f and g. We can easily show that $f*g=g*f$ a.e. and $(f*g)*h=f*(g*h)$ a.e. for any integrable functions f, g, h. For example, to prove the first identity, let y be such that $g(y-x)f(x)$ is integrable with respect to x, so y does not lie in the exceptional set of measure zero. For such y let $t=y-x$ so that $\int g(y-x)f(x)\,dx$ becomes

$$\int g(t)f(y-t)\,dt = (2\pi)^{1/2}(f*g)(y)$$

as we may translate the variables.

For any integrable function f we may define the *Fourier transform* as the function $\hat{f}$ given by

$$\hat{f}(s) = (2\pi)^{-1/2}\int e^{-ist}f(t)\,dt$$

Then $\hat{f}$ is a continuous function, by the Lebesgue dominated convergence theorem with $|f|$ as dominating function. Also, $|\hat{f}| \leqq (2\pi)^{-1/2}\int |f|\,dt$. The convolution and the Fourier transform are related by the identity

$$\widehat{(f*g)} = \hat{f}\hat{g}$$

true for integrable functions. To see this note that

$$\widehat{(f*g)}(s) = (2\pi)^{-1}\int dt \int e^{-ist}f(t-x)g(x)\,dx$$

Then the modulus of the integrand here, $|f(t-x)g(x)|$, is integrable by the argument given for the convolution. So by Fubini's theorem we may interchange the order of integration to get

$$\widehat{(f*g)}(s) = (2\pi)^{-1}\int g(x)\left(\int e^{-ist}f(t-x)\,dt\right)dx$$
$$= (2\pi)^{-1}\int g(x)\left(\int e^{-is(t+x)}f(t)\,dt\right)dx$$
$$= (2\pi)^{-1}\int e^{-isx}g(x)\,dx \int e^{-ist}f(t)\,dt$$
$$= \hat{f}(s)\hat{g}(s)$$

as required.

The Fourier transform has been defined here for functions of $L^1(\mathbb{R})$. We may extend the definition by an approximation argument to the functions of $L^2(\mathbb{R})$ and show that $\|\hat{f}\|_2 = \|f\|_2$ (Parseval's theorem). Note that the Fourier transform though continuous and bounded need not be integrable. But if it is (i.e., when both f and $\hat{f}$ are integrable functions), we have the Fourier inversion theorem:

$$f(x) = (2\pi)^{-1/2}\int \hat{f}(t)e^{ixt}\,dt \text{ a.e.}$$

The proof is not very difficult and depends on Fubini's theorem.

VII. GENERAL MEASURES

Measures arise in various ways and in various spaces and the theory outlined above can be applied in the main provided we have an appropriate class of sets defined to be measurable. So we shall suppose we have a set or space X and on it a σ-algebra $\mathcal{S}$ of sets. On the sets of $\mathcal{S}$ our measure μ is defined and so μ is presumed to be a nonnegative set function which takes the value 0 on the empty set and which is countably additive (i.e., if $\{A_n\}$ is a sequence of disjoint sets of $\mathcal{S}$ we have $\mu(\cup_{n=1}^{\infty}A_n)=\sum_{n=1}^{\infty}\mu(A_n)$). Then the triple $\{X, \mathcal{S}, \mu\}$ is called a *measure space*. This measure space is said to be σ *finite* if we may write X as $X=\cup_{n=1}^{\infty}X_n$, $X_n\in\mathcal{S}$, and $\mu(X_n)<\infty$. It is said to be a *complete* measure space if for any set E of $\mathcal{S}$ with $\mu(E)=0$ every subset of E also belongs to $\mathcal{S}$ (and then of course has measure zero also).

EXAMPLE 1. $X=\mathbb{R}$, $\mathcal{S}=\mathcal{M}$, $\mu=m$, gives Lebesgue measure on the real line. This is σ finite and complete.

EXAMPLE 2. $X=\mathbb{R}$, $\mathcal{S}=$ Borel sets, $\mu=m$, gives Borel measure. This is also σ finite but is not complete.

EXAMPLE 3. $X=\mathbb{R}^2$, $\mathcal{S}=\mathcal{M}\times\mathcal{M}$, $\mu=m\times m$, gives planar measure. This is σ finite but not complete.

EXAMPLE 4. $X=\mathbb{N}$ the set of natural numbers; $\mathcal{S}$ is the class of all subsets of $\mathbb{N}$. Let $\sum a_n<\infty$ with $a_n\geqq 0$ and define $\mu(E)=\sum a_k$ where the summation is over all integers k in E. This measure is obviously complete and is finite for the whole space (and so trivially σ finite).

Results such as $m(\lim E_i) = \lim m(E_i)$ hold as before for monotone sequences of sets. A real function f is measurable if $f^{-1}(\alpha, \infty)$ belongs to $\mathcal{S}$, the family of measurable sets. Integration theory proceeds as before, the nonnegative simple function $\phi = \sum_{i=1}^{n} a_i \chi_{A_i}$ having the integral $\int \phi \, d\mu = \sum_{i=1}^{n} a_i \mu(A_i)$ as in Section III. Then the integral $\int f \, d\mu$ of a nonnegative measurable function is the least upper bound of such integrals for all such functions $\phi \leqq f$. The integration of functions not necessarily nonnegative and of complex-valued functions is defined as before and the same theorems hold: Fatou's lemma, the Lebesgue monotone convergence theorem, the Lebesgue dominated convergence theorem, and their variants for series. Of course the Riemann integral will not exist in general so comparisons cannot be made. Also the result noted in Section V holds in general: $\int_E |f| \, d\mu$ tends to zero as $\mu(E) \to 0$.

The constructions and inequalities obtained in Section IV will hold good for the general L^p spaces $L^p(X, \mathcal{S}, \mu)$. So will the results of Section VI, as already noted, for a pair of measures μ and ν provided we assume these to be σ finite.

For some purposes it is convenient to use a complete measure and it is an important fact that a measure may be extended in a unique way so as to be complete. We replace the σ-algebra $\mathcal{S}$ by the larger class $\mathcal{S}'$ of sets of the form $E \cup N$ where $E \in \mathcal{S}$, $E \cap N = \varnothing$, and $N \subseteq M$ where $M \in \mathcal{S}$ with $\mu(M) = 0$. Then we check that $\mathcal{S}'$ so defined is a σ algebra and define μ on $\mathcal{S}'$ by $\mu(E \cup N) = \mu(E)$. This clearly extends μ and is easily seen to give an unambiguous definition. With this construction $\{X, \mathcal{S}', \mu\}$ is a complete measure space.

If $\mu(X) = 1$ the triple $\{X, \mathcal{S}, \mu\}$ is called a probability measure space. Then a particular set of terms is used, for historical reasons, and the emphasis is on a particular type of result. In particular for "almost everywhere" one writes "almost surely," abbreviated a.s.; measurable functions are called random variables; their Fourier transforms $\phi yx = \int e^{ixt} f(t) \, d\mu(t)$ are called characteristic functions. Note that Jensen's inequality will apply directly with [0, 1] replaced by X; also that the product of probability spaces is again a probability space.

We shall consider now some further types of convergence which may be applied to sequences of functions. Let $\{f_n\}$ be a sequence of measurable functions and f a measurable function (all on the measure space $\{X, \mathcal{S}, \mu\}$). Then f_n tends to f *in measure* if for every positive ε,

$$\lim \mu\{x: |f_n(x) - f(x)| > \varepsilon\} = 0$$

If μ is a probability measure this is termed convergence *in probability*. It is easily seen that a sequence $\{f_n\}$ can tend at most to one limit function f (up to sets of measure zero). Indeed the sequence $\{f_n\}$ defines the function f a.e., in the sense that if $\{f_n\}$ is a Cauchy sequence with respect to convergence in measure so that for any $\varepsilon > 0$,

$$\lim_{n,m \to \infty} \mu\{x: |f_n(x) - f_m(x)| > \varepsilon\} = 0$$

then there exists a measurable function f so that $f_n \to f$ in measure and also for some subsequence $\{n_i\}$, $f_{n_i} \to f$ a.e. This is proved by a careful choice of ε's as elements of a convergent series.

An analog of Fatou's lemma holds for convergence in measure: Let $\{f_n\}$ be a sequence of nonnegative measurable functions and let f be a measurable function such that $f_n \to f$ in measure; then

$$\int f \, d\mu \leqq \liminf \int f_n \, d\mu$$

The proof depends on applying the original Fatou's lemma of Section III to subsequences $\{f_{n_i}\}$ tending to f a.e. There is a corresponding analog of the Lebesgue dominated convergence theorem. Let $\{f_n\}$ be a sequence of measurable functions such that $|f_n| \leqq g$, an integrable function, and let $f_n \to f$ in measure, where f is measurable. Then f is integrable, $\lim \int f_n \, d\mu = \int f \, d\mu$, and $\lim \int |f_n - f| \, d\mu = 0$ (i.e., $f_n \to f$ in the mean). This result is easily proved. Since there exists a subsequence $\{f_{n_i}\}$ with limit f a.e., we have $|f| \leqq g$ so $f \in L^1(\mu)$. Also, for each n, $g + f_n \geqq 0$ and $g + f_n \to g + f$ in measure, so by the version of Fatou's lemma just given

$$\int g \, d\mu + \int f \, d\mu \leqq \liminf \int (g + f_n) \, d\mu$$

and

$$\int f \, d\mu \leqq \liminf \int f_n \, d\mu$$

Similarly, $g - f_n \geqq 0$, whence

$$\int g \, d\mu - \int f \, d\mu \leqq \liminf \int (g - f_n) \, d\mu$$

and so

$$\int f \, d\mu \geqq \limsup \int f_n \, d\mu$$

$$\geqq \liminf \int f_n \, d\mu \geqq \int f \, d\mu$$

Therefore equality holds and the first result follows. Also it is easily seen that $|f_n - f| \to 0$ in measure. But $|f_n - f| \leqq 2g$ and so the second result follows from the first.

In $L^p(\mu)$, convergence in the sense of the norm is termed convergence in the mean of order p (i.e., $f_n \to f$ in the mean of order $p (p \geqq 1)$ if $\lim \|f_n - f\|_p = 0$). (It may be defined for any $p > 0$, but for $0 < p < 1$ the norm

notation is not appropriate.) This convergence and convergence in measure are easily shown to be related: if $f_n - f$ in the mean of order p then $f_n \to f$ in measure. For suppose not. Then there exist $\varepsilon > 0$, $\delta > 0$ such that $\mu\{x: |f_n(x) - f(x)| > \varepsilon\} > \delta$ for infinitely many n. But then $\int |f_n - f|^p \, d\mu \geqq \varepsilon^p \delta$ for infinitely many n, contradicting convergence in the mean of order p.

Another important kind of convergence is almost uniform convergence. Let $\{f_n\}$ be a sequence of measurable functions and let f be a measurable function. Then we say that $f_n \to f$ almost uniformly (abbreviated a.u.) if for any $\varepsilon > 0$ there exists a set E with $\mu(E) < \varepsilon$ and such that on the complement of E, $f_n \to f$ uniformly (i.e., given $\varepsilon' > 0$ there exists N such that for $n > N$, $|f_n(x) - f(x)| < \varepsilon'$ for all $x \in \complement E$ (N depending on ε' but not on x)). Obviously uniform convergence (on the whole space) implies almost uniform convergence (just take $E = \varnothing$ above). In the opposite direction, the sequence $\{x^n\}$ converges to zero almost uniformly on [0, 1] but not uniformly.

If $f_n \to f$ a.u., however, then $f_n \to f$ in measure. For if not then there exist positive ε and δ such that

$$\mu\{x: |f_n(x) - f(x)| > \varepsilon\} > \delta$$

for infinitely many n. But since there exists E with $\mu(E) < \delta$ and with $f_n \to f$ uniformly on $\complement E$ we get a contradiction. We can also show easily that if $f_n \to f$ a.u. then $f_n \to f$ a.e. For if m is any positive integer we can find a set E_m with $\mu(E_m) < 1/m$ and on $\complement E_m$, $f_n \to f$ uniformly. But if $x \in \cup_{m=1}^{\infty} \complement E_m$ we have $x \in \complement E_N$, say, so $\lim f_n(x) = f(x)$. Since $\complement \cup_{m=1}^{\infty} \complement E_m = \cap_{m=1}^{\infty} E_m$, a set of measure zero, the result follows.

An important result in the opposite direction is given by Egorov's theorem: on a space of finite measure convergence almost everywhere implies almost uniform convergence. We can prove this by writing for each pair of positive integers k, n,

$$E_{k,n} = \{x: |f_m(x) - f(x)| < 1/k, m \geqq n\}$$

Then as $f_m \to f$ a.e., we have

$$\mu\left(\bigcap_{n=1}^{\infty} \complement E_{k,n}\right) = 0$$

and since all measures are finite,

$$\mu(\complement E_{k,n}) \to 0 \qquad \text{as} \quad n \to \infty$$

for each fixed k. So if $\varepsilon > 0$, for an appropriate n_k we have $\mu(\complement E_{k,n_k}) < \varepsilon/2^k$. Then if $E = \cap_{k=1}^{\infty} E_{k,n_k}$ we have $\mu(\complement E) < \varepsilon$, and on E, for each k, $|f_m - f| < 1/k$ for $m \geqq n_k$. So $f_n \to f$ a.u.

With a slight variation of this proof one can show that, for dominated sequences, convergence almost everywhere implies almost uniform convergence (i.e., if $f_n \to f$ a.e. where $|f_n| \leqq g$, an integrable function, then $f_n \to f$ a.u.).

Examples can be constructed showing that certain types of convergence do not in general imply certain other ones. For instance, let χ_n^i be the characteristic function of the interval $[(i-1)/n, i/n]$, $i = 1, \ldots, n$, with [0, 1] as the whole space and using Lebesgue measure and define $\{f_n\}$ to be the sequence $\chi_1^1, \chi_2^1, \chi_2^2, \chi_3^1, \chi_3^2, \chi_3^3, \ldots$. Then $f_n = \chi_k^i$ where $\frac{1}{2}k(k-1) < n \leq \frac{1}{2}k(k+1)$, whence $\int f_n \, dx = 1/k \to 0$ as n (and thus k) $\to \infty$, so that $f_n \to 0$ in the mean. But $f_n \to 0$ a.e.; indeed for no x in [0, 1] does $f_n(x) \to 0$. Nor does $f_n \to 0$ a.u. The space here, [0, 1], has nevertheless finite measure and the sequence is dominated by the integrable function $\chi_{[0,1]}$.

VIII. THE RADON–NIKODÝM THEOREM AND SIGNED MEASURES

New measures arise from given measures in natural ways. For example, let f be a nonnegative function measurable with respect to the measure μ on the σ-algebra $\mathcal{S}$, and define ν on $\mathcal{S}$ by

$$\nu(E) = \int_E f \, d\mu$$

Then ν is a nonnegative set function, vanishing on the empty set. That ν is countably additive follows from the corollary to the Lebesgue monotone convergence theorem. For if $\{E_n\}$ is a sequence of disjoint sets in $\mathcal{S}$ then

$$\begin{aligned} \nu\left(\bigcup_{n=1}^{\infty} E_n\right) &= \int \sum_{n=1}^{\infty} \chi_{E_n} f \, d\mu \\ &= \sum_{n=1}^{\infty} \int \chi_{E_n} f \, d\mu \\ &= \sum_{n=1}^{\infty} \nu(E_n) \end{aligned}$$

So ν is a measure. The two measures μ, ν are related by the following continuity condition. If μ_1, μ_2 are any two measures on the σ-algebra $\mathcal{S}$ and $\mu_2(E) = 0$ whenever $\mu_1(E) = 0$ we say that μ_2 is *absolutely continuous* with respect to μ_1, and write $\mu_2 \ll \mu_1$. Obviously for the measure ν constructed above, $\nu \ll \mu$.

If we remove the restriction that f be nonnegative we obtain measures with negative sign. So we have the following definition: a set function ν on the σ-algebra $\mathcal{S}$ is a *signed measure* if its values are real of infinite, ν takes at most one of the values ∞ and $-\infty$,

$$\nu(\varnothing) = 0 \qquad \text{and} \qquad \nu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \nu(E_i)$$

whenever $\{E_i\}$ is a disjoint sequence of sets of $\mathcal{S}$. Clearly every measure is a signed measure.

EXAMPLE. Let f be an integrable function on $\{X, \mathcal{S}, \mu\}$. Then $\nu(E) = \int_E f\, d\mu$ defines a signed measure. Clearly only the countable additivity needs checking, so for $E \in \mathcal{S}$ write $\nu^+(E) = \int_E f^+ d\mu$, $\nu^-(E) = \int_E f^- d\mu$. Then ν^+, ν^- are (finite) measure and

$$\begin{aligned}\nu\left(\bigcup_{i=1}^{\infty} E_i\right) &= \nu^+\left(\bigcup_{i=1}^{\infty} E_i\right) - \nu^-\left(\bigcup_{i=1}^{\infty} E_i\right) \\ &= \sum_{i=1}^{\infty} \int_{E_i} f^+\, d\mu - \sum_{i=1}^{\infty} \int_{E_i} f^-\, d\mu \\ &= \sum_{i=1}^{\infty} \int_{E_i} f\, d\mu = \sum_{i=1}^{\infty} \nu(E_i)\end{aligned}$$

In this example the signed measure ν has a decomposition as the difference of measures and there is a corresponding decompositions of the space into the sets $\{x\colon f(x) \geqq 0\}$ and $\{x\colon f(x) < 0\}$ on one of which ν acts like a measure and on the other $-\nu$ acts like a measure. More generally for any signed measure ν there is a decomposition $\nu = \nu_1 - \nu_2$ as the difference of measures for which ν_1 and ν_2 are *mutually singular* (written $\nu_1 \perp \nu_2$) (i.e., for some set $A \in \mathcal{S}$ we have $\nu_2(A) = \nu_1(\complement A) = 0$). Then ν_1 and ν_2 are said to be the *Jordan decomposition* of ν and are uniquely defined by ν. There is a corresponding decomposition of the space X as the union of disjoint sets A, B of $\mathcal{S}$ such that ν is nonnegative on the measurable subsets of A, $-\nu$ is nonnegative on the measurable subsets of B. This, the *Hahn decomposition*, is unique up to sets for which all subsets in $\mathcal{S}$ have zero ν measure. The example given above where ν is defined using an integrable function displays both the Jordan and Hahn decompositions. In the general case the Hahn decomposition is established first using an exhaustion argument to find the "largest" set on which ν acts like a nonnegative measure and the Jordan decomposition follows.

By analogy with the functions of bounded variation discussed in Section V the measure ν^+ is called the positive variation of ν, ν^- the negative variation, and $|\nu| = \nu^+ + \nu^-$ the total variation of the signed measure ν. Obviously in the example above, $|\nu|$ is given by $|\nu|(E) = \int_E |f|\, d\mu$.

The converse to the situation presented in this example is provided by the Radon-Nikodým theorem. It states that if $\{X, \mathcal{S}\, \mu\}$ is a σ-finite measure space and ν is a σ-finite measure such that $\nu \ll \mu$, then there exists a nonnegative measurable function f on X such that, for each $E \in \mathcal{S}$, $\nu(E) = \int_E f\, d\mu$. This f is unique in the convention given before; any other function with the same property agrees with f almost everywhere. The proof proceeds by reduction to the case for finite measures: write X as the union of a sequence of disjoint sets on which both μ and ν are finite, find the corresponding function f for each set, and put these functions together to get the corresponding "density function" on the whole of X. In the case of a finite measure, f is obtained as the supremum of functions g which are nonnegative, measurable, and satisfy $\int_E g\, d\mu \leqq \nu(E)$ for each E in $\mathcal{S}$. However, taking the supremum of such functions involves an uncountable family of functions so a nonmeasurable function could apparently appear. The proof reduces the operation to a countable one. The resulting function f can be regarded as the derivative of the measure ν with respect to μ and we write $f = d\nu/d\mu$. The theorem can be extended to signed measures. If μ, ν are signed measures we say ν is absolutely continuous with respect to μ if $\nu(E) = 0$ whenever $|\mu|(E) = 0$. Then ν can be written in terms of a derivative with respect to the measure μ by writing $\nu = \nu^+ + nu^-$, finding the derivatives of ν^+ and ν^- separately, and taking $d\nu/d\mu$ as their difference.

The analogy between ordinary derivatives and the Radon–Nikodým derivative $d\nu/d\mu$ goes further. The chain rule applies: if λ, μ, ν are σ finite measures such that $\nu \ll \mu$ and $\mu \ll \lambda$ then $\nu \ll \lambda$ and

$$d\nu/d\lambda = (d\nu/d\mu)(d\mu/d\lambda)$$

where equality is in the sense that this equation must hold almost everywhere (in the sense of λ). If the measure ν is defined by the integral of a function f with respect to μ this theorem takes the following form: if $\mu \ll \lambda$ where μ and λ are σ finite then there exists a measurable function g such that if $f \in L^1(X, \mu)$ then $fg \in L^1(X, \lambda)$ and for each E

$$\nu(E) = \int_E f\, d\mu = \int_E fg\, d\lambda$$

If we take any nonnegative integrable function f on the real line we may take multiple of f so as to get a function $p(x)$ such that $\nu(E) = \int_E p\, dx$ defines a probability measure with $p = d\nu/dm$ as the probability density. The Radon–Nikodým derivative of one probability measure with respect to another is important in connection with conditional probabilities and conditional expectations.

The Radon–Nikodým theorem shows that for a finite measure ν the relation $\nu \ll \mu$ is a genuine continuity property. For since $\nu(E) = \int_E f\, d\mu$ where f is an integrable function we have the result noted in Section VII that $\nu(E) \to 0$ as $\mu(E) \to 0$; more precisely, given $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $\mu(E) < \delta$ we have $\nu(E) < \varepsilon$.

Example 4 in Section VII exhibits the opposite property: the measure μ defined there is concentrated on a set of zero measure (the integer points) and so cannot be given as the

integral of a density or derivative. If $\sum a_n = 1$ then we have a "discrete probability," a frequency occurrence in elementary probability. A discrete probability of the type described and Lebesgue measure are mutually singular.

We can show that given any two σ-finite measures μ and ν, we may decompose one of these, say ν, as $\nu = \nu_0 + \nu_1$ where $\nu_0 \perp \mu$ and $\nu_1 \ll \mu$. This is the Lebesgue decomposition of ν with respect to μ. It is easily proved using the Radon–Nikodým derivative f of $\lambda = \mu + \nu$ with respect to μ, so that $\mu(E) = \int_E f\, d\lambda$. If $A = \{x: f(x) > 0\}$ and $B = \{x: f(x) = 0\}$ we may write

$$\nu_0(E) = \nu(E \cap B), \qquad \nu_1(E) = \nu(E \cap A)$$

and obtain the desired decomposition. So the two constructions of new measures, one using densities and the other measures concentrated on sets of measure zero, together provide the most general σ-finite measures. The σ finiteness is essential to these results, for examples can be constructed where absolute continuity holds but no derivative exists; such examples depend on non-σ-finite measures.

IX. EXTENSIONS OF THE DEFINITION OF MEASURE

In the last section we extended the idea of measure to include measures taking negative values. These arose in a natural way in connection with the integrals of functions taking positive and negative values. If as in Section III we allow functions to take complex values we obtain complex-valued measures. But just as complex-valued measurable functions may be dealt with by considering their real and imaginary parts separately so the properties of a complex-valued measure $\mu = \mu_1 + i\mu_2$, where μ_1 and μ_2 are real measures, and the corresponding integration theory for μ may be deduced from that for μ_1 and μ_2.

More interesting considerations arise when measures are allowed to take values in a vector space. We shall consider measures taking values in the n-dimensional real vector space $\mathbb{R}^n$. (More general spaces could be considered but the essential results are already exhibited in $\mathbb{R}^n$.) An important definition is that of an atomic measure. A set A in a measure space $\{X, \mathcal{S}, \mu\}$ is an atom for μ if $\mu(A) \neq 0$ and every measurable subset of A has measure zero or $\mu(A)$. The *range* of the measure μ is the set of numbers $\{mu(E): E \in \mathcal{S}\}$. Then the theorem of Liapounoff states that the range of a finite nonatomic measure with values in $\mathbb{R}^n$ is closed and convex. More precisely: let $\mu_1, \ldots, \mu_n$ be finite nonnegative nonatomic measures on X. Then the set of points in $\mathbb{R}^n$ of the form $(\mu_1(A), \mu_2(A), \ldots, \mu_n(A))$, with A ranging over the measurable subsets of X, is closed and convex. There are various proofs. In one, we consider the set $W = \{g: 0 \leq g \leq 1\}$ of L^∞ and the mapping T from W to $\mathbb{R}^n$ given by $Tg = (\int g\, d\mu_1, \ldots, \int g\, d\mu_n)$. The set W is convex and closed in the appropriate sense. Let

$$\underline{\lambda} = (\lambda_1, \ldots, \lambda_n) \in T(W)$$

Then if we show that $W_0 = T^{-1}\underline{\lambda}$ contains a characteristic function the result is proved since if E and F are measurable sets then for $0 \leq t \leq 1$

$$(t\mu_1(E) + (1-t)\mu_1(F), \ldots,$$
$$t\mu_n(E) + (1-t)\mu_n(F))$$

lies in $T(W)$. Now for any convex set a point is an extreme point if it cannot be written as the convex combination of other points of the set. There is a general result that in $\mathbb{R}^n$ every closed bounded convex set has extreme points and may be considered as the set of convex combinations of these extreme points. Now W_0 is convex, closed, and bounded; the proof ends with the demonstration that the extreme points of W_0 consist of characteristic functions, using the nonatomic property of the measures μ_i.

The result obtained is nontrivial even for ordinary measures ($n = 1$). It is obviously dependent on the nonatomic property: one could choose X as a single point with measure 1, to get a trivial counterexample. The result outlined above has applications in control theory where one wishes to know that a convex combination of states of the system is again a state of the system. It is also of importance in the investigation of higher dimensional vector spaces.

X. EXTENSIONS OF MEASURES AND LEBESGUE–STIELTJES INTEGRALS

In Section II we saw how starting with a "measure" (the length) defined on intervals we can, using an outer measure, extend this measure to a σ algebra containing the intervals. It is convenient to regard this as a two-stage process. First we extend the definition of length to give a measure on the set of finite unions of intervals, which form a ring, closed under finite unions and set differences. We then extend this measure to a σ ring closed under countable unions and under set differences. If this σ ring contains the whole space it is a σ algebra. This passage from a ring to a σ ring or σ algebra can be done in general and the result is essentially unique. An important example of this construction is provided by the construction of Lebesgue–Stieltjes measures. Take a monotone increasing finite-valued function g and for an interval $I = [a, b)$ define a measure on I by $\mu_g(I) = g(b) - g(a)$. Now, measures have continuity properties since if $I_1 \subseteq I_2 \subseteq \cdots$ we have $\mu(\cup I_i) = \lim \mu(I_i)$. So if $I_i = (a_i, b_i)$ we see that

g must be left-continuous (i.e., $g(x) \to g(x_0)$ whenever $x \to x_0, x < x_0$). Then μ_g is indeed a measure on the ring of finite unions of such intervals and extends to a measure on the Borel sets of the real line. We could, at least for the case g bounded, have chosen intervals of the form $(a, b]$ and g right-continuous. This form of the definition is more common in probability theory. The measure μ_g given by the construction outlined above is the Lebesgue–Stieltjes measure defined by g.

It is clear that points of discontinuity of g correspond to atoms of the measure μ_g, as defined in Section IX. Absolute continuity was defined for functions in Section V and for measures in Section VIII. These definitions are drawn together by the following result. Let g be a monotone increasing absolutely continuous function. Then g defines a measure on a σ algebra. If we assume that this measure has been completed, as defined in Section VII, to get the complete extension μ_g then μ_g is defined on the σ algebra $\mathcal{M}$ of Lebesgue measurable sets and $\mu_g \ll m$. Indeed the Radon-Nikodym derivative of μ_g and the derivative of g (which exists a.e.) correspond: for any a, b we have $g(b) - g(a) = \int_a^b g'\,dt$ and so $g' = d\mu_g/dm$. In this case integrals with respect to μ_g reduce to integrals with respect to Lebesgue measure:

$$\int_E f\,d\mu_g = \int_E f g'\,dt$$

These results are especially important in probability theory where the finiteness of the measure simplifies the results. Indeed let $\{X, \mathcal{S}, \mu\}$ be a probability measure space and f a finite-valued measurable function, so f is a random variable. Define a function F by $F(x) = \mu f^{-1}(-\infty, x]$ (i.e., $F(x) = \mu\{t\colon f(t) \leqq x\}$, so that $F(-\infty) = 0 \cdot F(\infty) = 1$). Then F is the distribution function of f and is a monotone increasing right-continuous function. The measure μ_F it defines agrees with the measure μf^{-1} on the intervals of $\mathbb{R}$.

XI. THE RADON–NIKODÝM PROPERTY FOR BANACH SPACES

In this section we consider measures taking their values in a Banach space, that is a normed space which is complete (the examples of L^p, L^∞ were considered in Section IV). This is more general than the finite dimensional case considered in Section X. We consider first the question of integrating functions taking values in a Banach space, with respect to a real measure. If we then set $m(A) = \int f\,d\mu$ we see that such a theory of integration gives rise to a Banach space–valued measure. Which measures are formed in this way depends on the Banach space in question, since the existence of such a function f implies that a version of the Radon–Nikodým theorem holds. The most useful version of this integral is the Bochner integral. First, we describe the details of such measures. Let Y be a Banach space with norm $\|.\|$. We will suppose that we have a space X with measurable sets $\mathcal{S}$. Then $m\colon \mathcal{S} \to Y$ is a vector measure if $m(\varnothing) = 0$ and whenever $\{A_i\}$ is a countable family of disjoint sets of $\mathcal{S}$ then $m(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty m(A_i)$ where this sum is norm convergent, that is, the sequence of vectors $\sum_{i=1}^n m(A_i)$ converges in the normed space $(Y, \|.\|)$ as $n \to \infty$. We will suppose that the space X is equipped with a finite measure μ: for some purposes it is more convenient to assume, as we will, that $\mu(X) = 1$, that is: μ is a probability measure. The *average range* of m is the set in Y given by $AR(m) = \{m(A)/\mu(A)\colon A \in \mathcal{S}, \mu(A) > 0\}$. As in the scalar case we say the vector-valued measure m is *absolutely continuous* with respect to μ, $(m \ll \mu)$ if $\mu(A) = 0$ implies $m(A) = 0$ (zero vector).

Then to set up a theory of integration we consider simple functions as before of form $f = \sum_{i=1}^n x_i \chi_{A_i}$ with $x_i \in Y$, $A_i \in \mathcal{S}$. Then for any set $A \in \mathcal{S}$, we can define for such an $f \int_A f\,d\mu = \sum_{i=1}^n x_i \mu(A \cap A_i)$. Then generally a function $f\colon X \to Y$ is *Bochner integrable* if there is a sequence $\{f_n\}$ of simple functions with $\lim_{n\to\infty} f_n(x) = f(x)$ a.e. (μ) and $\lim_{n\to\infty} \int \|f(x) - f_n(x)\|\,d\mu = 0$. This provides an unambiguous definition if we set $\int f\,d\mu = \lim \int f_n\,d\mu$, and we then write $f \in L_Y^1(X, \mathcal{S}, \mu)$, and we say $\int f\,d\mu$ is the *Bochner integral* of f. Regarding measurability, a Y-valued function is said to be *strongly measurable* if it is the limit a.e. of simple functions. A weaker requirement is that each scalar valued function $F \circ f$ where F is a continuous linear functional on Y should be measurable in the usual sense. For functions f with $\|f\|$ integrable and which satisfy a condition on the range (always true for integrable functions), the definitions are equivalent. Much of the theory extends; for example, we have a dominated convergence theorem: we require $\|f_n(x)\| \leq g(x)$ a.e. (μ). where g is integrable and $\lim f_n(x) = f(x)$ a.e., then f is integrable, $\int f\,d\mu = \lim \int f_n\,d\mu$ and $\lim \int \|f_n - f\|\,d\mu = 0$ (i.e. f_n converges to f in the mean).

The Radon–Nikodým property turns out to be very important for when considering such vector-valued measures and functions; whether it holds depends on the geometry of Y. Let K be a closed bounded convex set in the Banach space Y. Then K has the Radon–Nikodým property (RNP) for $\{X, \mathcal{S}, \mu\}$ if for any Y-valued measure m which is absolutely continuous with respect to μ and whose average range $AR(m)$ lies in K there exists a function $f \in L_Y^1(X, \mathcal{S}, \mu)$ such that $m(A) = \int_A f\,d\mu$ for each $A \in \mathcal{S}$. More generally, if E is a closed convex (possibly unbounded) set of Y (e.g., $E = Y$), then E has the RNP for $\{X, \mathcal{S}, \mu\}$ if each closed bounded subset K of E has the RNP for $\{X, \mathcal{S}, \mu\}$. Finally, K has the RNP

if it has the RNP for each $\{X, \mathcal{S}, \mu\}$ for μ a probability measure. An example of a space *without* the RNP is at hand.

EXAMPLE 1. [Bourgin]: Let $X = [0, 1]$, $\mathcal{S}$ = Lebesgue measurable sets, μ = Lebesgue measure. Let $Y = L^1[0, 1]$ and define m by $m(A) = \chi_A$ for each $A \in \mathcal{S}$. Then $m \ll \mu$, $AR(m)$, lies in the closed unit ball of Y but Y has not the RNP. For if it had, there would exist a function $f \in L^1_Y(\mu)$ with $m(A) = \int_A f\, d\mu$ for each $A \in \mathcal{S}$. So for each x, $f(x)$ is a real-valued measurable function taking the value $f(x)(s)$ at $s \in [0, 1]$. Let I be an interval in $[0, 1]$, then for all $A \in \mathcal{S}$

$$\begin{aligned}\int_A \left(\int \chi_I(s) f(t)(s)\, ds \right) dt &= \int \chi_I(s) \left(\int_A f(t)\, dt \right) ds \\ &= \int \chi_I(s)\, m(A)(s)\, ds \\ &= \int \chi_I(s)\, \chi_A(s)\, ds \\ &= \int_A \chi_I(t)\, dt\end{aligned}$$

So

$$\int_I f(t)(s)\, ds = \int \chi_I(s) f(t)(s)\, ds = \chi_I(t)$$

outside a set of measure zero, and for such t **not** in I, $\int_I f(t)(s)\, ds = 0$. Allowing I to vary over all subintervals I_r of $[0, 1]$ with rational end-points we get an exceptional set of measure zero. Any set on which the function $f(t)$ is positive can be approximated by an interval, and choosing a subinterval I_r of this interval with $t \notin I_r$ we see that $f(t)$ must vanish a.e. for almost all t. This contradicts the fact that $\int_A f\, d\mu = \chi_A \neq 0$ for A a set of positive measure.

The RNP is closely related to the convergence of martingales, which we now describe. Let $\{\mathcal{S}_n\}$ be a sequence of sub σ-algebras of $\mathcal{S}$, with $\mathcal{S}_n \subseteq \mathcal{S}_m$ whenever $n < m$. Suppose $f_n \in L^1_Y(X, \mathcal{S}, \mu)$ for each n. Suppose also that the functions f_n are strongly $\mathcal{S}_n$ measurable for each n and that $\int_A f_n\, d\mu = \int_A f_m\, d\mu$ for each A in $\mathcal{S}_n$ provided $n < m$. Then the sequence $\{f_n, \mathcal{S}_n\}$ is a Y-valued martingale. A closed bounded convex set K in Y has the *martingale convergence property* (MCP) for $\{X, \mathcal{S}, \mu\}$ if whenever $\{f_n, \mathcal{S}_n\}$ is a martingale such that $\cup_{n=1}^{\infty} \mathcal{S}_n$ generates $\mathcal{S}$ and $f_n \in L^1_K(X, \mathcal{S}_n, \mu)$ for each n then there exists $f \in L^1_K(X, \mathcal{S}, \mu)$ such that $\lim_{n\to\infty} \|f_n(x) - f(x)\| = 0$ a.e., (μ). The corresponding statements for closed convex sets and the definition of "Y has MCP" follow exactly as for RNP. In fact a closed bounded convex set has the RNP if, and only if, it has the MCP. To see that a closed convex set K has the MCP provided it has the RNP we let $m_n(A) = \int_A f_n\, d\mu$ where μ is a probability measure and $\{f_n, \mathcal{S}_n\}$ form a martingale and $A \in \mathcal{S}_n$. Since μ is a probability measure $AR(m_n)$ lies in K. Then $\{m_n(A)\}$ converges for each A in $\cup \mathcal{S}_n$. By a limiting argument we get $\lim m_n(A) = m(A)$, which is a measure by the Vitali-Hahn-Saks Theorem. By construction $m_n \ll \mu$ and in the limit $m \ll \mu$ and $AR(m) \subseteq K$. So by the RNP there exists $f \in L^1(X, \mathcal{S}, \mu)$ such that $\int_A f\, d\mu = m(A)$ for each A in $\mathcal{S}$. By the martingale property $\int_A f\, d\mu = \int_A f_n\, d\mu$ for each $A \in \mathcal{S}_n$ and by a theorem of Lévy we deduce $\lim_{n\to\infty} \|f_n(x) - f(x)\| = 0$ a.e. (μ). This theorem of Lévy uses conditional expectations whose existence depends on the classical Radon–Nikodým result given in Section VII.

A third property of Banach spaces which turns out to be related is that of dentability. A bounded set D in Y in *s-dentable* if for each $\varepsilon > 0$ there exists x_ε in D with $x_\varepsilon \notin s\text{-}co(D \backslash U_\varepsilon(x_\varepsilon))$, where $U_\varepsilon(y)$ denotes the ball radius ε and center Y and

$$s\text{-}coB = \left\{ \sum_{i=x}^{\infty} \alpha_i x_i : x_i \in B, \alpha_i \geq 0, \sum_{i=1}^{\infty} \alpha_i = 1 \right\}.$$

Replacing *s-co* by closed convex hull, we get the stronger definition of D dentable. Drawing a diagram we see that D dentable implies that D is in some sense rotund for some part of its boundary. Dentability can be defined equivalently in terms of *slices*. For a bounded set D in Y, the *slice* $s(D, f, \alpha) = \{x : x \in D, f(x) > \sup_{y\in D} f(y) - \alpha\}$ where $\alpha > 0$, and f is a continuous linear functional on Y. Then it is an easy consequence of the Hahn–Banach theorem that D is dentable if, and only if, it has slices of arbitrarily small diameter. Now it follows easily from the definitions that if D is dentable then it is s-dentable. But the converse is not true, as the following example shows.

EXAMPLE 2. Let $Y = C[0, 1]$ the Banach spaces of functions continuous on $[0, 1]$ and with the norm $\|f\| = \max |f(x)|$ and let D be the unit ball $U_1[0]$ of Y. Then with $f(x) \equiv 1$ as x_ε, for any $\varepsilon > 0$, we see that D is s-dentable. However, D is not dentable for if $f \in D$ and for any fixed n we choose function $f_n^1, \ldots, f_n^n$ with $f_i^n(t) = f(t)$ for $t \notin [i - 1/n, i/n]$ and $|f_i^n(t) - f(t)| > 1/2$ for some $t \in (i - 1/n, i/n)$, for each i. Then $\|f_i^n - f\| > 1/2$ but $\| \sum_{i=1}^n 1/n\ f_i^n - f\| \leq 2/n$. So $f \in \overline{co}(D \backslash U_{1/2}(f))$. So D is not dentable. However, it can be shown by a geometrical argument that if K is a closed convex set in Y with interior K (int K) nonempty then if K is not dentable int K is not s-dentable. See Davis and Phelps for details. So every bounded set of Y is dentable if, and only if, every bounded set is s-dentable. This turns out to be the important property. Indeed, if a closed bounded convex set K has every subset s-dentable then K has the RNP.

To outline the proof: consider the special case when m is in the form

$$m(A) = \sum_{i=1}^{n} x_i \mu(A \cap B_i),$$

with $\{B_i\}$ being a partition of X into sets of positive measure and with $x_i \in K$. Then for $\mu(A) > 0$ and $A \subseteq B_i$ we have

$$\frac{m(A)}{\mu(A)} = x_i = \frac{m(B_i)}{\mu(B_i)}.$$

Set $f = \sum_{i=1}^{n} x_i \chi_{B_i}$; then for each A in $\mathcal{S}$ we have

$$\int_A f \, d\mu = \sum_{i=1}^{n} x_i \mu(A \cap B_i) = m(A).$$

So $f = dm/d\mu$ and K has the RNP. In general m will not be of this simple form, but we can approximate it by a sequence of such measures, obtaining a convergent sequence of such derivatives f_n. So we need to know that we can partition X into sets B_i such that for subsets B of B_i, with $\mu(B) > 0$, and for a suitable x in K we have $\|m(B)/\mu(B) - x_i\| < \varepsilon$. Then we use a sequence of such partitions with ε's tending to zero to obtain the approximation. So we let $E = \{m(B)/\mu(B)\colon \mu(B) > 0, B \subseteq B_i\}$, a subset of K as $AR(m) \subseteq K$. So E is s-dentable and so we can find $x_i \notin s\text{-}co(E \backslash U_\varepsilon(x_i))$, with $x_i \in E$, $x_i = m(B)/\mu(B)$, say. Suppose we can find sets C in B_i with $\|m(C)/\mu(C) - x_i\| \geq \varepsilon$. Maximizing such a family of disjoint sets C we could write $B_i = \cup_{i=1}^{\infty} C_i$ and then

$$x_i = \frac{m(B)}{\mu(B)} = \sum_{j=1}^{\infty} \frac{\mu(C_j)}{\mu(B} x_j$$

with x_j

$$= \frac{m(C_j)}{\mu(C_j)} \in E \quad \text{and} \quad \sum_{i=1}^{\infty} \frac{\mu(C_i)}{\mu(B)} = 1,$$

contradicting the s-dentability of E.

It can be shown fairly easily that if K has the MCP, then its subsets are s-dentable (Bourgin). Suppose not, so there is a subset D of K which is not s-dentable. Then a nonconvergent martingale can be constructed inductively, using the interval $[0, 1]$, $\mu =$ Lebesgue measure. So there exists a positive ε such that for each $x \in D$, $x \in s\text{-}co(D \backslash \cup_\varepsilon (x))$. At the first stage choose $x_0 \in D$, define $f_0 = x_0$ a constant function and let $\mathcal{S}_0 = \{[0, 1), \varnothing\}$, the minimal σ-algebra. By the property of D there exist $\{t_j\colon 0 < t_j < 1, \Sigma t_j = 1\}$ and points y_j of D with $\|x_0 - y_j\| \geq \varepsilon$, $\Sigma t_j y_j = x_0$. Partition $[0, 1]$ into half-open intervals B_j, one for each t_j with $\mu(B_j) = t_j$. Let $\mathcal{S}_1$ be the σ-algebra generated by $\{B_j\}$, and $f_1 = \sum_{j=1}^{\infty} y_j \chi_{B_j}$. At the next stage of the induction we partition each B_j to get a larger σ-algebra, and a function f_2. By the non-s-dentable property of D, this martingale so constructed will not converge.

This result establishes the equivalence of the properties RNP, MCP, dentability of subsets and s-dentability of subsets for a Banach space Y. Such spaces, to some extent, have the nice properties of finite dimensional spaces. Among the various other properties of such spaces is the fact that they possess the Krein–Milman property referred to in Section IX; that is, for any closed bounded convex set K in Y, K is the closed convex hull of its extreme points. Indeed we saw earlier that the space $C[0, 1]$ has a unit ball $U_1[0]$, which is not dentable: $U_1[0]$ has just two extreme points, the constant functions $+1$ and -1, so it does not possess the KMP. To prove the KMP it is sufficient to show that every bounded convex set has an extreme point. For dentable sets this can be done using the fact that they have slices of arbitrarily small diameter. A nested decreasing sequence of such slices is found the intersection of which yields the desired extreme point. For further details, related results, and references see Bourgin (1983) and Phelps (1988). We have seen that the spaces $C[0, 1]$, $L^1[0, 1]$ have not the RNP. However, the spaces L^p of Section IV for which $1 < p < \infty$ have the RNP, as have all reflexive spaces. It is not known whether the KMP implies the RNP. Another important property of finite-dimensional spaces is that convex functions are differentiable almost everywhere. To see how this extends, we say that the real-valued function f on a Banach space Y is *Frechét differentiable* at y if there exists a linear functional $\phi'(y)$ on Y such that for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\|\varphi(y + x) - \varphi(y) - \varphi'(y)(x)\| \leq \varepsilon \|x\|$ whenever $\|x\| \leq \delta$. Then the space Y is said to be an *Asplund space* if every continuous convex function f on a nonempty open set D in Y is Frechét differentiable at each point y of some dense set E of D where the set E is a G_δ set, that is: $E = \cap_{i=1}^{\infty} G_i$ when the sets G_i are dense open sets of D. Then the space Y is an Asplund space, if, and only if, the Banach space Y^* of continuous linear functionals on Y has the RNP.

XII. MEASURE AND FRACTALS

Fractals are often introduced and thought of in pictorial terms. However, the underlying measure theory is important. We start with Hausdorff measure on $\mathbf{R}$. This is defined in two steps: first define the "approximating measure":

$H^*_{s,\delta}(A) = \inf \sum \ell(I_k)^s$ where the infimum is taken over all coverings of the set A by intervals $\{I_k\}$ with $\ell(I_k) \leq \delta$. Then let $H^*_s(A) = \lim H^*_{s,\delta}$ as $\delta \to 0$. This limit exists and defines Hausdorff outer measure. This *is* an outer measure, $H^*_s(\{x\}) = 0$ for each point, $H^*_s(A + x) = H^*_s(A)$, for each set A, so it is invariant under translation and, significantly, it "scales" in this way: $H^*_s(kA) = k^s H^*_s(A)$ for any positive k. Then H^*_s is a *metric outer measure*,

that is if two sets A and B are a positive distance apart, $H_s^*(A \cup B) = H_s^*(A) + H_s^*(B)$. Then Borel sets are H_s^* measurable in the usual sense: the proof requires a little care in showing that intervals are H_s^* measurable. It is easily seen that H_1^* is just Lebesgue measure. Hausdorff measure, H_s, defined by H_s^*, is used to analyze sets of zero Lebesgue measure. For it can be shown fairly easily that if $H_s^*(A) < \infty$ then $H_q^*(A) = 0$ for $q > s$. So if $0 < H_s^*(A) < \infty$ then $H_q^*(A) = \infty$ for $q < s$. So we can define (Hausdorff) dimension of a set E by: $\dim(E) = \inf\{s: H_s^*(E) = 0\}$. This provides a dimension, usually noninteger, for Borel sets in $\mathbf{R}$. All this may be generalized with, in the definition, $\ell(I_k)^s$ being replaced by $h(\ell(I_k))$ with a suitable function h, for example a monotonic increasing function $h(t)$ for $t \geq 0$ with $h(0+) = h(0) = 0$.

Hausdorff dimension can be difficult to calculate. The real numbers have dimension = 1, finite sets of points have dimension zero. For the Cantor Set, P, described in Section II, the calculation is straightforward. This set is "self-similar" in the sense that the subsets in the closed intervals $J_{1,1}$, $J_{1,2}$ left after the first "open third" $I_{1,1}$ has been removed are copies of the Cantor set scaled by 1/3 and translated. So using the "scaling property" of H described above, we get $H_s(P) = \frac{2}{3^s} H_s(P)$ and so $s = \frac{\log 2}{\log 3}$, provided we can show that $0 < H_s(P) < \infty$. This part needs care in transforming a arbitrary covering of P into coverings of the subsets of $J_{1,1}$, $J_{1,2}$ mentioned in Section III, and using the metric outer measure property. This method and result generalize immediately to the "Cantor-like set" P_ξ obtained when we remove, not the middle third, but a central open interval of positive length $1 - 2\xi$ at the first stage, with the residual intervals being of length ξ at stage one, ξ^2 at stage two, etc. The same argument gives the Hausdorff dimension of P_ξ as $-\frac{\log 2}{\log \xi}$, a number between 0 and 1. This shows that sets in $\mathbf{R}$ exist with all possible positive values of Hausdorff dimension. In forming the Cantor set we removed numbers whose expansion to base 3 contained a 1. If we remove instead those containing a 2 or those containing a zero, we get somewhat different sets whose Hausdorff dimensions were found by Best. Similar results for expansions to other bases are to be found in Weymann. Sets of finite positive H_s measure are called s-sets, or *fractals* in this context, though some writers use "fractals" to refer to sets which are self-similar at all levels of magnification.

This generalizes easily to two dimensions: from a closed square one deletes a central open cross to leave four closed squares placed at the corners: these are then similarly reduced, etc. The resulting "Cantor square" along with many examples of self-similar sets are considered in the very accessible book of Lauwerier. In two dimensions one may consider the Hausdorff dimension of a curve. With some restrictions to avoid pathological space-filling curves one finds that the curve C of length L has dimension one and Hausdorff measure $H_1(C) = L$ (Falconer, p. 24). This relates to the functions of bounded variation discussed in Section V. In three dimensions we may consider the motion of minute particles in a liquid. This erratic motion is modeled probalistically by Brownian motion. In Falconer (p. 144) we have that Brownian paths are s-sets of dimension one, with probability one. Applications to many areas are to be found in Mattila, together with a comparison of different definitions of "dimension," and which also contains a sizeable bibliography.

SEE ALSO THE FOLLOWING ARTICLES

CALCULUS • COMPLEX ANALYSIS • CONVEX SETS • DIFFERENTIAL EQUATIONS, ORDINARY • FRACTALS • INTEGRAL EQUATIONS

BIBLIOGRAPHY

Best, E. P. (1992). On sets of fractional dimension III. *London Math. Soc.* **47**(2), 436–454.

Bourgin, R. D. (1983). "Geometric Aspects of Convex Sets with the Radon–Nikodým Property" (Lecture Notes in Mathematics, 993), Springer-Verlag, Berlin.

Cohn, D. L. (1980). "Measure Theory," Birkhäuser, Boston and Basel.

Davis, W. J., and Phelps, R. R. (1974). The Radon-Nikodým property and dentable sets in Banach spaces. *Proc. Amer. Math. Soc.* **45,** 119–122.

De Barra, G. (1981). "Measure Theory and Integration," Ellis Horwood, Chichester, England.

Diestel, J., and Uhl, J. J. (1977). "Vector Measures," Mathematical Surveys 15. *Amer. Math. Soc.*, Providence, RI.

Dinculeanu, N. (1967). "Vector Measures," Pergamon, London.

Dunford, N., and Schwartz, J. T. (1958). "Linear Operators, Part I," Interscience Publications Inc., New York.

Falconer, K. J. (1985). "The Geometry of Fractal Sets," Cambridge Univ. Press, Cambridge, U.K.

Gelbaum, B. R. (1982). "Problems in Analysis," Springer-Verlag, New York.

Hewitt, E., and Stromberg, K. (1965). "Real and Abstract Analysis," Springer-Verlag, New York.

Lauwerier, H. (1991). "Fractals," Penguin Books, Baltimore.

Mattila, P. (1995). "Geometry of Sets and Measures in Euclidean Spaces, Fractals and Rectifiability," Cambridge Univ. Press, Cambridge, U.K.

Pesin, I. N. (1970). "Classical and Modern Integration Theories," Academic Press, New York.

Pfeffer, W. F. (1977). "Integrals and Measures," Dekker, New York.

Phelps, R. R. (1988). "Convex Functions. Monotone Operators and Differentiability," Lecture Notes, University of Washington, Seattle.

Rogers, C. A. (1970). "Hausdorff Measures," Cambridge Univ. Press, London and New York.

Rudin, W. (1966). "Real and Complex Analysis," McGraw-hill, New York.

Weymann, H. (1971). Das Hausdorff-Mass von Cantormengen, *Math. Ann.* **193,** 7–20.

Wheeden, R. L., and Zygmund, A. (1977). "Measure and Integral," Dekker, New York.



Nonlinear Programming

Jon W. Tolle
University of North Carolina at Chapel Hill

I. Overview
II. Theoretical Aspects
III. Computation
IV. Applications

GLOSSARY

Constraint functions Functions that are used to define the feasible set.
Decision vector Vector of unknown variables on which the objective function is defined.
Feasible set Set of decision vectors that satisfy the constraints and over which the objective function is defined.
Global optimal solution Particular choice of the decision vector that solves the optimization problem.
Lagrangian function Function of the decision vector and the multipliers with critical points that are related to optimal solutions.
Local optimal solution Particular choice of the decision vector that yields the least value of the objective function in some relatively open subset of the feasible set.
Multipliers Coefficients of the constraint gradients in the first-order necessary conditions.
Objective function Function that is to be minimized.
Sequential quadratic programming Computational technique for solving nonlinear programs.

NONLINEAR PROGRAMMING is the study and solution of an optimization problem in which a nonlinear function of several variables is minimized over the solution set of a finite number of functional inequalities and equations. The problem can be written as:

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to:} && g_i(x) \le 0, \quad i = 1, \ldots, M \\ &&& h_j(x) = 0, \quad j = 1, \ldots, P \\ &&& x \in \Re^N, \end{aligned}$$

where the objective function f and the constraint functions, g_i and h_j, are (usually) continuously differentiable functions. The components of the vector x are called the decision variables. Such nonlinear problems find a wide range of applicability in the physical and social sciences, engineering, and the decision sciences.

I. OVERVIEW

A. Optimization and Nonlinear Programming

A significant proportion of the mathematical models formulated and studied by scientists and engineers involve some form of optimization. For physical scientists, nature seems inherently to be an optimizing force; the laws of conservation and equilibrium can be interpreted as

Encyclopedia of Physical Science and Technology, Third Edition, Volume 10
Copyright © 2002 by Academic Press. All rights of reproduction in any form reserved.

optimality conditions. Engineers build controlling devices that perform with minimal energy expenditure and design plant schedules that maximize some measure of production efficiency. Economists and other social scientists theorize that much of human behavior and development, whether individual or collective, is guided by optimization processes. Finally, modern managerial scientists use optimization methods to organize and reduce large amounts of data to make rational decisions on policy issues in the public and private sectors.

Mathematically, the most general form of an optimization model can be characterized by a set X and a real-valued function f defined on X. The set X, often called the *feasible set*, consists of the possible realizations of the model. For example, X could consist of a set of permissible trajectories for a satellite launch vehicle, the possible flight schedules for an international cargo airline, or a set of possible purchases by a consumer in a given market. Each $x \in X$ is a vector, the components of which are termed the *decision variables* of the model. The function f, called the *objective function*, is a measure by which the vectors in X are compared. Thus for these examples, $f(x)$ might represent the energy required to traverse trajectory x, the total cost for the implementation when x is an airline schedule, or the value of a consumer's purchase when x is the vector with components that are quantities of commodities.

The problem to be solved in an optimization model is

$$\text{minimize} \quad f(x), \quad x \in X.$$

That is, an $x^* \in X$ is sought such that for all other $x \in X$, $f(x^*) \leq f(x)$. The fact that the problem is formulated as a minimization rather than a maximization is not important because the above is mathematically equivalent to

$$\text{maximize} \quad -f(x), \quad x \in X.$$

The problem as stated is much too general to yield useful information. In order to be able to analyze the problem, more must be known about the properties of the set X and the function f. The general field of optimization is broken down into subfields according to these properties. Nonlinear programming is one of these subfields; traditionally, it refers to the case in which the set X is assumed to be functionally prescribed. That is, X is the intersection of a finite number of sets each of which is the solution set of a functional equation or inequality. The equations or inequalities that prescribe the feasible set are called constraints. In an optimization model, an inequality constraint may be used to represent the limits on an available resource or may be a production quota that must be exceeded. An equality constraint might represent the requirement that a consumer spend or save all of his or her income or that a trajectory begin from a designated location.

Using a functional representation of the constraint set the standard nonlinear programming problem, hereafter denoted NLP, can be formulated as

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to:} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, M, \\ & h_j(x) = 0, \quad j = 1, \ldots, P, \\ & x \in \Re^N. \end{aligned}$$

The functions g_i and h_j are called the *constraint functions* and, like f, are generally assumed to be continuously differentiable. Nonlinear programming problems are distinguished from the special types of optimization problems called *linear programming* problems, in which the objective and constraint functions are affine. It is also conventional to separate nonlinear programming, which requires the decision vectors to be finite-dimensional, from problems such as those of optimal control, in which X is a subset of an infinite-dimensional space. However, the computation of the solutions to these latter problems depends on the solution of finite-dimensional approximations (see Section IV.C).

B. The Solution of a Nonlinear Program

A vector, $x^* \in X$, satisfying $f(x^*) \leq f(x)$ for all $x \in X$ is called a (global) optimal solution to NLP and the corresponding number $v^* = f(x^*)$ is called the optimal value of NLP. One of the distinguishing characteristics of nonlinear programming is that local optimal solutions are possible. Hence, $x^L \in X$ is a local optimal solution if there is an $\epsilon > 0$ such that $f(x^L) \leq f(x)$ for all $x \in X \cap \{x : |x - x^L| < \epsilon\}$ ($|z|$ represents the Euclidean length of the vector z). Unfortunately, for many nonlinear programs, global optimal solutions cannot be identified until all local optimal solutions are known. Hereafter, the term "optimal solution" will refer to a local solution unless specified otherwise.

The major theoretical questions to be answered concerning a given NLP relate to the existence, characterization, and stability of solutions. Existence refers to the determination of conditions on the objective and constraint functions under which global and local solutions are guaranteed to exist. The characterizations of a solution are theoretical conditions on a point x that are either necessary or sufficient for it to be a global or local solution. These characterizations are important for identifying the solution(s) and its properties and in the construction of algorithms for computing the solution. Stability refers to the sensitivity of the solution with respect to the perturbation of the parameters which define the objective and constraint functions. Stability is important in the practical

use of optimization because these parameters are usually not known precisely. For instance, they might be the observed mean of historical data or they might be the result of a physical measurement with its attendant errors. It is desirable, in these cases, to have assurances that the optimal solution and optimal value are not wildly inaccurate as a result of small errors in the determination of these parameters.

In most cases, it is the numerical values of an optimal solution that are ultimately needed. Thus, the computational question of how best to numerically approximate an optimal solution is equally important to the questions of theory. A computationally efficient numerical algorithm that will yield a good approximate solution to NLP must be available. Consequently, the techniques of numerical analysis and computational mathematics play a major role in nonlinear programming.

The following simple one-dimensional example illustrates the major theoretical questions in nonlinear programming:

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to:} && x - \alpha \geq 0, \\ &&& x - \beta \leq 0, \end{aligned}$$

where f is twice differentiable and $0 < \alpha < \beta$. Since the objective function is continuous and $X = [\alpha, \beta]$ is compact, the existence of a global solution is guaranteed. Elementary calculus shows that any local minimum point must satisfy $f'(x) = 0$ if $\alpha < x < \beta$, $f'(x) \geq 0$ if $x = \alpha$, or $f'(x) \leq 0$ if $x = \beta$. These necessary conditions for optimality are examples of the characterizations of the solution. Note that $\{\alpha < x < \beta, f'(x) = 0, f''(x) > 0\}$ is a sufficient (but not necessary) set of conditions for x to be a local minimum. For an example of the ideas of stability, specify $f(x) = xe^{-x}$, $\alpha = 1/2$, and $\beta > 1$. The global optima are

$$x^* = \begin{cases} \alpha, & \beta \leq \beta^*, \\ \beta, & \beta \geq \beta^*, \end{cases}$$

where $\beta^* > 1$ satisfies

$$\beta^* e^{-\beta^*} = (1/2)e^{-(1/2)}.$$

Moreover the optimal value as a function of β is

$$v^*(\beta) = \begin{cases} (1/2)e^{-(1/2)}, & \beta \leq \beta^*, \\ \beta e^{-\beta}, & \beta \geq \beta^*. \end{cases}$$

$v^*(\beta)$ is a continuous function of the parameter β and

$$\frac{dv^*}{d\beta} = \begin{cases} 0, & \beta < \beta^*, \\ (1-\beta)e^{-\beta}, & \beta > \beta^*. \end{cases}$$

Thus, the global optimal value is differentiable except when $\beta = \beta^*$.

C. Special Types of Nonlinear Programs

The set of nonlinear programs can also be subdivided further into classes associated with special properties of the objective and constraint functions. Some of the more common classes are described in this section.

In theory, the simplest nonlinear program is the unconstrained program (i.e., the problem in which there are no constraints) where the objective function is minimized over all of $\Re^N$. These problems are important because they include many of the important *least-square models* and also because the computational algorithms for solving them are used as the foundations for algorithms for solving the more general problems. Among the constrained nonlinear programs, the easiest ones to deal with are the *convex programs*. These problems have a unique global optimal solution and no local solutions; much of the theory is analogous to that found in linear programming. Of special interest among the convex programs are the *convex quadratic programs*. In these problems, the objective function has the form

$$f(x) = \frac{1}{2}x^t Ax + a^t x,$$

where A is an $N \times N$ symmetric positive definite matrix and a is a fixed N-vector, and all of the constraints functions are affine. As will be seen, one approach to solving a general NLP is to approximate it by a quadratic program.

Two classes of problems that are currently on the forefront of research in the field of nonlinear programming are nondifferentiable and large-scale programs. *Nondifferentiable programs* are those for which the objective function is only piecewise differentiable. This class has many applications, one of the most important of which is the case in which f is itself the solution of an optimization problem; for example,

$$f(x) = \max_{\alpha \in A}\{(c_\alpha)^t x\},$$

where $\{c_\alpha\}$ is a family of vectors. The term *large-scale problems* refers to problems in which some or all of the parameters, N, M, and P are large. The theory for this class is not changed but special computational techniques are necessary to ensure that solving such a problem is feasible in terms of computer time and storage.

Although there are important applications and results associated with each of these special types of nonlinear programs, the theory and computation presented in the remainder of this article deals primarily with the general problem, the major exception is the specialization of the theory to convex programs.

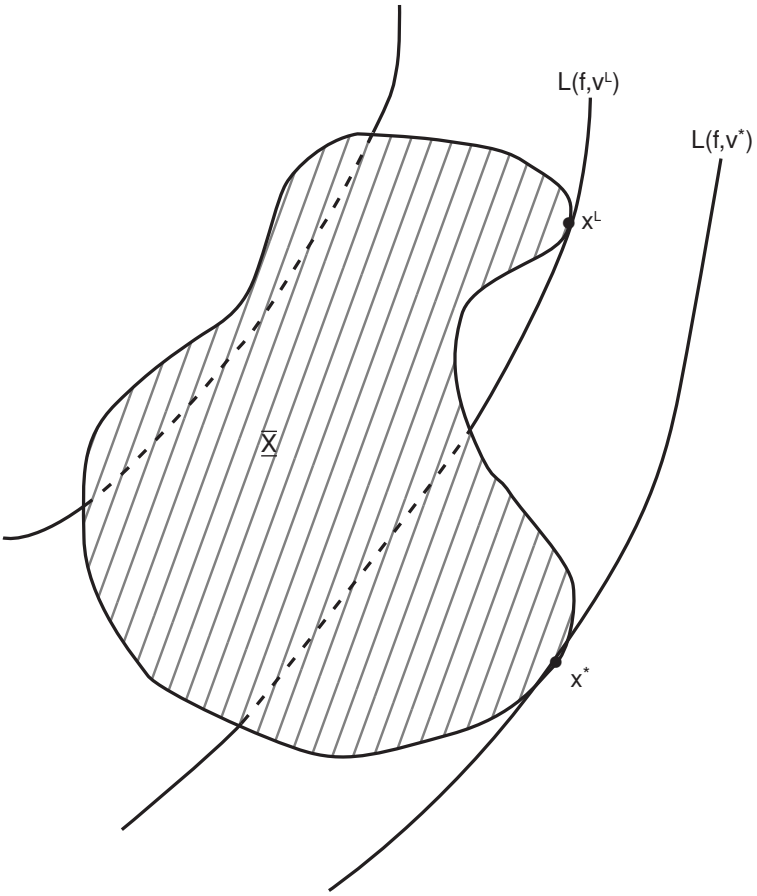


FIGURE 1 Level sets of the objective function with local and global minima.

II. THEORETICAL ASPECTS

A. The Geometry

The theoretical part of nonlinear programming is based on the geometry of the feasible set X and the underlying geometry of the objective function. This geometry can be used to motivate the basic theorems of nonlinear programming (presented later in this section) and also the algorithms used to solve the problems.

The *level sets* of a function will be important in the following development. Let w be a real-valued function defined on $\Re^N$. For each real number α, the following level sets are defined:

$$L(w, \alpha) = \{x : w(x) = \alpha\},$$

$$L^-(w, \alpha) = \{x : w(x) \leq \alpha\}.$$

With this notation, the feasible set for NLP can be written as a finite intersection of these level sets:

$$X = \left\{ \bigcap_{i=1}^{M} L^-(g_i, 0) \right\} \cap \left\{ \bigcap_{j=1}^{P} L(h_j, 0) \right\}.$$

Because the f, g_i, and h_j are assumed to be continuous, the level sets and, therefore X, will all be closed sets.

Now v^* is an optimal value for NLP if the set

$$X^* = L(f, v^*) \cap X$$

is nonempty and for each $v < v^*$, the sets

$$L(f, v) \cap X$$

are empty. Each $x^* \in X^*$ is then an optimal solution of NLP. Similarly, $x^l \in X$ is a local optimal solution for NLP with local optimal value $v^l = f(x^l)$ if for some $\epsilon > 0$ the set

$$L(f, v) \cap X \cap \{x : |x - x^l| < \epsilon\}$$

is empty for each $v < v^l$. An example of the geometry is illustrated in Fig. 1.

The set of optimal solutions is simplified when the feasible set and the level sets $L^-(f, \alpha)$ are *convex* sets. Mathematically, a subset $C \subset \Re^N$ is convex if, for every pair of vectors x and y in C and all $\lambda \in [0, 1]$, the vector $\lambda x + (1-\lambda)y$ is also in C. Thus a convex set is one that has no indentations or protuberances. If the feasible set and

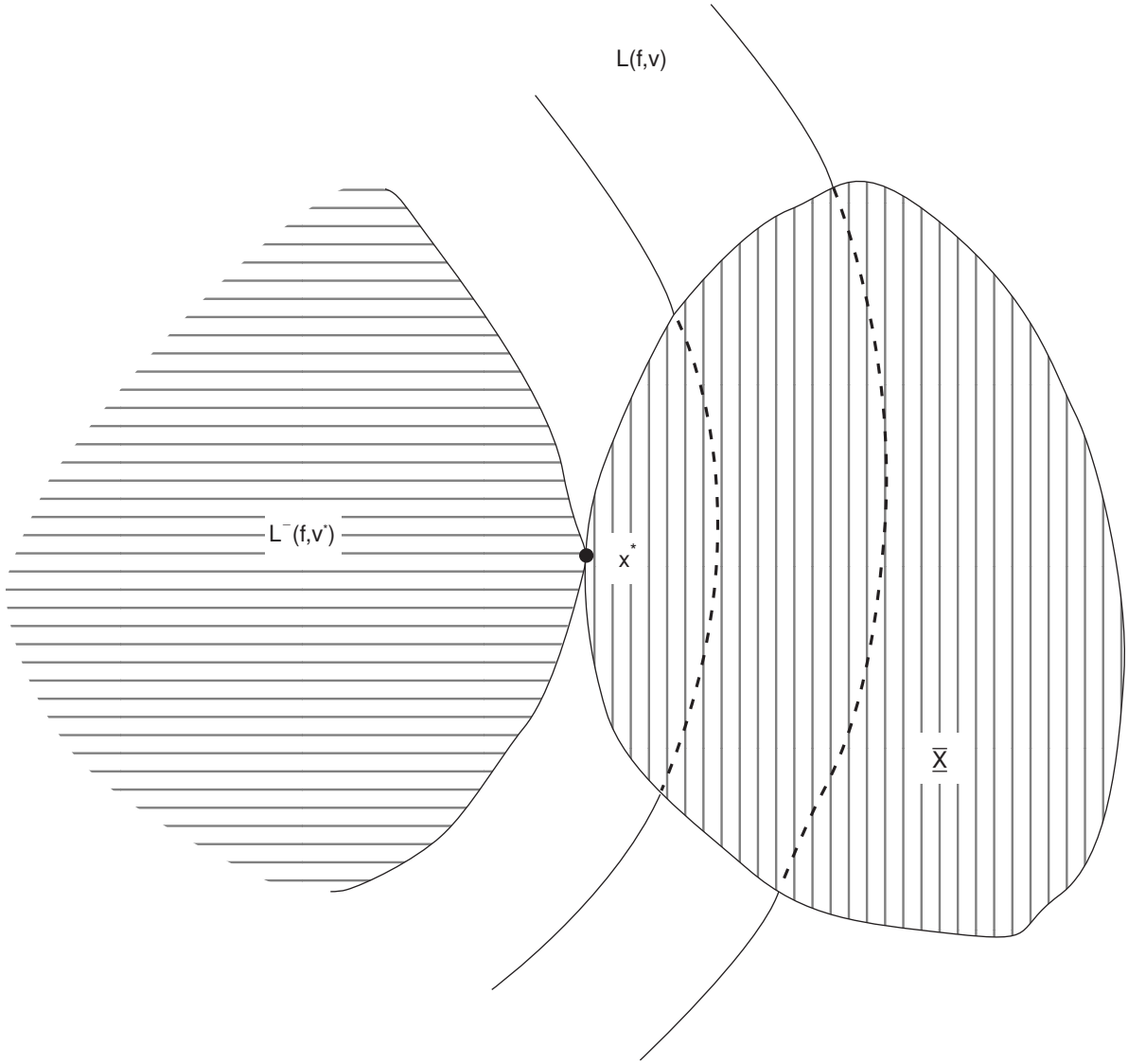


FIGURE 2 Global minimum for a convex program.

the level sets are all convex then, for any optimal solution x^l with optimal value v^l, the sets

$$L^-(f, v) \cap X$$

are empty for all $v < v^l$ and hence there are no local solutions that are not also global solutions. This situation is illustrated in Fig. 2.

These geometric conditions characterize the optimal solutions completely. However, they are not very useful in practice because the level sets cannot be easily mapped. In the next section, these geometric conditions are transformed into algebraic conditions, which will in turn yield qualitative and quantitative results for the nonlinear programs.

B. First-Order Conditions

In order to establish first-order optimally conditions the assumption that the objective and constraint functions are continuously differentiable will be used. This allows the functions to be approximated linearly about a given vector, which in turn leads to a local polyhedral approximation of the feasible set.

Given a continuously differentiable function w defined on $\Re^N$ and a vector $\hat{x}$ the affine (linear) approximation to w at $\hat{x}$ is

$$\hat{w}(x) = w(\hat{x}) + \nabla w(\hat{x})^t (x - \hat{x}),$$

where $\nabla w(\hat{x})$ is the *gradient* of w, that is, the N-dimensional column vector of partial derivatives of w at

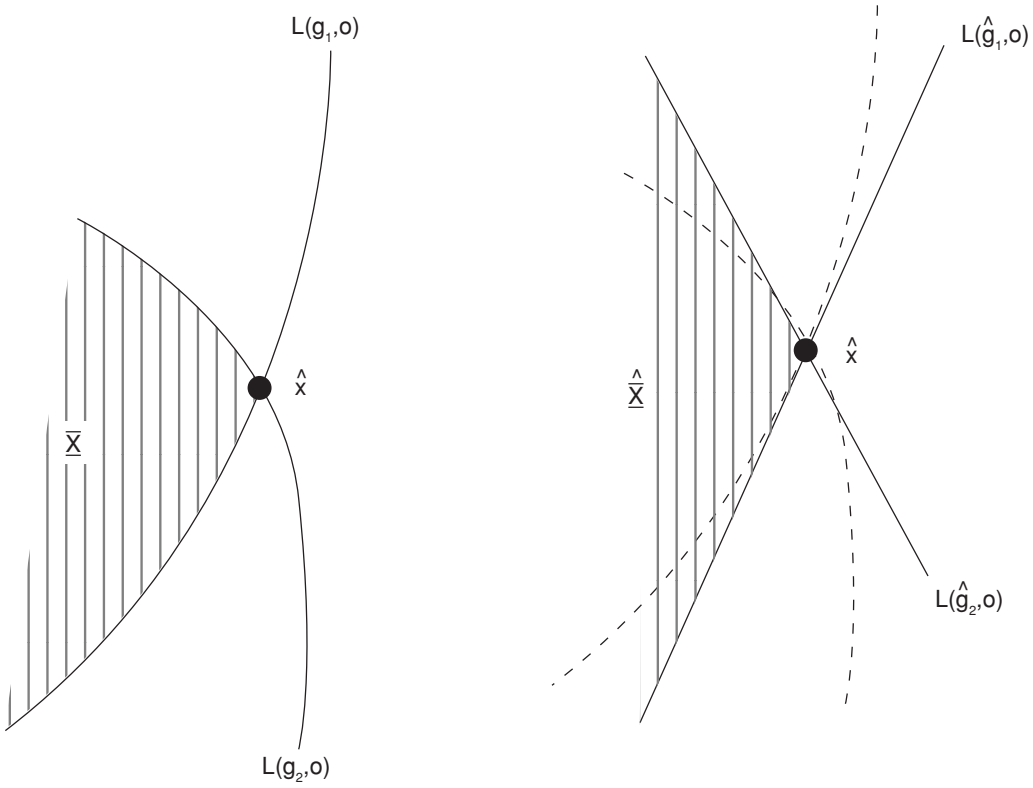


FIGURE 3 Linear approximation of the feasible set.

$\hat{x}$. If the objective and constraint functions in NLP are replaced by their affine approximations at a point $\hat{x} \in X$ then the resulting optimization problem is a linear program with a polyhedral feasible set $\hat{X}$, which serves as a local approximation of NLP (although the approximation may be poor). Figure 3 illustrates this approximation in a simple inequality-constrained case.

It can easily be shown that in the inequality-constrained case, $\hat{x}$ is an optimal solution to the approximating linear program if and only if the vector $-\nabla f(\hat{x})$ is a nonnegative combination of the gradients of the *active constraints* at $\hat{x}$ (i.e., those constraints that have value 0 at $\hat{x}$). As is seen in Fig. 4, this condition means that $\hat{f}$ cannot have a direction of decrease into the interior of $\hat{X}$ from $\hat{x}$ and hence $-\nabla f(\hat{x})$ must lie in the cone $\hat{K}$ generated by the gradients of the active inequality constraints. If equality constraints are present, the situation is slightly more complicated in that $-\nabla f(\hat{x})$ must be a combination of all of the active gradients with the coefficients of the active inequality constraint gradients being nonnegative and the coefficients of the equality constraints being unrestricted in sign; that is,

$$\nabla f(x) + \sum_{i=1}^{M} \mu_i \nabla g_i(x) + \sum_{j=1}^{P} \omega_j \nabla h_j(x) = 0$$

for some $\mu_i \geq 0$ and some ω_j.

The question that now arises is the extent to which the optimality of $\hat{x}$ for the approximating problem is related to the optimality of $\hat{x}$ for NLP. The answer is contained in the following fundamental result of nonlinear programming.

Theorem 1 (First-Order Necessary Conditions): Let $\hat{x} \in \Re^N$ and suppose that the set of active constraint gradients at $\hat{x}$,

$$\{\nabla g_i(\hat{x}) : i \text{ such that } g_i(\hat{x}) = 0\}$$
$$\cup \{\nabla h_j(\hat{x}),\ j = 1, \ldots, P\},$$

is linearly independent. Then $\hat{x}$ is a local optimal solution only if there exist vectors $\hat{\mu} \in \Re^M$ and $\hat{\omega} \in \Re^P$ such that

$$\nabla f(\hat{x}) + \sum_i \hat{\mu}_i \nabla g_i(\hat{x}) + \sum_j \hat{\omega}_j \nabla h_j(\hat{x}) = 0, \quad (1)$$

$$\hat{\mu}_i g_i(\hat{x}) = 0, \quad i = 1, \ldots, M, \quad (2)$$

$$\hat{\mu}_i \geq 0, \quad i = 1, \ldots, M, \quad (3)$$

$$g_i(\hat{x}) \leq 0, \quad i = 1, \ldots, M, \quad (4)$$

$$h_j(\hat{x}) = 0, \quad j = 1, \ldots, P. \quad (5)$$

The condition on the linear independence of the active constraint gradients is called a *constraint qualification*. It assures that the geometry at $\hat{x}$ is not so pathological

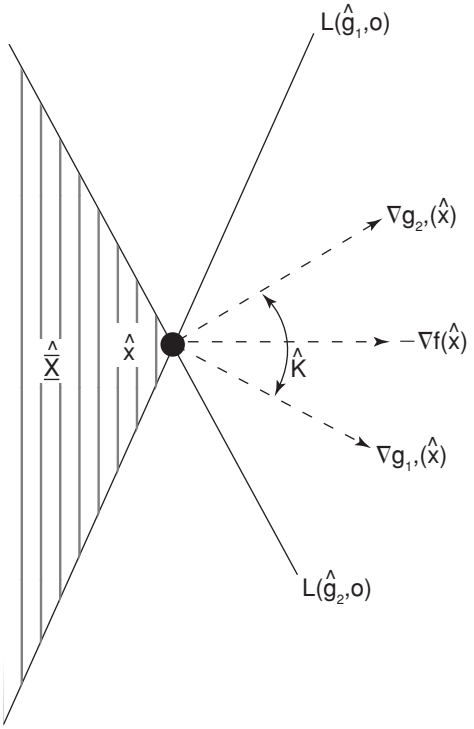


FIGURE 4 $-\nabla f(\hat{x})$ in the cone defined by the active constraint gradients.

that the set $\hat{X}$ is not a good local approximation of X. If this condition holds and $\hat{x}$ is a local optimal solution of NLP, then it is a global optimal solution to the linear approximating problem and hence conditions in Eqs. (1) and (3) hold. It is also a consequence of this constraint qualification that for a given $\hat{x}$ the corresponding $\hat{\mu}$ and $\hat{\omega}$ are unique. Conditions in Eq. (2) are called the *complementary slackness* conditions; they merely expresses the condition that $\hat{\mu}_i = 0$ if $g_i(\hat{x}) \neq 0$ [i.e., inactive constraints do not effect Eq. (1)]. Conditions in Eqs. (4) and (5) are just the feasibility conditions on $\hat{x}$.

Note that this characterization of $\hat{x}$ is a necessary condition only. It is easy to construct examples for which $\hat{x}$ satisfies Eqs. (1)–(5) but is not a local optimal solution for NLP. This is a consequence of using only linear approximations that are too coarse to reflect the true state of affairs. In order for these first-order conditions to always be sufficient for optimality as well as necessary, the problem must be convex.

A real-valued function w defined on $\Re^N$ is said to be *convex* if for every x and y in $\Re^N$ and every λ, $0 \leq \lambda \leq 1$, the following inequality holds:

$$w(\lambda x + (1-\lambda)y) \leq \lambda w(x) + (1-\lambda)w(y).$$

It can easily be established that if w is convex then the level set $L^-(w, \alpha)$ is a convex set for every α for which it is nonempty and that $L(w, \alpha)$ is convex if and only if w is an affine function. For these reasons, NLP is called a *convex program* if the functions f and g_i, $i = 1, \ldots, M$ are convex and the functions h_j, $j = 1, \ldots, P$ are affine. If NLP is convex, then X and $L^-(f, \alpha)$ are convex sets and, as indicated in the preceding section, every local optimal solution is a global optimal solution. The consequences of these facts are summarized in the following theorem.

Theorem 2: Let NLP be a convex program and assume that there is an $x^c \in X$ with $g_i(x^c) < 0$ for $i = 1, \ldots, M$. Then $\hat{x}$ is a global optimal solution of NLP if and only if there are vectors $\hat{\mu}$ and $\hat{\omega}$ such that Eqs. (1)–(5) hold.

The condition involving x^c is a standard constraint qualification for convex programs, sometimes called the Slater condition. Theorems 1 and 2 are important in that they severely limit the set of possible optimal points for NLP. These conditions are the higher-dimensional analogues of the conditions given in the example of Section I.B for locating a minimum of a nonlinear function of one variable on a closed interval. Most computational algorithms for solving NLP generate $\hat{x}$, $\hat{\mu}$, and $\hat{\omega}$ that approximately satisfy Eqs. (1)–(5).

To illustrate the application of the necessary conditions, consider the convex quadratic program

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}x^t A x + a^t x \\ \text{subject to:} \quad & Bx - b \leq 0, \\ & Dx - d = 0, \end{aligned}$$

where A is an $N \times N$ positive definite matrix, B is an $M \times N$ matrix, and D is a $P \times N$ matrix of rank P. This is a convex program, and assuming that the constraint qualification of Theorem 2 holds, the solution to this problem is the solution of the conditions

$$\begin{aligned} Ax + B^t\mu + D^t\omega + a &= 0, \\ \mu^t(Bx - b) &= 0, \\ \mu &\geq 0, \\ Bx - b &\leq 0, \\ Dx - d &= 0. \end{aligned}$$

If there are no inequality constraints then this system reduces to the $(P + N)$-dimensional linear system

$$\begin{aligned} Ax + D^t\omega &= -a, \\ Dx &= d, \end{aligned}$$

which has a unique solution; namely, (x^*, ω^*), the global optimal solution and its multiplier. Thus, in this case

solving the NLP reduces to solving a linear system of equations.

Historically, the necessary conditions in theorem 1 are a part of classical mathematics for the case when there are only equality constraints; the $\hat{\omega}_j$ are known as Lagrange multipliers and Eq. (1) is then a direct consequence of the implicit function theorem. The inequality-constrained case was treated by Karush, Kuhn, and Tucker and the theory is sometimes referred to by their names. It should be noted that a more general necessary condition can be derived without assuming any constraint qualification. This result, called the Fritz John condition, differs from that of theorem 1 in that there is a coefficient, possibly zero, in front of $\nabla f(x)$ in Eq. (1).

C. Second-Order Conditions

In this section, it is assumed that the objective and constraint functions are twice continuously differentiable. For unconstrained nonlinear functions, second-order conditions are easily derived. Suppose $\hat{x}$ is a critical point of f (i.e., $\nabla f(\hat{x}) = 0$) and let $Hf(\hat{x})$ denote the Hessian matrix of f at $\hat{x}$, the symmetric matrix of second-order partial derivatives. Then $\hat{x}$ is an unconstrained local minimum of f if $Hf(\hat{x})$ is positive definite and only if $Hf(\hat{x})$ is positive semidefinite.

To state analogous conditions for the constrained problem, the Lagrangian function is introduced. This function of the $N + M + P$ variables (x, μ, ω) defined by

$$\mathcal{L}(x, \mu, \omega) = f(x) + \sum_i \mu_i g_i(x) + \sum_j \omega_j h_j(x)$$

is central to the theoretical development of the subject. Conditions in Eqs. (1)–(5) show that $(\hat{x}, \hat{\mu}, \hat{\omega})$ is a critical point of $\mathcal{L}(x, \mu, \omega)$ with respect to x and ω and satisfies

$$\frac{\partial \mathcal{L}}{\partial \mu_i}(\hat{x}, \hat{\mu}, \hat{\omega}) = g_i(\hat{x}) \leq 0.$$

The second-order conditions for NLP are given in theorems 3 and 4. In these theorems, $I(x)$ will represent the index set of active inequality constraints at x; that is,

$$I(x) = \{i : g_i(x) = 0\}.$$

Theorem 3 (Second-Order Necessary Conditions): Let $\hat{x}$ satisfy the conditions of Theorem 1 with corresponding multipliers $\hat{\mu}$ and $\hat{\omega}$. If $\hat{x}$ is a local minimum then for any vector z satisfying

$$\nabla g_i(\hat{x})^t z = 0, \quad i \in I(\hat{x}),$$
$$\nabla h_j(\hat{x})^t z = 0, \quad j = 1, \ldots, P,$$

it is the case that

$$z^t H_{xx} \mathcal{L}(\hat{x}, \hat{\mu}, \hat{\omega}) z \geq 0$$

where $H_{xx}\mathcal{L}(\hat{x}, \hat{\mu}, \hat{\omega})$ is the $N \times N$ Hessian matrix of $\mathcal{L}$ with respect to x.

The second-order necessary condition requires the Hessian of the Lagrangian to be positive semidefinite on the tangent subspace to the feasible set at $\hat{x}$. In a manner analogous to that of the unconstrained case, a second-order sufficiency condition can be obtained by the more stringent requirement that the Hessian of $\mathcal{L}$ be positive definite on this subspace.

Theorem 4 (Second-Order Sufficiency Conditions): Let $\hat{x}$, $\hat{\mu}$, and $\hat{\omega}$ satisfy conditions in Eqs. (1)–(5) of theorem 1 and suppose that for $i \in I(\hat{x})$, $\hat{\mu}_i > 0$. Further, suppose that for every nonzero z satisfying

$$\nabla g_i(\hat{x})^t z = 0, \quad i \in I(\hat{x}),$$
$$\nabla h_j(\hat{x})^t z = 0, \quad j = 1, \ldots, P,$$

it is the case that

$$z^t H_{xx} \mathcal{L}(\hat{x}, \hat{\mu}, \hat{\omega}) z > 0.$$

Then $\hat{z}$ is an isolated local minimum of NLP.

$\hat{x}$ is an isolated local minimum if there is a neighborhood of $\hat{x}$ that contains no other local solution of NLP. There are other slightly weaker versions of the second-order sufficient conditions that do not require that $i \in I(\hat{x})$ implies $\hat{\mu}_i > 0$. However, this restriction, called strict complementary slackness, is required in many important theoretical applications, including those of the next section. For twice differentiable functions it is the case that they are convex if and only if their Hessian matrices are positive semidefinite. Thus for convex programs, the Hessian of the Lagrangian is positive semidefinite and the second-order conditions are redundant, as is to be expected in light of theorem 2.

D. Stability and Duality

As stated earlier, the stability of the optimal solution and its optimal value are of major importance in applications of nonlinear programming models. Fortunately, the optimal solution and its optimal value are stable under reasonable conditions. To formally state this result, it is necessary to consider the perturbed version of NLP,

$$\begin{array}{ll} \text{minimize} & \hat{f}(x, p) \\ \text{subject to:} & \hat{g}_i(x, p) \leq 0, \quad i = 1, \ldots, M, \\ & \hat{h}_j(x, p) = 0, \quad j = 1, \ldots P, \end{array}$$

where p is a Q-vector, $\hat{f}(x, 0) = f(x)$, $\hat{g}_i(x, 0) = g_i(x)$, and $\hat{h}_j(x, 0) = h_j(x)$. Furthermore, it is assumed that

$\hat{f}$, $\hat{g}_i$, and $\hat{h}_j$ are continuously differentiable functions of p near $p=0$. The vector p represents the parameters of the problem. The following theorem gives conditions under which the optimal solution and multipliers are smooth functions of the perturbation.

To state the theorem, we define a *regular solution* of NLP to be a solution at which the linear independence of the active constraint gradients, strict complementary slackness, and the second-order sufficient conditions a11 hold.

Theorem 5 (Basic Perturbation Theorem): Let x^* be a regular optimal solution to NLP with multipliers μ^* and ω^*. Let $I(x^*)$ be the index set of active constraints at x^*. Then there is an $\epsilon > 0$ and functions $x^*(p)$, $\mu^*(p)$, $\omega^*(p)$ defined and continuously differentiable on the set $E = \{p : |p| \leq \epsilon\}$ with $x^*(0) = x^*$, $\mu^*(0) = \mu^*$, and $\omega^*(0) = \omega^*$. $x^*(p)$ is a local optimal solution to the corresponding perturbed problem and $\mu^*(p)$ and $\omega^*(p)$ are its multipliers. Moreover, the second-order sufficient conditions hold at $x^*(p)$ and $I(x^*(p)) = I(x^*)$.

One of the most important cases of perturbation occurs when p is an M-vector and $\hat{g_i}(x, p) = g_i(x) - p_i$. Under the assumptions of theorem 5, the optimal solution as a function of p, $x^*(p)$, is smooth and it can be shown that the optimal value as a function of p, $v^*(p) = f(x^*(p))$, satisfies.

$$\nabla_p v^*(0) = -\mu^*.$$

In other words, the instantaneous rate of change in the optimal value as a function of a shift in the value of g_i is the negative of the ith multiplier. This gives an interpretation of the multiplier in terms of the model. For example, if the ith constraint is a bound on a resource and the objective is measured in dollars, the value of additional units of that resource in terms of decreased optimal value is linearly approximated as μ_i^* dollars per unit. A similar result holds for pertubations of the equality constraints. As a result of this interpretation, the optimal multipliers are often called *shadow prices*.

The preceding observations on the properties of the multipliers lead, in linear programming, to the formulation of a *dual* linear program with the multipliers as the optimal solution. Similar, but much less complete results can be obtained for nonlinear programs. The appropriate formulation involves the Lagrangian function defined earlier. It can be shown that NLP is equivalent to the min-max problem

$$\min_x \max_{\mu \geq 0, \omega} \mathcal{L}(x, \mu, \omega).$$

The *dual problem* can then be defined as the max-min problem

$$\max_{\mu \geq 0, \omega} \min_x \mathcal{L}(x, \mu, \omega).$$

If the program is convex and the Slater condition holds then x^* is an optimal solution to NLP with multipliers μ^* and ω^* if and only if (x^*, μ^*, ω^*) is a saddle point for $\mathcal{L}(x, \mu, \omega)$ that is,

$$\mathcal{L}(x^*, \mu, \omega) \leq \mathcal{L}(x^*, \mu^*, \omega^*) \leq \mathcal{L}(x, \mu^*, \omega^*)$$

for all x, $\mu \geq 0$, and ω. Thus the values of the primal and dual problems are equal at optimality. This result fails to hold under less restrictive hypotheses on the problem NLP.

III. COMPUTATION

A. Basic Concepts

Finding a numerical approximation to the solution of a nonlinear program can be a difficult task. If the number of variables is large or the functions involved are highly nonlinear, the computation can be time consuming, even on the fastest computers. Moreover, in the case of nonconvex programs, the presence of local solutions can make the determination of a global solution problematic. The methods and procedures described as follows are concerned with approximating local solutions only.

The algorithms for approximating the solutions of NLP are all iterative in nature; that is, given an initial estimate of the solution, x^0, a sequence $\{x^k\}$ of approximate solutions is generated with each iterate, x^k, being determined successively from information gathered at the preceding iterations. It is desired that at each iteration, the new iterate is a better approximation, in some sense, of a local solution than the previous ones. In theory, the iterations should converge to a local solution, say x^*. In most algorithms approximations to the optimal multipliers, (μ^k, ω^k), are computed along with x^k at each iteration and the algorithms terminate when the approximate solutions and multipliers satisfy the first-order necessary conditions to some predetermined degree of accuracy.

In constrained optimization, in addition to minimizing f, the optimal solution must also satisfy the feasibility requirements. It is possible (even probable when there are nonlinear equality constraints) that an iterate will not be feasible. Ideally, an algorithm should be designed in such a way that, given a current iterate x^k, the new iterate x^{k+1} will be no more infeasible than x^k and will also satisfy $f(x^{k+1}) \leq f(x^k)$. Unfortunately, this is not always possible and therefore, a successful algorithm should balance the goals of feasibility and decreasing f. In the algorithms described in this article, the iterates are of the form

$$x^{k+1} = x^k + \alpha_k d^k,$$

where the vector d^k gives the direction in which a "step" is taken and and the scalar α_k, called the *step length parameter*, determines how far in this direction the next iterate lies. Another approach, called a *trust region method* can also be employed to determine the step from x^k to x^{k+1}; a reference to these methods can be found in the Bibliography.

In judging the effectiveness of an algorithm, there are three major criteria: (1) robustness, (2) rate of convergence, and (3) efficiency. The first two terms deal with theoretical issues. The term "robustness" refers to the likelihood that the algorithm will yield a sequence of iterates that converge, in theory, to a local solution regardless of the starting point x^0, whereas "rate of convergence" refers to the speed with which the iterates converge, for example, how fast the theoretical error terms, $|x^k - x^*|$, tend to zero. Finally, the "effectiveness" is concerned with practical considerations in implementing the algorithm; it includes such problems as computer time and storage requirements and numerical stability. The descriptions that follow discuss only the first two of these criteria because the third is very dependent on the particular platform and software used.

B. Unconstrained Optimization Algorithms

Any description of algorithms for solving NLP must begin with the computational algorithms for unconstrained optimization problems because the latter are fundamental to the design of the former. The algorithms for unconstrained problems are much less complex because there is no question of feasibility that must be taken into account in the choice of step direction and length. There are two basic approaches to solving the unconstrained problem to be discussed here: the basic descent method and variants of Newton's method. The former emphasizes decreasing f while the latter attempts to solve the necessary conditions, $\nabla f(x) = 0$. Each is discussed briefly as are hybridizations of the two approaches that attempt to incorporate their best properties.

For the descent method, the choice of the step direction is a direction d^k satisfying

$$\nabla f(x^k)^t d^k < 0. \tag{6}$$

Movement in this direction, called a *descent direction* will, for at least for a short distance, decrease f, and hence α_k can be chosen so that

$$f(x^{k+1}) = f\left(x^k + \alpha_k d^k\right) < f(x^k).$$

The specific choice $d^k = -\nabla f(x^k)$, gives the direction of greatest local rate of decrease in f and leads to the *steepest descent* method. In practice, α_k is chosen so as to minimize a polynomial approximation to $\phi(\alpha) = f(x^k + \alpha d^k)$ subject to the need to reduce the value of f. Under fairly reasonable conditions, these descent methods yield a convergent sequence of iterates and the limit point satisfies the equation $\nabla f(x^*) = 0$. These methods are fairly robust; their convergence not depending on the initial iterate x^0. However, the convergence rate of these descent methods is typically very slow and, consequently, they are ill-suited for general use.

The second general method is that of Newton's method and its modifications. The pure Newton method determines the step, d^k, for solving $\nabla f(x) = 0$ by

$$d^k = -[Hf(x^k)]^{-1} \nabla f(x^k),$$

where $Hf(x^k)$ is the Hessian matrix of f at x^k, and requires $\alpha_k = 1$. It is well known that if the initial iterate x^0 is sufficiently close to a local solution x^* at which the gradient $\nabla f(x^k)$ is nonzero, the iterates will converge to x^* very rapidly. In particular, they will converge at a quadratic rate, which roughly means that the number of decimal places of accuracy is doubled at each step. The biggest drawback to the use of Newton's method is its lack of robustness. Unless the vector x^0 is close to a local solution, these iterates may fail to converge or else converge to a nonminimum point.

There are a number of methods designed to obtain both the robustness of descent methods and the rapid local convergence of Newton's method. If f is a convex function, its Hessian matrix is positive definite and the Newton step is then also a descent step. Applied with a line search to determine α_k, this approach, called a *damped Newton method*, can both be robust and yield rapid local convergence. For nonconvex problems a standard approach for obtaining robustness without sacrificing rapid convergence is to use a *quasi-Newton* or *secant* method. In this type of algorithm, the Hessian matrix, $Hf(x^k)$ is approximated by a *positive definite* matrix, B_k, and the step direction is given by

$$d^k = -[B_k]^{-1} \nabla f(x^k).$$

Because B_k is positive definite, the direction d_k is a descent direction [satisfies Eq. (6)] and a line search procedure can be used to determine α_k. Because it is a descent method, this algorithm will be robust and if the B_k are relatively good approximations to the Hessians of f, the convergence rate should be close to that of Newton's method. The class of matrices that satisfy the generalized secant condition

$$B_{k+1}(x^{k+1} - x^k) = \nabla f(x^{k+1}) - \nabla f(x^k) \tag{7}$$

provide good approximations to the Hessian of f at x^{k+1}. These approximations are implemented in an algorithm by an "updating formula" of the form

$$B_{k+1} = B_k + E_k,$$

where E_k is a low rank matrix chosen so that the matrix B_{k+1} is positive definite and Eq. (7) holds. It has been shown, under relatively mild restrictions, that the descent method using appropriate secant updates will be robust and converge at a rate approaching that of Newton's method. Most modern unconstrained optimization algorithms now utilize a version of this procedure. The references on unconstrained optimization contain details for these methods.

C. Penalty and Barrier Functions

An early approach to solving NLP was to attempt to convert the constrained problem into a constrained one by incorporating the constraints into the objective function. There are basically two types of methods included in this approach: penalty methods and barrier methods. The two cases are considered separately.

Penalty methods, as the name suggests, incorporate the constraints into the objective function in such a way that infeasibility is penalized. The most common method replaces NLP by the problem of minimizing

$$P(x; \rho) = f(x) + \rho(\|h(x)\| + \|g^+(x)\|) \qquad (8)$$

where

$$g_i^+(x) = \begin{cases} 0 & \text{if } g_i(x) \le 0, \\ g_i(x) & \text{else,} \end{cases}$$

and $\|\cdot\|$ represents any norm. $\rho > 0$ is a parameter that plays a major role in the algorithm. Given a value of the parameter ρ, the minimization of P will take into consideration values of x that are infeasible but will penalize them proportionally to their "distance" from feasibility, as measured by $h(x)$ and $g^+(x)$, and to the value of ρ. If a sequence $\rho_k \to \infty$ is chosen and $x(\rho_k)$ is an unconstrained minimum of $P(x; \rho_k)$ with $x(\rho_k) \to \hat{\mathbf{x}}$, it can be shown that $\hat{\mathbf{x}}$ is a local minimum of NLP. This observation forms the basis for an algorithm for solving NLP in which x^k is taken as an approximation to $x(\rho_k)$ and is used as the starting point for minimizing $P(x, \rho_{k+1})$.

The choice of norm to be used in the definition of P is important. From an ease of computation standpoint, the Euclidean norm is the most convenient. However, the L_1 norm,

$$\|z\| = \sum_i |z_i|,$$

provides an important theoretical advantage. Although the use of the L_1 norm forces $P(x; \rho)$ to be nondifferentiable on the boundary of the feasible region, it can be shown that the following property holds.

Theorem 6: There exists a $\rho^* > 0$ such that for all $\rho > \rho^*$, an unconstrained minimum of the L_1 penalty function is an optimal solution for NLP, and conversely.

A penalty function that has the property described in this theorem is called an *exact* penalty function. This type of penalty function is important because it obviates the use of limiting processes to obtain a solution to NLP. This is not the panacea that it appears, however, because exact penalty functions are typically so complicated that they do not readily lend themselves to computational procedures; their application is less efficient than applying other techniques for solving NLP. For example, the nondifferentiability of the L_1 penalty function precludes using the gradient descent methods described above. However, the L_1 penalty function does have a use as a "merit function" in other optimization algorithms for solving NLP; as is described as follows.

Barrier methods are applied primarily to inequality-constrained problems. In particular, they are applicable when the feasible region is known to have a strictly feasible point x^0, that is, a point that satisfies $g(x^0) < 0$. The most common barrier function is the so-called *logarithmic barrier function*,

$$B(x; \rho) = f(x) - \rho \sum_{j=1}^{p} \log(-g_j(x)), \qquad (9)$$

where ρ is again a positive parameter. Another type of barrier function is the inverse barrier function,

$$V(x; \rho) = f(x) + \rho \sum_{j=1}^{p} 1/(-g_j(x)).$$

An optimization procedure for minimizing $B(x, \rho)$ for fixed positive ρ started at x^0 cannot reach the boundary of X because the log term approaches $+\infty$ as the boundary is neared; that is, the boundary defines a barrier that the optimization procedure cannot cross. As ρ gets smaller, the penalty for approaching the boundary gets smaller, so if $x(\rho k)$ represents any set of minima of $B(x; \rho_k)$ corresponding to a sequence $\rho_k \to 0$ then a limit point $\hat{\mathbf{x}}$ must be an optimal solution of NLP. There is no exact barrier function because, if the solution to NLP is on the boundary of X, it can only be reached in the limit. Nevertheless, barrier functions have also found applications to solving NLP in the guise of the interior point methods that are described as follows.

D. The Sequential Quadratic Programming Algorithm

At one time, it was common to generate approximations to the solution of NLP by linearly approximating the objective function and the constraints at a current iterate

and to use this information to generate the next iterate. These methods, as typified by the gradient projection and reduced gradient algorithms, are comparable to the descent method described for unconstrained optimization and hence are not competitive for the more complicated nonlinear constrained problems.

A more general approach that uses higher order information and has been shown to be particularly effective is the *sequential quadratic programming* algorithm. In this type of algorithm an approximation to NLP is constructed in which the constraints are approximated linearly and a quadratic approximation to the Lagrangian function is employed as an objective function. Specifically, if x^k is a current iterate for NLP, not necessarily feasible, with corresponding multiplier approximations μ^k and ω^k the following quadratic programming problem results:

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2} d^t B_k d + \nabla_x \mathcal{L}(x^k, \mu^k, \omega^k)^t d \\ \text{subject to:} \quad & \nabla g_i(x^k)^t d + g_i(x^k) \le 0, \quad i = 1, \ldots, M, \\ & \nabla h_j(x^k)^t d + h_j(x^k) = 0, \quad j = 1, \ldots, P, \end{aligned}$$

where B_k is the Hessian to the Lagrangian with respect to x at (x^k, μ^k, ω^k), $H_{xx}\mathcal{L}(x^k, \mu^k, \omega^k)$, or an approximation thereof. If this problem is solved to obtain a solution d^k with associated multipliers y^k and z^k, then the new iterate for NLP is given by

$$x^{k+1} = x^k + \alpha_k d^k,$$

for an appropriate choice of step length α_k. Updates for the multipliers are given by

$$\mu^{k+1} = y^k \quad \text{and} \quad \omega^{k+1} = z^k.$$

The use of the Lagrangian in the objective function permits quadratic effects in the constraints to have an effect on the iteration.

As a justification for this approach, it can be seen that if the true Hessian of the Lagrangian is used in a problem with equality constraints only, then the iterate (x^{k+1}, ω^{k+1}) is identical to that obtained by using the Newton step to find a solution of the first-order conditions of NLP:

$$\begin{aligned} \nabla f(x) + \sum_{j=1}^{P} h_j(x)\omega_j &= 0 \\ h(x) &= 0, \end{aligned}$$

starting at (x^k, ω^k). A similar result holds for the case when inequality constraints are present. Thus, depending on the choice of the step length parameter, this method will yield convergence at a rate similar to that of Newton's method when started near a solution.

In the actual implementations of the algorithm, the matrices $\{B_k\}$ are often a quasi-Newton approximations just as in the algorithms for unconstrained optimization; that is, they satisfy the conditions:

$$\begin{aligned} B_{k+1}(x^{k+1} - x^k) = \nabla_x \mathcal{L}(x^{k+1}, \mu^{k+1}, \omega^{k+1}) \\ - \nabla_x \mathcal{L}(x^k, \mu^{k+1}, \omega^{k+1}). \end{aligned}$$

These matrices are updated by a rank-two matrix at each iteration and are usually chosen to be positive definite. Since the true Hessian is not positive definite except on a subspace of $\Re^N$ (see theorem 4), the use of positive definite quasi-Newton approximations will usually lead to a slower rate of convergence. Another aspect of this algorithm that requires careful implementation is the determination of the step length parameter α_k. In unconstrained optimization, the parameter is chosen so that the objective function is decreased at each step. As was described in the first section of this article, in constrained optimization there is also the requirement of achieving feasibility, or at least decreasing infeasibility. At any step this may be inconsistent with decreasing the objective function and therefore it is not clear how the choice of α_k should be made. This is where the *merit function*, mentioned in the preceding section, comes into play. In the sequential quadratic programming method, a decrease in a merit function is used to determine the step length parameter for a given iterate. The choice of merit function depends upon the particular version of the algorithm but is usually taken to be one of the standard penalty functions. For example, if the L_1 penalty function is used then at a given step, then α_k is chosen so that

$$P\left(x^k + \alpha_k d^k; \rho\right) < P(x^k, \rho).$$

Because P is decreased at each step and for large ρ a minimum of P is a minimum of NLP, this choice of α_k is justified. However, the value of ρ must generally be adjusted in the algorithm to assure that this property holds.

In spite of the complications involved in its implementation, versions of the sequential quadratic programming algorithm are considered the most effective general purpose algorithms currently available for solving the NLP.

E. Interior Point Methods

Recently, there have been developments in algorithms for NLP that involve generalizing the interior point algorithms that have been so successful in linear programming. In one approach to interior point methods, the inequality constraints are incorporated into the objective function by the use of a barrier function. A similar formulation results from considering the Newton steps for solving a perturbed

form of the first order conditions. A basic version of the latter development for the case where there are only inequality constraints is given here; more examples of the method can be found in the Bibliography.

The first order system to be solved in this method is the system of equations derived from Eqs. (1)–(5), in which a slack variable z has been added to covert the inequalities to equalities:

$$\begin{aligned} \nabla f(x) + \sum_{i=1}^{M} g_i(x)\mu_i &= 0, \\ g(x) + z &= 0, \\ g_i(x) \cdot z_i &= \beta, \quad i = 1, \ldots, M, \\ z &\geq 0, \\ \mu &\geq 0. \end{aligned}$$

This system is perturbed because the complementary slackness conditions are not satisfied at the zero level but at the β level for some $\beta > 0$. As in barrier function methods, it is assumed that there is a family of solutions $(x(\beta), z(\beta), \mu(\beta))$ to this set of equations. Under certain assumptions (usually convexity of the f and g_i functions), this set of solutions defines a curve called the *central path* that tends to a solution of NLP as $\beta \to 0$. The idea here is to use Newton steps to solve this system while decreasing β towards zero. Given a current iterate, (x^k, z^k, μ^k) with $z^k > 0$ and $\mu^k > 0$, the set of equations becomes

$$\begin{bmatrix} H_{xx}\mathcal{L}(x^k, \mu^k) & Jg(x^k)^t & 0 \\ Jg(x^k) & 0 & I \\ 0 & Z_k & U_k \end{bmatrix} \begin{bmatrix} dx \\ du \\ dz \end{bmatrix} = \begin{bmatrix} -\nabla_x \mathcal{L}(x^k, \mu^k) \\ -g(x^k) - z^k \\ \beta e - Z_k \mu^k \end{bmatrix},$$

where $Jg(x^k)$ is the Jacobian of the vector function g, Z_k and U_k are the diagonal matrices with diagonals that are z^k and μ^k, e is the vector of ones, and $\mathcal{L}$ is the Lagrangian function. The iterates are now updated by the formulas:

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k\, dx, \\ z^{k+1} &= z^k + \alpha_k\, dz, \\ \mu^{k+1} &= \mu^k + \alpha_k\, du, \end{aligned}$$

where α_k is chosen to maintain the slack variables and multiplier variables as positive and to decrease an appropriate merit function. The key issue in an implementation of this algorithm is how progress toward the point on the central path corresponding to β is mixed with decrease of β. As in the the case of sequential quadratic programming, an approximation of the Hessian of the Lagrangian can be used.

IV. APPLICATIONS

A. Nonlinear Models

The description of actual nonlinear optimization models is hindered by the fact that the nonlinearities that occur in the model are usually due to technical theoretical concepts special to the particular field and hence are not easily explained to the nonspecialist. In other cases the objective or constraint functions may have little or no theoretical basis but be nonlinear functions that have been fit to historical data. Linear optimization models in production, blending, and network problems are discussed under the topic of linear programming. Many nonlinear programming models are generalizations of these linear problems in which a more accurate simulation of reality is obtained by allowing nonlinearities in the objective and constraint functions.

In models for which the objective or constraint function is a cost function, nonlinearities very often occur when economies of scale are present (i.e., when the per unit cost of a particular item decreases as the amount purchased or produced increases). In blending problems, a particular quality of the blend, for instance, the octane rating of gasoline, is a nonlinear function of the components of the mixture. In network flow problems, nonlinearities may represent increasing per unit cost of shipment as a function of the flow from node to node. Another type of nonlinearity results from the modeling of a learning process in production models; that is, the efficiency of production increases as production rises.

B. Statistical Applications

One of the most important uses of nonlinear programming is in the estimation of the parameters for a statistical distribution using the observed data. This is called maximum likelihood estimation. A related problem is the familiar regression problem in which parameters, say $\theta_i, i = 1, \ldots, N$, are estimated so that a particular nonlinear function $w(\theta)$ "best" fits a set of observed data; that is, given values of a control variable $z^k, k = 1, \ldots, L$, and a sequence of corresponding experimental responses $y^k, k = 1, \ldots, L$, values of θ are determined so that some measure of the difference vector $e \in \Re^L$, with

$$e_k = y^k - w(\theta, z^k), \quad k = 1, \ldots, L,$$

is minimized.

The most common measure is the Eucclidean norm of e, in which case, the optimization problem is

$$\begin{aligned} \sup \quad & \sum_{k=1}^{L} [y^k - w(\theta, z^k)]^2 \\ \text{subject to:} \quad & \theta \in \Theta, \end{aligned}$$

which is the so-called least squares pregression problem. Here Θ is the feasible set, which expresses any relations or restriction on the parameters. This problem, expecially for the case in which $\Theta = \Re^L$, has been exhaustively studied and many software packages are now available containing algorithms that compute a solution in an efficient manner.

C. Discretization of Infinite-Dimensional Problems

The increase in available computer power has allowed the use of finite-dimensional nonlinear programming techniques to be applied to solve discretized infinite-dimensional optimization problems. Problems generated in this manner are typically of very large dimension and often have special structure. The limiting behavior of the optimal solutions as the number of variables increases is of significance in this type of problem.

As an illustration of this problem, the simple optimal control problem of minimizing an energy function for a prescribed trajectory is formulated as follows.

$$\begin{aligned} \text{minimize} \quad & \int_a^b E(u(t))\,dt \\ \text{subject to:} \quad & dy(t)/dt = w(y(t), u(t)), \\ & y(a) = y_a, \\ & y(b) = y_b, \\ & |u(t)| \le 1, \end{aligned}$$

where a, b, y_a, and y_b are given scalars, and E and w are given real-valued functions. The object is to choose the control function $u(t)$ that solves the problem. To discretize this problem, $[a, b]$ is first partitioned so that

$$a = t_0 < t_1 < t_2 < \cdots < t_N = b.$$

Then, the following finite-dimensional nonlinear program approximates the optimal control problem.

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{N} E(u_j)(t_j - t_{j-1}) \\ \text{subject to:} \quad & y(t_j) - y(t_{j-1}) = (t_j - t_{j-1}) w(y(t_{j-1}), u_j), \\ & \qquad\qquad\qquad\qquad\qquad\qquad j = 1, \ldots, N, \\ & y(t_0) = y_a, \\ & y(t_N) = y_b, \\ & |u_j| \le 1, \quad j = 1, \ldots, N. \end{aligned}$$

The control $u(t)$ is approximated in this model by the piecewise constant function

$$\hat{u}(t) = u_j, \quad \text{for } t_{j-1} \le t < t_j.$$

In theory the approximation should improve as N gets large. Unfortunately, the number of variables and constraints in the nonlinear program increases and the problem becomes more difficult to solve. Problems of this type with tens of thousands of variables are not uncommon.

SEE ALSO THE FOLLOWING ARTICLES

APPROXIMATIONS AND EXPANSIONS • DYNAMIC PROGRAMMING • LINEAR OPTIMIZATION • MATHEMATICAL MODELING

BIBLIOGRAPHY

Avriel, M. (1976). "Nonlinear Programming, Analysis and Methods," Prentice-Hall, Englewood Cliffs, New Jersey.

Biegler, L., Nocedal, J., and Schmid, C. (1995). "A reduced Hessian method for large-scale constrained optimization," *SIAM J. Optimization* **5,** 314–347.

Boggs, P., Kearsely, A., and Tolle, J. (1999). "A global convergence analysis of an algorithm for large scale nonlinear programming problems," *SIAM J. Optimization* **9,** 833–862.

Boggs, P., and Tolle, J. (1995). "Sequential quadratic programming," *Acta Numerica*. 1–51.

Dennis, J., and Schnabel, R. (1983). "Numerical Methods for Nonlinear Equations and Unconstrained Optimization," Prentice-Hall, Englewood Cliffs, New Jersey.

El-Alem, M. (1995). "A robust trust-region algorithm with a nonmonotonic penalty parameter scheme for constrained optimization," *SIAM J. Optimization* **5,** 348–378.

El-Bakry, A., Tapia, R., Tsuchiya, T., and Zhang, Y. (1996). "On the formulation and theory of the Newton interior point method for nonlinear programming," *J. Optimization Theory Applications* **89,** 507–541.

Nash, S., and Sofer, A. (1996). "Linear and Nonlinear Programming," McGraw-Hill, New York.

Nocedal, J. (1991). "Theory of algorithms for unconstrained optimization," *Acta Numerica*. 199–242.



Number Theory, Algebraic and Analytic

H. E. Rose
University of Bristol

I. Algebraic Number Fields
II. Diophantine Equations
III. Elliptic Curves
IV. Diophantine Approximation and Transcendence
V. Prime Number Theory
VI. Partitions
VII. Computational Number Theory

GLOSSARY

Congruence If $a, b, m \in \mathbb{Z}$ and $m > 0$ then $a \equiv b \pmod{m}$ stands for m divides $b - a$; written as "a is congruent to b modulo m."

Divisibility If $a, b \in \mathbb{Z}$, we write $b \mid a$ when b divides a exactly.

Equivalence relation A relation $\simeq$ on a set S that satisfies, for all $a, b, c \in S$, $a \simeq a$, if $a \simeq b$ then $b \simeq a$, and if $a \simeq b$ and $b \simeq c$ then $a \simeq c$. The congruence defined above is an example. The subset $\{x : x \simeq a\}$ of the set S, where a is a fixed member of S, is called an *equivalence class*. The relation $\simeq$ splits S into a disjoint union of such classes.

Greatest common divisor If $a, b \in \mathbb{Z}$, not both zero, then (a, b) denotes the largest integer that divides both a and b. It exists, as $\mathbb{Z}$ has unique factorization.

Integer Member of the set $\{\ldots -2, -1, 0, 1, 2, 3, \ldots\}$. To distinguish from other sets defined later we often use the term *rational integer* instead of integer. $\mathbb{Z}$ denotes the set of all integers with the usual operations and rules of addition, subtraction, and multiplication. It is an integral domain.

Order $\bigcirc, \sim$**:** $f(x) = \bigcirc(g(x))$, f has order g, stands for $\lim_{x\to\infty} |f(x)|/g(x)$ is bounded as $x \to \infty$, and $f(x) \sim g(x)$ stands for $\lim_{x\to\infty} f(x)/g(x) = 1$.

Polynomial An expression of the form $a_0x^n + a_1x^{n-1} + \cdots + a_n$ where each a_i is a constant in the underlying ring or field and x is a variable; n is called the *degree* of the polynomial. Polynomials in two or more variables are defined similarly. A polynomial is called *irreducible* if it cannot be written as a product of polynomials of smaller degree. Finally a polynomial is called *homogeneous* if each of its summands has the same degree; for example, $x^4 + 3xy^3 - xyzt$ is homogeneous of degree 4.

Prime number p is a prime if $p \in \mathbb{Z}$, $p > 1$, and if $a > 0$ and a divides p then $a = 1$ or $a = p$.

Rational function A real function of the form F/G where F and G are polynomials.

Rational number A number of the form a/b where $a, b \in \mathbb{Z}$, $(a, b) = 1$, and $b \neq 0$. $\mathbb{Q}$ denotes the set of all rational numbers with the usual operations and rules of addition, subtraction, multiplication, and division. It is a field.

Unique factorization Every element in $\mathbb{Z}$ can be represented uniquely as a product of primes. An integral domain with this property (using irreducible elements rather than primes) is called a *unique factorization domain*.

THE CENTRAL CONCERNS of number theory are the properties of the integers and rational numbers. Questions relating to individual real or complex numbers, integer matrices, algebraic number fields, points on curves, lattice points in convex regions, and similar entities are also considered. The subject has a very long history, as long as mathematics itself, and is as actively researched today as at any time in the past.

Number theory is not an organized theory in the usual sense but is a vast collection of individual topics and results, with some coherent subtheories and a long list of unsolved problems. Some of these topics are highly specialized, for example, giving a single solution to a Diophantine equation, whereas others have wide applicability, the Euclidean algorithm being an example. Methods range over virtually all mathematical disciplines, although group and field theory, linear algebra, and both real and complex analysis are the most commonly used. Problems and results are studied entirely for their own sake; the fact that many theorems are useful elsewhere is incidental.

In this article we shall treat a few of the more important topics not considered in the previous article. For each of these topics major results have been established recently, and this progress is likely to continue in the future.

The reader is referred to the article quoted above for an account of the history of the subject and the basic definitions and results concerning the properties of the integers—in particular, the fundamental theorem of arithmetic, prime numbers, greatest common divisors, and congruences. The fundamental theorem states that every integer $n > 1$ can be expressed uniquely (except for the order of the factors) as a product of prime numbers. The greatest common divisor of two integers a and b is denoted by (a, b); it can be found using the Euclidean algorithm. The basic results of congruence theory concern the conditions for the solution of linear and quadratic congruences and include the law of quadratic reciprocity of C. F. Gauss (1777–1855).

I. ALGEBRAIC NUMBER FIELDS

Algebraic numbers and integers are generalizations of rational numbers and integers, and algebraic number fields are extensions of the rational field $\mathbb{Q}$. In both cases many properties are preserved in the new situations, but not all; for example, unique factorization is often lost. Although these new notions are important in their own right, they are also important because many problems concerning the rational numbers are best treated in this more general context. For example, to show that one of Fermat's equations (see Section II.A.5) has no rational solutions we work in an algebraic extension of $\mathbb{Q}$ in which the left-hand side of the equation can be factorized into linear factors. We shall define these new notions now and discuss their relationship to the rational case.

Definition: A complex number α is called an *algebraic number* if rational numbers $c_1, \ldots, c_n$ can be found to satisfy the polynomial equation

$$\alpha^n + c_1\alpha^{n-1} + \cdots + c_n = 0.$$

The positive integer n is called the *degree* of α provided that α does not satisfy another polynomial equation whose degree is less than n. If each c_i $(i = 1, \ldots, n)$ is a rational integer (i.e., each $c_i \in \mathbb{Z}$) then α is called an *algebraic integer*.

For example, the complex numbers $2/3$, $(2 + i)/3$, and $2^{1/3}$ are algebric numbers, and 5, $\sqrt{-5}$, and $(1 + \sqrt{5})/2$ are algebraic integers. (Note that the last number is an algebraic integer because it satisfies the equation $x^2 - x - 1 = 0$). The nonalgebraic numbers, that is the transcendental numbers, are discussed in Section IV.

The algebraic number fields are defined as follows. Let α be an algebraic number of degree n, and let $\mathbb{Q}(\alpha)$ denote the set of all complex numbers of the form

$$b_0 + b_1\alpha + \cdots + b_{n-1}\alpha^{n-1},$$

where each $b_i \in \mathbb{Q}$. It can be shown that this set is closed under the usual operations of complex addition, subtraction, multiplication, and division and so is a field. It is called an *algebraic number field of degree n* over $\mathbb{Q}$ and its algebraic structure is similar to that of $\mathbb{Q}$ itself.

Given a field $K = \mathbb{Q}(\alpha)$, let O_K denote the set of algebraic integers in K. This set is closed under sums and products and so is an integral domain called the *ring of integers* in K. If α is an algebraic integer, we also define the set $\mathbb{Z}[\alpha]$ to be the collection of all expressions $g(\alpha)$ where g is a polynomial with rational integer

coefficients. Each element of $\mathbb{Z}[\alpha]$ is an algebraic integer, and so $\mathbb{Z}[\alpha] \subseteq O_K$. Note that this inclusion can be strict. For example if $\alpha = \sqrt{5}$ then $\mathbb{Z}[\alpha] = \{a + b\sqrt{5}: a, b \in \mathbb{Z}\}$, but as noted above, the number $(1 + \sqrt{5})/2$ is an algebraic integer; hence it belongs to the ring of integers of the field $\mathbb{Q}(\sqrt{5})$ but not to the domain $\mathbb{Z}[\sqrt{5}]$.

It can be shown that for each algebraic number field $K = \mathbb{Q}(\alpha)$ of degree n its ring of integers O_K has a $\mathbb{Z}$-basis $\{\gamma_1, \ldots, \gamma_n\}$. That is, algebraic integers $\gamma_1, \ldots, \gamma_n$ can be found so that each integer in O_K can be expressed in the form $c_1\gamma_1 + \cdots + c_n\gamma_n$ where $c_i \in \mathbb{Z}$, $i = 1, \ldots, n$. In the example above, the $\mathbb{Z}$-basis for O_K, where $K = \mathbb{Q}(\sqrt{5})$, is $\{1, (1 + \sqrt{5})/2\}$.

A "number theory" has been developed for these rings of integers O_K; it is similar to that for the rational case, but two main aspects are different—units and factorization. A *unit* ε in an algebraic number field K is an algebraic integer that divides 1 (that is, $\varepsilon\varepsilon' = 1$ for some $\varepsilon' \in O_K$).

In $\mathbb{Q}$ the only units are 1 and -1, but most algebraic number fields contain infinitely many units. For example, the units of the field $\mathbb{Q}(\sqrt{2})$ are $(3 + 2\sqrt{2})^m$ for each $m \in \mathbb{Z}$.

An irreducible element, which is the generalization of the notion of a prime number in $\mathbb{Z}$, is defined as follows: ξ is *irreducible* (over K) if $\xi \in O_K$, ξ is not a unit of zero, and whenever we have $\xi = \beta\gamma$, and β and γ belong to O_K, then either β or γ is a unit. Clearly, each member of O_K can be written as a product of a unit and a number of irreducible elements; but in many cases this representation is not unique, that is, K does not have unique factorization. For example, If $K = \mathbb{Q}(\sqrt{10})$, we have

$$39 = 3 \times 13 = (7 + \sqrt{10})(7 - \sqrt{10}),$$

and 3, 13, $7 + \sqrt{10}$, and $7 - \sqrt{10}$ are all irreducible elements in this field. It is an open problem to list all algebraic number fields that have unique factorization. One important case has been settled recently; that is the imaginary quadratic fields $\mathbb{Q}(\sqrt{n})$ where n is a negative integer. The only fields in this class that have unique factorization are those where $n = -1, -2, -3, -7, -11, -19, -43, -67$, and -163 (see Section IV).

The lack of unique factorization in many number fields is a severe drawback; it can be partially retrieved by introducing ideals into the theory. Let K be an algebraic number field with ring of integers O_K. An *ideal* I is a nonempty subset of O_K with the properties: if $a, b \in I$ then $a - b \in I$, and if $a \in I$ and $c \in O_K$ then $ac \in I$. Further, if I and J are ideals we define the product ideal IJ to be the set of all finite sums $\Sigma_{a_i b_j}$, where $a_i \in I$ and $b_j \in J$. Also I is called a *prime ideal* when $I \neq O_K$; and if J_1 and J_2 are ideals with $J_1 J_2 \subseteq I$ then either $J_1 = I$ or $J_2 = I$. It can be shown that each ideal I where $I \neq O_K$ (O_K acts as the identity ideal) can be represented *uniquely* in the form $I = P_1 \ldots P_n$ where each P_i is a prime ideal.

There is a close connection between the arithmetic of O_K and the arithmetic of these ideals. This is illustrated by the following important result. An ideal I is called *principal* if $I = \{at: a \in O_K\} = \langle t \rangle$ for some fixed $t \in O_K$; t is called the *generator* (of I) in O_K. We have O_K is a unique factorization domain if and only if each ideal in O_K is principal. To illustrate this theory we shall consider two examples.

(a) Let $K = \mathbb{Q}(i)$ where $i = \sqrt{-1}$. In this case the ring of integers $O_K = \{a + bi: a, b \in \mathbb{Z}\}$; it is known as the set of *Gaussian integers*. This is a unique factorization domain and so all ideals are principal. For instances the ideal $\langle 2, 3 + 3i \rangle = \{2a + (3 + 3i)b: a, b \in O_K\}$ equals the principal ideal $\langle 1 + i \rangle$ as $1 + i$ divides both 2 and $3 + 3i$ in the Gaussian integers.

(b) In our second example, let $K = \mathbb{Q}(\sqrt{10})$. Its ring of integers does not have unique factorization, and it has nonprincipal ideals. For instance, as noted above, $39 = 3 \times 13 = (7 + \sqrt{10})(7 - \sqrt{10})$, and these last four entries are irreducible in O_K. Further the ideal $\langle 3, 7 + \sqrt{10} \rangle$ is not principal because there is no $t \in O_K$ that divides both 3 and $7 + \sqrt{10}$. For the ideals we have

$$\begin{aligned} \langle 3 \rangle &= \langle 3, 7 + \sqrt{10} \rangle \langle 3, 7 - \sqrt{10} \rangle \\ \langle 13 \rangle &= \langle 13, 7 + \sqrt{10} \rangle \langle 13, 7 - \sqrt{10} \rangle \\ \langle 7 \pm \sqrt{10} \rangle &= \langle 3, 7 + \sqrt{10} \rangle \langle 3, 7 - \sqrt{10} \rangle \end{aligned}$$

and the ideal $\langle 39 \rangle$ can be written uniquely as the product of the four prime ideals $\langle 3, 7 + \sqrt{10} \rangle$, $\langle 3, 7 - \sqrt{10} \rangle$, $\langle 13, 7 + \sqrt{10} \rangle$, and $\langle 13, 7 - \sqrt{10} \rangle$.

Algebraic number theory can be viewed in two ways. In the first it is a branch of modern algebra and provides some good examples of structures for that theory to work with. In the second way, it provides a more general context in which some problems concerning the rational numbers and integers, especially Diophantine equation problems, can be viewed. For a good introduction to the whole subject see Stewart and Tall (1979).

II. DIOPHANTINE EQUATIONS

This topic is the largest and most important in number theory; it concerns the solution of polynomial equations in some specified number system, often the integers $\mathbb{Z}$ or the rational numbers $\mathbb{Q}$. The name refers to the early Greek mathematician Diophantus, who was working in Alexandria in the middle of the third century. His surviving books show that he had a highly developed knowledge of the solution of equations in integers, especially those involving sums of squares. Nowadays although, some

methods recur, a large array of ideas and arguments are used; this is particularly true when integer solutions are sought. When solutions in an algebraic number field (for example $\mathbb{Q}$) are required, a more organized theory based on ideas from algebraic geometry has been developed; we shall discuss this in the next section. A good survey of the whole topic can be found in Mordell (1969).

We begin by listing some of the major results in Diophantine equation theory. The list is by no means complete, but it will give the reader some idea of the range of theorems and methods used.

A. Some Major Results in Diophantine Equation Theory

1. Linear Diophantine Equations

If $a, b, n \in \mathbb{Z}$, then the equation

$$ax + by = n$$

has a solution in integers x and y if and only if $(a, b) \mid n$, and if x_0, y_0 is a solution then the general solution is given by $x_0 + tb/(a, b)$, $y_0 - ta/(a, b)$ where $t \in \mathbb{Z}$. The Euclidean algorithm provides a good method for finding solutions of these equations. This result is one of the most useful in the whole of number theory.

2. Pell's Equation and Quadratic Forms

Suppose d is a positive integer and not square, then the equation

$$x^2 - dy^2 = 1$$

has infinitely many integer solutions x, y generated from a fundamental solution x_0, y_0 using the relation $x + y\sqrt{d} = (x_0 + y_0\sqrt{d})^n$ for $n \in \mathbb{Z}$. The fundamental solution can be found using the continued fraction expansion for $\sqrt{d}$ (see Section IV). It is an accident that Pell's name has been attached to this equation; J.-L. Lagrange (1736–1813) gave the first proof of its solvability.

A quadratic form (over $\mathbb{Z}$) is a function $F: \mathbb{Z}^n \to \mathbb{Z}$ given by the equation

$$F(x_1, \ldots, x_n) = \sum_{i,j=1}^{n} a_{ij} x_i x_j,$$

where, for $1 \le i, j \le n$, a_{ij} is an integer satisfying $a_{ij} = a_{ji}$. An extensive theory has been developed for solutions of equations of the form

$$F(x_1, \ldots, x_n) = m.$$

It relies to some extent on the theory of Pell's equation discussed above. The results are generally straightforward, if a little complicated to state. Equations can have either a finite or an infinite number of solutions (this is related to the underlying geometry); in the latter case the solutions are generated by a finite set of fundamental solutions. Accounts are given in Rose (1999) [two-variable integer case] and Cassels (1978) [general field and integral domain cases].

3. Thue–Siegel–Roth Theorem

This is an example of a major class of equations that have only finitely many solutions. One version of the theorem is as follows: Suppose f is an irreducible homogeneous polynomial, without multiple roots, in the rational number variables x and y, and having degree $n > 2$. Then if the equation $f(x, y) = m$, $m \in \mathbb{Z}$, is soluble, it has only finitely many solutions. The proof uses Diophantine approximation theory (see Section IV). T. A. Skolem (1887–1963) has given an ingenious method for solving some of these equations; see Borevich and Shafarevich (1966). Also, G. Faltings has greatly extended this result; see Section III.

4. Goldbach Conjecture

Here, we look for prime number solutions only; the result has not yet been fully established. C. Goldbach (1690–1764) conjectured that every even positive integer larger than 2 can be expressed as a sum of two prime numbers. For example $10 = 7 + 3$, $100 = 53 + 47$, $1000 = 509 + 491$, etc. At the present time, no counterexample has been found and J.-R. Chen has proved that every large even integer is the sum of a prime a number that has at most two prime factors. Also, I. M. Vinogradov (1891–1983) has shown that every large odd positive integer is the sum of three primes. The proofs of both of these results are long and complex.

5. Fermat's "Last Theorem"

First proposed in 1637 and finally proved in 1995, this is undoubtedly the most famous and well-researched Diophantine equation. We shall give a brief history, for more information see Edwards (1977), Ribenboim (1979), Cornell (1997) (for a detailed account of the whole proof), and Singh (1997) (for a popular account including the recent advances).

Although of little intrinsic interest in itself, Fermat's so-called Last Theorem has had a profound influence of the development of pure mathematics over the past three and a half centuries. It states: the equation F_n

$$x^n + y^n = z^n$$

has no solution in integers x, y and z if $xyz \neq 0$ and $n > 2$. (There are solutions when $n = 2$.) In 1637, P. de Fermat

(1601–1673) proposed this result and claimed to have a proof, which has not survived, although he did prove the result in the case $n=4$ and so showed that n can be restricted to be an odd prime number. The next step was taken by L. Euler (1707–1783) who proved Fermat's result for $n=3$. He was the first to realize that the solution to the problem would come by working in a number system larger than the integers; he used the complex numbers in his proof. In the middle of the nineteenth century, E. E. Kummer (1810–1893) greatly extended this work and so proved the result for a large number of cases (and as a by-product helped to lay the foundation of modern algebra). Building on this and using the power of modern computers, Fermat's Last Theorem had been established for all $n < 4{,}000{,}000$ by 1980. Also by Falting's work (see next section), it was known that even if one of the equations F_n was solvable, it could only have a finite number of solutions.

In 1985, G. Frey noted an apparently simple connection between Fermat's result and some properties of a class of elliptic curves (see Section III). Frey's elliptic curves have the form

$$y^2 = x(x - a^n)(x - c^n).$$

Using these curves, he made the following conjecture: if Fermat's Last Theorem is *false* then so is the Taniyama and Weil Elliptic Curve conjecture (see the last paragraph of Section III). He noted that if $a^n - c^n = b^n$ and $a, b, c \in \mathbb{Z}$, then the discriminant Δ of the curve has the form $\Delta = (abc)^{2n}$, and so he conjectured that if an elliptic curve whose discriminant was a $2n$-power existed, then it could not satisfy the Taniyama and Weil conjecture. This implication was proved by K. Ribet in 1990. At about this time, A. Wiles began work on proving that the Taniyama and Weil conjecture is *true*, one consequence of which would be, by Frey and Ribet's result, that Fermat's result was finally proved. His work was successful for in 1995, and with some input from R. Taylor, he published his proof of the Taniyama and Weil conjecture for a large class of elliptic curves, which includes the Frey curves, and so finally established Fermat's Last Theorem 358 years after it had been first proposed. This work will surely rank as one of the greatest achievements of twentieth century mathematics.

6. Sums of Squares

Here we ask: Can a positive integer m be expressed as a sum of k squares, or l cubes, or in general as a sum of nth powers? We shall look at the squares case first. Lagrange's famous four-squares theorem states that every positive integer is a sum of four squares. The usual proof uses the so-called "infinite descent method," which is a kind of mathematical induction. By a simple counting argument we find, for our given m, positive integers t, x, y, z, w satisfying $x^2 + y^2 + z^2 + w^2 = tm$. Secondly, given this equation, a procedure is devised to give a new solution t_1, x_1, y_1, z_1, w_1 to this equation, which satisfies $0 < t_1 < t$. The result follows by finitely many applications of this procedure.

The two-square result first proved by Fermat, is: the equation

$$x^2 + y^2 = n$$

has a solution x, y provided those factors of n that are congruent to 3 modulo 4 occur in the prime factorization of n with an even power. So for example, 1, 2, 4, 5, 8, 9, 10, 13, 16, 17, and 18 are the positive integers less than 20 that can be expressed as sums of two squares. This can also be established using the infinite descent method, although many other proofs exist. The proofs of both of these results rely on the fact that the product of two sums of 2 or 4 squares is itself a sum of 2 or 4 squares, respectively. For example, we have

$$(x^2 + y^2)(z^2 + w^2) = (xz + yw)^2 + (xw - yz)^2$$

(a result known to Diophantus in the third century AD.) No similar identity is available for an odd number of squares, and so in this case different methods involving quadratic form theory are used. The three-squares theorem, first provided by Gauss, states that n can be expressed as a sum of three squares proved it is not of the form $4^a(8b + 7)$. For example, every positive integer less than 30 is a sum of three squares except 7, 15, 23, and 28.

7. Waring's Problem

More generally we can ask if all positive integers are sums of kth powers with $g(k)$ summands where g depends only on k. This is known as Waring's problem. In 1770, E. Waring (1736–1798) postulated that $g(2) = 4$ (this is Lagrange s thoerem), $g(3) = 9$, $g(4) = 19$, and so on. In 1909, D. Hilbert (1862–1943) showed that g exists for all k. It has been established that $g(3) = 9$ and $g(5) = 37$; very recently it has been shown that $g(4) = 19$, the value Waring conjectured. Further, for most $k \geq 6$, $g(k) = 2^k + A - 2$, where A denotes the largest integer less than $(3/2)^k$. For the exact result see Hardy and Wright (1954) pages 335–337.

One aspect distinguishes the case $k = 2$, discussed in Subsection 6, and the case $k > 2$; in the latter case there are a few exceptional integers that require more summands than usual. For instance only 23 and 239 need nine cubes, and if $n > 8042$ then, in all probability, six cubes are sufficient. If we let $G(k)$ denote the least number of

summands needed to express all sufficiently large positive integers as a sum of kth powers, then clearly

$$G(k) \le g(k) \quad \text{and} \quad G(2) = g(2) = 4.$$

For most k the precise value of $G(k)$ is not known. $G(3)$ lies between 4 and 7 with the most likely values 4 or 5, and $G(4) = 16$. G. H. Hardy (1877–1947) and J. E. Littlewood (1885–1977) conjectured that $G(k) \le 2k + 1$ if $k \ne 2^m$, where $m > 1$, and $G(2^m) = 2^{m+2}$ again where $m > 1$. At the moment the best result known is: as $k \to \infty$, $G(k) \le (2 + c_k)k \ln k$, where $c_k \to 0$ as $k \to \infty$. For an account of this work see Vaughan (1981) and Ribenboim (1989), pages 236–245, which gives the current status of this problem when $k \le 10$.

III. ELLIPTIC CURVES

In this section we discuss the solution of two variable Diophantine equations defined over an algebraic number field. Compared with the material discussed in the previous section, a more organized theory has been developed; it uses some basic ideas from algebraic geometry. Also, it is the subject of considerable current research interest.

Let K be an algebraic number field, for example the rational field $\mathbb{Q}$, and let $A^n(K)$ denote the set of n-tuples $(x_1, \ldots, x_n)$ where each $x_i \in K$. $A^n(K)$ is called the n-dimensional *affine space* over K. On the set $A^3(K) - \{(0, 0, 0)\}$ we define the relation $\simeq$ by: $(x, y, z) \simeq (x', y', z')$ if and only if $x = x't$, $y = y't$, $z = z't$ for some nonzero $t \in K$. This is an equivalence relation, and the set of equivalence classes is called the two-dimensional *projective space* $P^2(K)$ over K. Points in this space are denoted by $(x{:}y{:}z)$. Note that by the relation $\simeq$ it is only the ratios of x, y, and z that matter; so, for example, (2:3:5), $(-4{:}-6{:}-10)$, and (26:39:65) all represent the same point.

A *curve* C in $P^2(K)$ is the set of points $(x{:}y{:}z)$ that satisfy a homogeneous polynomial equation $f(x, y, z) = 0$, and the *degree* of C is the degree of the polynomial f. A point P on C is called *singular* if the partial derivatives ($\partial f/\partial x$, etc.) at P with respect to all three variables x, y, and z are zero; unique tangents can be drawn to the curve at points that are not singular. Two curves C and C' are said to be *birationally equivalent* if each can be transformed into the other by a rational function and the correspondence is one-to-one and onto except at a finite number of points. For example $x' = 1/(x - 1)$, $y' = 1/y$, $z' = 2z$ is birational. Note that the degrees of C and C' need not be identical.

We come now to the important notion of the genus of a curve.

Definition: The genus g of a curve C that has degree n and has N singular points is given by $g = \frac{1}{2}(n - 1)(n - 2) - N$.

It can be shown that g is a nonnegative integer and is unaltered when C is birationally transformed into a new curve C'. The genus provides an important classification, into three main classes, for the collection of all curves defined over the field K. The first class contains curves of genus zero. All curves in this class can be reduced (by birational transformations) to lines or conics, and the linear and quadratic form material discussed in the previous section applies to the associated Diophantine equations.

The second class contains the curves of genus one. Using birational transformations, all curves in this class that have at least one point whose coordinates lie in K can be represented by equations of the form

$$y^2z = x^3 + axz^2 + bz^3, \tag{1}$$

where $a, b \in K$. They are called *elliptic curves* because they can be parametrized using elliptic functions. The *discriminant* Δ of the curve (1) is $-16(4a^3 + 27b^2)$. Many of the basic properties of this curve are determined by this quantity, for example the curve (1) is elliptic if and only if $\Delta \ne 0$ and, in the real plane, it has one component if $\Delta < 0$ and two if $\Delta > 0$. The last main class contains curves having genus larger than one. Until recently little was known about these equations. In 1922 L. J. Mordell (1888–1972) conjectured that each equation in this class has only finitely many points whose coordinates lie in K. This remarkable conjecture was established by Faltings in 1983, using advanced techniques from algebraic geometry. We have already mentioned one consequence of this work. If $n > 3$, the genus of each of the curves associated with the Fermat equation F_n (see Section II.A.5) is larger than one, and so none of these equations can have infinitely many solutions. This fact provided initial support to Wiles and others in their eventual solution of Fermat's result. Table I summarizes the main classes for the field $\mathbb{Q}$.

A. Elliptic Curves

Points on elliptic curves have an important algebraic structure that we shall discuss now. If the point $(x{:}y{:}z)$ lies on the curve C (defined over the field K) and $x, y, z \in K$, then we call the point *rational*; and we let $C(K)$ denote the set of rational points on C. Using the so-called *chord-tangent method*, $C(K)$ can be given a group structure as follows. Let us suppose that C is represented in the standard form, Eq. (1). In this case it is a cubic curve and any straight line will intersect it in at most three points. Further, if two of these points are rational than the coefficients of the

TABLE I Properties of Homogeneous Three Variable Diophantine Equations Defined Over $\mathbb{Q}$

Genus	Parametrization	Maximum number of solutions in $\mathbb{Z}$	Maximum number of solutions in $\mathbb{Q}$
0	By rational functions	Infinite	Infinite
1	By elliptic functions	Finite	Infinite but finitely generated
≥ 2	—	Finite	Finite

equation of the line through these points will belong to K, and so the third point will also be rational. This conclusion is valid even if the first two points coincide at P, that is, the line is the tangent at P.

Using this chord-tangent process we can define an addition operation on $C(K)$. Let O be the point (0:1:0), it lies on all curves of the form, Eq. (1), and it will act as the identity of the group. Suppose P_1, $P_2 \in C(K)$ and the line P_1P_2 meets C again at Q; as noted above, $Q \in C(K)$. Now construct the line OQ; it will meet C in one further rational point. This new point is called the *sum* of P_1 and P_2 and is denoted by $P_1 + P_2$. It can be shown that $C(K)$ with this operation is an Abelian group. The proof is straight-forward except for associativity, which requires a result from algebraic geometry known as Bezout's theorem. Some sample constructions are given in Fig. 1. In this example, $P_1 + P_2 = 2P$ where $2P$ denotes the tangent point $P + P$.

In 1922 Mordell showed that the Abelian group $C(\mathbb{Q})$ has only a finite number of generators; that is, all rational points on $C(\mathbb{Q})$ can be constructed by the chord-tangent method from a finite set of points. Later, A. Weil extended this to all algebraic number fields K, and the result is now called the Mordell–Weil theorem. The group $C(K)$ can be finite or infinite. Recently B. Mazur has classified all groups that can occur in the finite case when $K = \mathbb{Q}$; they are either cyclic or direct products of two cyclic groups and have orders not larger than 16.

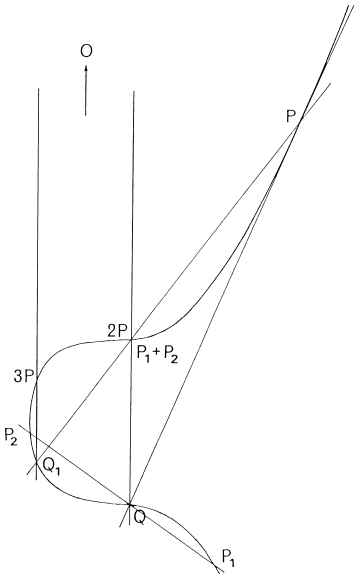


FIGURE 1 Sample constructions of rational points on elliptic curves by the chord-tangent method.

The *rank* of $C(K)$ is the number of infinite order generators of $C(K)$; it is finite by the Mordell–Weil theorem. This number has been studied extensively, and recently some results have been established. For many curves the rank is zero or one; but examples have been given in which the rank is at least 22. For a fixed field K, it is not known if the rank can be arbitrarily large or, conversely, if a constant m exists such that the rank of all elliptic curves over K is less than m.

There are many conjectures concerning these curves, the most important involve L-functions, a detailed definition of which can be found in Silverman (1986). Roughly speaking, an L-function for a curve C encodes "local" properties (that is, properties of C defined over finite fields, working modulo a prime for example) in the hope of obtaining "global" properties of C (that is, properties of C valid over the rational numbers or the complex numbers). Two main collections of conjectures have been studied; the first, due to B. Birch and H. P. F. Swinnerton–Dyer, relates the behavior of the L-function of a curve C near 1 to the rank of C over the rational numbers—there is good numerical evidence for this conjecture and it has been proved recently for some classes of curves. The second main collection of conjectures concerns the domain of definition of the L-functions in the complex plane. It is a simple matter to see that these functions are defined at most points z in the right hand side of the plane (to be precise, when the real part of z is larger than $3/2$). The Taniyama and Weil conjecture states that the domain of definition can be extended to cover the whole of the complex plane except for the point 1. It is known that if this conjecture holds for the curve C in question, and a set of related curves, then the curve is "modular" which implies that it possesses a number of useful arithmetic properties including a special type of parameterization. It is this conjecture (for a subset of all curves, called the *semistable* curves, which contains the Frey curves related to Fermat's Problem) that A. Wiles (and R. Taylor) proved in 1993/1995 [see Wiles (1995)] and which finally established that Fermat's Last Theorem is valid in all cases (see Section A.5). At the time of writing (November 1999), it has been announced by C. Breuil, B. Conrad, F. Diamond, and R. Taylor that the Taniyama and Weil conjecture is valid for all elliptic curves defined over the rational numbers, no details are available.

IV. DIOPHANTINE APPROXIMATION AND TRANSCENDENCE

These topics are concerned with the number-theoretic properties of the real and complex numbers. The starting point for this work is the question: How close can a rational number a/b approximate to a real number α assuming some restriction on the size of a and b; that is, given α can integers a and b be found to satisfy

$$\left|\alpha - \frac{a}{b}\right| < \frac{1}{f(b)} \tag{2}$$

for some suitably chosen monotone increasing function f? This is the basic question in Diophantine approximation theory. We answer this question and discuss some related topics in the subsections below.

Continued fractions play an important role in this theory; they are defined as follows. Let α be a real number and let $[\alpha]$ denote the largest integer c satisfying $c \le \alpha$; then we can write

$$\alpha = [\alpha] + 1/\alpha_1,$$

where $\alpha_1 > 1$ provided α is not an integer. We repeat this construction by setting

$$\alpha_n = [\alpha_n] + 1/\alpha_{n+1}$$

$(n = 1, 2, \ldots)$ provided α_n is not an integer. If α is a rational number this process will stop in a finite number of steps, otherwise it will continue indefinitely. We write q_0 for $[\alpha]$, q_n for $[\alpha_n]$, and $[q_0, q_1, \ldots, q_n]$ for the expression

$$q_0 + 1/(q_1 + 1/(q_2 + \cdots + 1/q_n \ldots)).$$

This is called the nth *continued fraction convergent* to α. For example, the fifth continued fraction convergent to $\sqrt{2}$ is $[1, 2, 2, 2, 2, 2]$.

A. Rational Approximation

1. Best Approximation

By a "good" rational approximation a/b to a real number α we mean that a/b is close to α, and a and b are relatively small. We define the *best approximation* to a real number α relative to n to be the rational number a/b closest to α satisfying $0 < b < n$. For example, $22/7$ is a good approximation to π; it is in fact the best approximation relative to any $n \le 56$. All good approximations to a real number α are continued fraction convergents to α.

2. Hurwitz's Theorem

The theorem of A. Hurwitz (1859–1919) puts a limit on how good approximations can be in general. It states that if α is irrational then the inequality (2) has infinitely many solutions a, b provided $f(b) \le b^2\sqrt{5}$. This result can be derived using some elementary theorems concerning continued fractions [see Rose (1999)]. It is best possible for if $\alpha = (1 + \sqrt{5})/2$ (its continued fraction representation is $[1, 1, 1, \ldots]$) and $f(b) > b^2\sqrt{5}$ then inequality (2) has only finitely many solutions. If this α and any real number whose continued fraction expansion is eventually all ones is excluded, then Hurwitz's theorem can be improved by replacing $\sqrt{5}$ by $\sqrt{8}$. Then $\sqrt{8}$ is best possible when $\alpha = \sqrt{2}$. This process continues; finally it can be shown that there are uncountably many real numbers α for which the inequality (2) has infinitely many solutions if $f(b) = 3b^2$, but only finitely many if the number 3 is replaced by a larger number.

3. Sets of Real Numbers Modulo 1

Let $((\alpha))$ denote the fractional part of α, that is $((\alpha)) = \alpha - [\alpha]$. P. L. Chebyshev (1821–1894) was the first to ask: Given α how is the set $\{((n\alpha)): n = 1, 2, \ldots\}$ distributed in the unit interval? Using the methods discussed above, he was able to show that this set is dense in the unit interval and that the distribution is uniform provided α is irrational. His methods apply to a number of similar problems.

B. Transcendental Numbers

A real or complex number that satisfies no polynomial equation with algebraic coefficients is called *transcendental*. J. Liouville (1809–1882) was the first to show that transcendental numbers exist, although using the diagonal argument of G. Cantor (1845–1918) we now know that almost all real or complex numbers have this property. On the other hand there are many specific numbers whose transcendental status is unknown. Examples of Euler's constant γ, $e + \pi$, and $\pi^{\sqrt{2}}$.

Liouville began with the inequality (2) and showed that if α is an algebraic number of degree $n (n > 1)$ then (2) has only finitely many solutions if $f(b) = b^{n+k}$ and $k > 0$. Using this he was able to show that certain real numbers are transcendental; for example the number whose decimal expansion is $0.11000100\ldots$, that is, all zeros except for 1's at the $n!$ digit places, $n = 1, 2, \ldots$. A number of this type is now called a *Liouville number*.

C. Hermite (1822–1901) introduced the main method for establishing the transcendence of a number in 1873. It begins by assuming that α is algebraic, satisfying the polynomial equation $f(\alpha) = 0$; that is the contrary of the required result. Then two properties of a formula F constructed using the coefficients of f are derived that contradict one another, and so transcendence is established. For example, the contradictory properties could be (a) F is a

positive integer and (b) F tends to zero. Hermite used this method to show that e $(=\sum_{n=1}^{\infty} 1/n!)$ is transcendental. F. Lindemann (1852–1939) extended this to show that if $\alpha_1, \ldots, \alpha_n$ are distinct algebraic numbers and $c_1, \ldots, c_n$ are nonzero algebraic numbers, then

$$c_1 e^{\alpha 1} + \cdots + c_n e^{\alpha_n} \neq 0. \tag{3}$$

Using this result he was able to show that the following numbers are transcendental where α is algebraic, nonzero, and not equal to one in the last two cases:

$$\pi, e^{\alpha}, \sin \alpha, \cos \alpha, \sinh \alpha, \cosh \alpha,$$

$$\arcsin \alpha, \text{ across}\, \alpha, \text{ and } \ln \alpha.$$

A famous problem first proposed by the ancient Greeks and known as "squaring the circle" was: Construct a square equal in area to a given circle using only a ruler and compass. Lindemann's result shows that this is impossible as $\sqrt{\pi}$ is transcendental. Further extensions of the method enabled A. O. Gelfond (1906–1968) and T. Schneider to show that if α and β are algebraic numbers, $\alpha \neq 0$ or 1, and β is irrational then α^{β} is transcendental.

One consequence of this result is: If α, β, γ and δ are nonzero algebraic numbers and no linear relation with rational coefficients exists between $\ln \beta$ and $\ln \delta$, then

$$\alpha \ln \beta + \gamma \ln \delta \neq 0.$$

In 1966, A. Baker extended this by providing a similar result with n rather than 2 summands [see Lindemann's result, inequality (3)]. A number of important applications have followed. For example, the number $e^{\alpha} \beta^{\gamma}$ is transcendental provided α, β and γ are algebraic and nonzero.

We shall mention two further applications.

(a) In Section III, we introduced the elliptic curves; Baker's result provides an effective finite bound on the number of integer (as opposed to rational number) solutions of the associated Diophantine equations.

(b) Baker's result enables us to list all those imaginary quadratic fields that have unique factorization (see Section I).

1. Mahler's Classification

The set of all real and complex numbers is divided into four classes A, S, T, and U as follows:

1. The *height* h of a polynomial f with rational integer coefficients, not all zero, is the maximum of the absolute values of its coefficients.
2. Given γ, n, and $h > 1$, let f be the polynomial with rational integer coefficients, degree at most n, and height at most h, such that $|f(\gamma)|$ takes the least nonzero value. Now let $\omega(n, h)$, ω_n, and ω, be given by

$$|f(\gamma)| = h^{-n\omega(n,h)}$$

$$\omega_n = \lim_{h\to\infty} \max_{1\leq k\leq h} \omega(n, k)$$

$$\omega = \lim_{n\to\infty} \max_{1\leq m\leq n} \omega_m$$

Finally, let ν be the least positive integer n such that $\omega_n = \infty$, and let $\nu = \infty$ if $\omega_n < \infty$ for all n.

There are four possibilities for the values of ω and ν (note ω and ν cannot both be finite). The scheme of K. Mahler (1888–1972) defines

- γ to be an A-number if and only if $\omega = 0$ and $\nu = \infty$,
- γ to be an S-number if and only if $0 < \omega < \infty$ and $\nu = \infty$,
- γ to be a T-number if and only if ω and ν are both infinite, and
- γ to be a U-number if and only if $\omega = \infty$ and ν is finite.

The following results concerning this classfication have been proved.

(a) If two numbers α and β are algebraicly dependent (i.e., $g(\alpha, \beta) = 0$ for some two-variable polynomial g) then they belong to the same class.

(b) The A-numbers are exactly the algebraic numbers, and so the S-, T-, and U-numbers are transcendental.

(c) Almost all real and complex numbers are S-numbers.

(d) The Liouville numbers are U-numbers; in the exampleof a Liouville number given above $\omega_1 = \infty$.

(e) T-numbers exist; this is a recent result of W. M. Schmidt. Many problems remain, for example it is known that π is an S- or T-number, but which? For further details on all of the results in this section see Baker (1975), LeVeque (1955), and Schmidt (1980).

V. PRIME NUMBER THEORY

The properties of the prime numbers have been studied since the time of the early Greeks Pythagoras and Euclid; even so, many questions remain unanswered today. The first major result, which is attributed to Euclid, states that the set of prime numbers is infinite. The other major result from this period is the so-called "Sieve of Eratosthenes," which is an effective method for enumerating all primes less than some fixed integer.

Two pre-eminent results have been established in modern times. They are the theorem of P. G. L. Direchlet

(1805–1859) concerning primes in arithmetic progressions and the Prime Number Theorem (PNT), which estimates the density of the primes. We shall discuss these, but we shall begin by considering some of the many unsolved problems in prime number theory. Most of the conjectures associated with these problems are backed up with ample numerical evidence, but this should be treated with some skepticism because in most cases this evidence involves only relatively small numbers. For example, most of the available numerical evidence suggests that $Li(x) > \pi(x)$ for all x. [These two prime density functions are defined below.] But it is now known that $Li(x) - \pi(x)$ changes sign infinitely often, and the first change occurs before $x = 6.69 \times 10^{370}$.

A. Conjectures Concerning Prime Numbers

1. The Twin Prime Conjecture

A brief study of a table of prime numbers shows that there are many pairs of primes p, q where $q = p + 2$; examples are 3, 5; 101, 103; 1997, 1999; and $10^9 + 7$, $10^9 + 9$. If we let $\pi_2(x)$ denote the number of prime pairs less than x;then, for instance, $\pi_2(10^3) = 35$ and $\pi_2(10^6) = 8164$. The twin prime conjecture states that $\pi_2(x) \to \infty$ as $x \to \infty$, and the likely value of $\pi_2(x)$ is $\bigcirc(f_2^x (\ln t)^{-2}\, dt)$. Progress has been made on this conjecture, for Chen has shown that there are infinitely many pairs of integers p, $p + 2$ where p is prime and $p + 2$ has at most two factors.

2. Primes in Intervals

It is a simple matter to show that a prime occurs between n and $2n$ for all integers $n > 0$. This is known as "Bertrand's Postulate"; for a proof see Rose (1999), page 231. Using the Prime Number Theorem this result can be extended to show that for all large n there is a prime between n and $(1 + \varepsilon)n$ for any given fixed $\varepsilon > 0$. But no similar result for a smaller interval has been established. For example, it is not known if a prime occurs between n^2 and $(n + 1)^2$ for all large n.

3. Extensions of Dirichlet's Theorem

Very little information is available concerning the set D of those n for which $an + b$ is prime. (Dirichlet's theorem tells us that it is infinite.) For example, does D itself contain infinitely many primes, or any primes at all? Another proposed extension replaces the linear form $an + b$ by a quadratic one. For instance, it is not known if there are infinitely many primes of the form $n^2 + n + 1$ or $n^2 + n + p$ for some prime p.

A number of other problems concerning the primes have been discussed elswhere. These include the Goldbach conjecture (see Section II.A.4), problems concerning formulas taking prime values, and Fermat and Mersenne numbers.

We shall discuss now the two major results mentioned earlier and their connections with the Riemann hypothesis.

B. Dirichlet's Theorem

Suppose $(a, b) = 1$; this famous theorem first established in 1837 states that there are infinitely many prime numbers in the arithmetic progression $A = \{an + b, n = 1, 2, \ldots\}$. It is proved by considering a series of the form $\sum \chi(p)/p$ where the sum is taken over the primes p in A. The function χ is used to pick out elements of A and is defined using some basic group theory (it is called a multiplicative character). Second, standard analytic techniques are applied to this series to show that it diverges to infinity. Dirichlet's theorem follows, for if there were only finitely many primes in the progression A then the series would have a finite sum.

C. The Prime Number Theorem (PNT)

Let $\pi(x)$ denote the number of primes p satisfying $p \le x$. J. S. Hadamard (1865–1963) and C. de la Vallee Poussin (1866–1962) proved in 1896 that

$$\pi(x) \sim x / \ln x.$$

A more precise version of this result is

$$\pi(x) = Li(x) + \bigcirc(x/(\ln x)^2),$$

where

$$Li(x) = \lim_{\epsilon \to 0} \int_0^{1-\varepsilon} + \int_{\varepsilon+1}^{x} \frac{dt}{\ln t}.$$

There are two main proof of PNT, an analytic one and an elementary one. A typical analytic proof begins with the result: if s is a complex variable whose real part is larger than one then

$$\ln \zeta(s) = s \int_2^{\infty} \frac{\pi(x)}{x(x^s - 1)}\, dx. \tag{4}$$

[The zeta function $\zeta(s)$ is defined below.] An estimate for $\pi(x)$ is then-obtained by "solving" this equation for $\pi(x)$. We shall consider the error term in this process later. Elementary proofs of PNT rely on the following formula, which was established by A. Selberg in 1949.

$$\sum_{p \le x} \ln^2 p + \sum_{pq \le x} \ln p \ln q = 2x \ln x + \bigcirc(x),$$

where the sums in this formula are taken over prime values only. This is used to improve some standard estimates, and the result follows after a lengthy argument. The best estimates for $\pi(x)$ are achieved using analytic methods.

D. The Riemann Hypothesis

For a complex variable $s = \sigma + it$ we define the Riemann zeta function $\zeta(s)$ by

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

for $\sigma > 1$ and $-\infty < t < \infty$. Using analytic continuation, the domain of definition can be extended to the whole complex plane, except the point 1. The connection between the prime numbers and the zeta function is given by the following result [due to Euler.]

$$\zeta(s) = \prod_p (1 - p^{-s})^{-1},$$

where the product is taken over all primes p. Taking logs of both sides we obtain

$$\ln \zeta(s) = -\sum_p \ln(1 - p^{-s})$$

$$= \sum_p p^{-s} + \text{finite error term},$$

that is, a sum taken over the prime numbers only has been expressed by standard analytic terms. The result (4) follows from this easily.

We can see from the foregoing equation that it is important to know where the zeros of the zeta function are. It has zeros at $s = -2, -4, \ldots$, and the remaining zeros lie inside the region of the complex plane bounded on the left by the imaginary axis and on the right by the vertical line passing through the point $s = 1$; this is called the *critical strip*. The hypothesis of B. Riemann (1826–1866) states that all the zeros in the critical strip lie on the central line, that is, the vertical line through the point $s = \frac{1}{2}$. Hardy showed that there are infinitely many zeros on this line, but it is not known if there are any zeros off the line. This is one of the most important unsolved problems in the whole of mathematics. One major consequence of the establishment of this hypothesis would be greatly improved estimates for many theorems in analytic number theory. We give two examples. The error term $x/(\ln x)^2$ in the prime number theorem $\pi(x) = Li(x) + \bigcirc(x/(\ln x)^2)$ could be replaced by $\sqrt{x} \ln x$ if Riemann's hypothesis was available. Second, it is known that the first quadratic nonresidue n of a prime p satisfies $n < p^{1/4+\varepsilon}$ for large p and $\varepsilon > 0$. (The integer n is a quadratic nonresidue modulo p if the congruence $x^2 \equiv n \pmod{p}$ is insoluble). If Riemann's hypothesis was available then this inequality could be replaced by $n < c(\ln p)^2$ for some constant c. For further applications of this hypothesis see Titchmarsh (1987), for the latest numerical evidence see Odlyzko (1991), and for an informal but detailed discussion of the whole topic of prime numbers see Ribenboim (1989).

VI. PARTITIONS

Unlike much of the previous material, partition theory deals with additive problems. These are often very easy to state, but much of the development is as difficult as any in number thory. Every positive integer can be expressed as a sum of smaller integers in a number of ways. If $n > 0$ and

$$n = a_1 + a_2 + \cdots + a_r$$

where each a_i is a positive integer, then the set $\{a_1, a_2, \ldots, a_r\}$ is called a *partition* of n and the individual terms $a_1, a_2, \ldots$ are called *parts*. The order of the parts is unimportant, although it is usual to write them in decreasing order. The total number of partitions of n is denoted by $p(n)$. For example, $p(5) = 7$ as

$$5, \quad 4+1, \quad 3+2, \quad 3+1+1, \quad 2+2+1,$$
$$2+1+1+1, \quad 1+1+1+1+1$$

is the set of partitions of 5.

There are many relationships between subsets of the set of partitions of an integer n. The most famous, due to Euler, states that the number of partitions of n all of whose parts are odd integers equals the number of partitions of n where no part is repeated. In the example above, the first three partitions have distinct parts while the first, fourth, and seventh have exclusively odd parts. A number of similar properties will be discussed below.

Euler's result is remarkable because there is no obvious connection between the two sets, but it becomes almost a triviality once it has been expressed in terms of generating functions. We shall define these now. Let $\{c_1, c_2, \ldots\}$ be an infinite sequence of positive integers; then the power series

$$f(q) = \sum_{n=0}^{\infty} c_n q^n$$

is called the *generating function* for this sequence. These power series are treated formally, but in most cases they are absolutely convergent when $0 < q < 1$. To prove Euler's result let $p_1(n)$ denote the number of partitions of n with exclusively odd parts and let $p_2(n)$ denote the number of partitions of n with distinct parts; then we can show that

$$\sum_{n=0}^{\infty} p_1(n) q^n = \prod_{n=1}^{\infty} (1 - q^{2n-1})^{-1}$$

and

$$\sum_{n=0}^{\infty} p_2(n) q^n = \prod_{n=1}^{\infty} (1 + q^n)$$

Now

$$\sum_{n=1}^{\infty}(1+q^n)=\prod_{n=1}^{\infty}\frac{(1-q^{2n})}{(1-q^n)}$$
$$=\prod_{n=1}^{\infty}(1-a^{2n-1})^{-1}$$

and this gives $p_1(n)=p_2(n)$ for all n.

To illustrate the theory we list some further results:

1. The number of partitions of n with at most m parts equal the number of partitions of n in which no part exceeds m.
2. The number of partitions of n in which no part occurs more than three times equals the number of partitions of n in which only odd parts may be repeated.
3. If $n\equiv 4 \pmod 5$ then $p(n)$ is divisible by 5. There are similar results for some other moduli, for example 7, 11, 25, 35, ...
4. The number of partitions of n with minimum difference two (that is $|a_i-a_j|>1$ for all i and j) equals the number of partitions of n in which each part has the form $5m+1$ or $5m+4$.

Most of these results are proved by deriving properties of the corresponding generating functions, as we did above for Euler's theorem. A typical result is: If $0<q<1$ then

$$\prod_{n=1}^{\infty}(1-q^n)=\sum_{m=-\infty}^{\infty}(-1)^m q^{m(3m-1)/2}.$$

This is called Euler's pentagonal number theorem. Using it we can give a recursive procedure for calculating $p(n)$, viz: If $n>0$, $p(0)=1$, and $p(m)=0$ if $m<0$, then

$$p(n)=p(n-1)+p(n-2)+\cdots$$
$$+(-1)^{m-1}p(n-k(3k-1)/2)$$
$$+(-1)^{m-1}p(n-k(3k-1)/2)+\cdots.$$

This provides a very efficient method for calculating $p(n)$.

The value of $p(n)$ increases rapidly with n, for instance $p(10)=42$, $p(20)=627$, and $p(30)=5604$. Hence, it is important to find an estimate of the approximate value of $p(n)$ for large n. One such estimate is

$$p(n)\sim\frac{1}{4n\sqrt{3}}e^{\pi\sqrt{2n/3}}.$$

A more accurate estimate as well as full details on all the topics mentioned in this section can be found in Andrews (1976).

VII. COMPUTATIONAL NUMBER THEORY

With the advent of modern computer systems, a new branch of number theory has developed over the past few decades which provides powerful algorithms for many aspects of the subject; Cohen (1993) gives a good introduction to the whole area. The most spectacular algorithms give very efficient methods for factorizing positive integers or for determining whether a given number is prime or not. Other algorithms include, for example, methods for finding the class group of an algebraic number field (that is, determining the ideal structure of the field, see Section I), efficient methods for dealing with large matrices, and methods for calculating the invariants associated with elliptic curves (for example the discriminant, see Section III). This work has a major bearing on cryptography, and so gives number theory a direct influence on modern commerce and public affairs. Many coding procedures used to keep the transmission of information secret are effective because it is very difficult to factorize a large positive integer, particularly if this integer has two or more large prime factors.

We shall illustrate the methods used in this theory by describing one modern procedure for factorizing a positive integer. It uses elliptic curves and was invented by H. Lenstra, further details can be found in Cohen (1993) or Rose (1999), among others. Before describing Lenstra's method, we need to consider an earlier method due to Pollard. When attempting to factorize an integer n, Pollard noted that in many cases, but by no means all, if a prime number p divides n, then $p-1$ has many small factors and hopefully all its prime factors are small. If all the prime factors of $p-1$ divide k, then $a^{p-1}-1$ divides a^k-1, Fermat's Theorem shows that p divides $a^{p-1}-1$, and hence p divides a^k-1 for $a=2, 3, \ldots$. It is usual to take k to be of the form $j!$ or $LCM\,(2, 3, \ldots, j)$ for some relatively small integer j. Therefore, for various choices of a and k we calculate (n, a^k-1), and if this gcd is larger than 1, then we have found a factor of n and we can continue the process until all factors have been located. Note that there are very efficient methods for calculating gcds, even if the integers involved are large.

Pollard's method fails when $p-1$ has a large prime factor. Lenstra introduced a similar method replacing the integers modulo a prime p by an elliptic curve C defined over a finite ring K. The expectation that $p-1$ has small factors in Pollard's method is replaced by the expectation that the order of points on the curve C defined over K has many small factors [if P is a point on $C(K)$, its order is the least positive integer t such that tP is the zero point, see Section III]. The ring K is the system of integers modulo n where n is the number to be factored. This integer is not prime (this is checked before the factorization algorithm is

applied because it is much easier to determine the primality status of n than it is to factorize it). Hence the ring K is not a field, that is division can fail. But this does not cause a problem because it is exactly at the point of failure that a factor of n appears. The method is as follows. We choose an elliptic curve C, a point P on C, and an integer k with many small factors, and then calculate the co-ordinates of the point kP modulo n. If this works then we have *failed* to find a factor of n, and so we repeat the process with a different C, or a different P, or a different k until we obtain a calculation that crashes. This particular calculation will have crashed because it involved a division by a factor m of n, this is equivalent to division by zero modulo n, but this is a *success* because we now have the factor m of n. Each single calculation is no more efficient than in Pollard's method. Lenstra's methods works well because (a) there is a great range of starting positions (that is, for curves C, points P and numbers k), and (b) each separate calculation can be performed extremely quickly. For example, this method was used to factorize $n = 2^{137} - 1$ on a standard issue desktop computer using the computer package PARIGP developed by H. Cohen *et al*., and in under 6 seconds it gave the answer

$$2^{137} - 1 = 32032215596496435569 \times 5439042183600204290159,$$

where the two integers on the right hand side are prime. Using traditional methods, this calculation would have taken many hours.

SEE ALSO THE FOLLOWING ARTICLES

ALGEBRA, ABSTRACT • APPROXIMATIONS AND EXPANSIONS • GROUP THEORY • NUMBER THEORY, ELEMENTARY

BIBLIOGRAPHY

Andrews, G. E. (1976). "The Theory of Partitions," Addison-Wesley, Reading, Massachusetts.

Baker, A. (1975). "Transcendental Number Theory," Cambridge Univ. Press, London.

Borevich, Z. I., and Shafarevich, I. R. (1966). "Number Theory" (translated from Russian by N. Greenleaf). Academic Press, New York.

Cassels, J. W. S. (1978). "Rational Quadratic Forms (LMS monograph 13)," Academic Press, London.

Cohen, H. (1993). "A Course in Computational Algebraic Number Theory," Springer, New York.

Cornell, G., Silverman, J. H., and Stevens, G. (1997). "Modular Forms and Fermat's Last Theorem," Springer, New York.

Edwards, H. M. (1977). "Fermat's Last Theorem," Springer, New York.

Hardy, G. H., and Wright, E. M. (1954). "An Introduction to the Theory of Numbers," 3rd ed. Clarendon Press, Oxford.

LeVeque, W. J. (1955). "Topics in Number Theory" (2 volumes). Addison-Wesley, Reading, Massachusetts.

Mordell, L. J. (1969). "Diophantine Equations," Academic Press, London.

Odlyzko, A. M. (1991). "The 10^{20}-th Zero of the Riemann Zeta Function and 70 Million of the Neighbors," AT&T Bell Labs., Murray Hill. N.J.

Ribenboim, P. (1979). "13 Lectures on Fermat's Last Theorem," Springer, New York.

Ribenboim, P. (1989). "The Book of Prime Number Records," Springer, New York.

Rose, H. E. (1999). "A Course in Number Theory," Revised 2nd ed. Clarendon press, Oxford.

Schmidt, W. M. (1980). "Diophantine Approximation," Lecture notes 785, Springer, Berlin.

Silverman, J. (1986). "The Arithmetic of Elliptic Curves," Springer, New York.

Singh, S. (1997). "Fermat's Last Theorem," Fourth Estate, London.

Stewart, I. N., and Tall, D. O. (1979). "Algebraic Number Theory," Chapman and Hall, London.

Titchmarsh, E. C. (1987). "Theory of the Riemann Zeta Function," 2nd ed. Clarendon Press. Oxford.

Vaughan, R. C. (1981). "The Hardy-Littlewood Method," Cambridge Univ. Press. London.

Wiles, A. (1995). "Modular elliptic Curves and Fermat's Last Theorem," *Ann. Math*. **141,** 443–551.



Number Theory, Elementary

Robert L. Page
University of Maine, Augusta

I. The Ancient Roots of Number Theory
II. Classical Problems and Results
III. Modern Directions
IV. Unsolved Problems and Conjectures

GLOSSARY

Absolute value Absolute value of a number n is given by $|n| = n$ if $n \geq 0$ and $|n| = -n$ if $n < 0$. Absolute value of a number is always nonnegative.

Composite number Natural number larger than 1 that is not prime. The first 10 composite numbers are 4, 6, 8, 9, 10, 12, 14, 15, 16, and 18. The number 1 is neither prime nor composite.

Definite integral $\int_a^b f(x)\,dx$ is the integral from a to b of the function f with respect to the variable x.

Divisor If a, b, and c are natural numbers with $a = bc$, b and c are divisors or factors of a. Since $12 = 1 \cdot 12 = 2 \cdot 6 = 3 \cdot 4$, it follows that 1, 2, 3, 4, 6, 12 are all factors or divisors of 12.

Greatest integer function $[x]$ is the largest integer whose value does not exceed x. We have $[-1.2] = -2$, $[0.5] = 0$, and $[6.8] = 6$.

***n* factorial** $n! = 1(2)(3)\cdots(n-1)(n)$. That is, $1! = 1$, $2! = 1(2) = 2$, $3! = 1(2)(3) = 6$, and $10! = 1(2)(3)\cdots(9)(10) = 3{,}628{,}800$.

Natural logarithm Natural logarithm of the number x is written in x and is the exponent t such that $e^t = x$, where $e = 2.71828\ldots$.

Natural number Member of the set $\{1, 2, 3, \ldots\}$.

Prime number Natural number larger than 1 whose only divisors are 1 and the number itself. The first 10 primes are 2, 3, 5, 7, 11, 13, 17, 19, 23, and 29.

Proper divisor Any divisor of a natural number except the number itself. Accordingly, 1, 2, 3, 4, and 6 are proper divisors of 12.

NUMBER THEORY is the study of the properties of integers. The numbers used to count, called the counting or natural numbers, are the set $\{1, 2, 3, 4, \ldots\}$. Closely related are the integers $\{\ldots, -3, -2, -1, 0, +1, +2, +3, \ldots\}$, which, along with the operations of addition and multiplication, are the foundation upon which arithmetic is constructed. From the integers, we may construct other systems of numbers, such as the rational numbers or the real numbers. However, it is the properties of the integers and the solution of problems that can be stated in terms of integers that form the subject matter of number theory, sometimes called the higher arithmetic.

I. THE ANCIENT ROOTS OF NUMBER THEORY

A. Arithmetica versus Logistica

As ancient cultures slowly developed the concept of number and applied it to the solution of everyday problems, it was inevitable that a few people would become interested in the properties of numbers, in relations among numbers and their patterns, and in numbers as logical entities rather than as practical tools.

The use of numbers in commerce and trade, construction of temples, surveying of land, and other practical pursuits came to be known as logistica. Such activities, which could be carried on by anyone with knowledge of the proper procedures, particularly commoners or slaves, was considered a less noble endeavor than was the study of abstract properties and relations among numbers, called arithmetica. The latter was considered the province of scholars and royalty, too precious to be entrusted to the common people. In ancient times, then, the goals of arithmetic were the same as those of number theory today: the study of the properties of integers.

B. Ancient Mathematical Records

Some records have survived through the centuries to shed light on the early history of mathematics. In Czechoslovakia in 1937, a wolf bone was found on which were carved 55 notches in groups of 5. The bone, which was about 30,000 years old, suggests that before mankind learned to write there was a compulsion to record numbers.

About 3000 B.C., a mace belonging to King Menes of Egypt was inscribed with hieroglyphic symbols representing 400,000 oxen, 422,000 goats, and 120,000 prisoners. On the Columna Rostrata in Rome, erected in honor of the victory over Carthage in 260 B.C., the symbol for 100,000 was engraved 31 times, signifying the number 3,100,000.

However, surviving records indicate that early mathematics consisted of more than just recording large numbers. Clay tablets from Babylonia containing columns of numbers, which were originally thought to be merely records of business accounts, were later discovered to be mathematical texts and tables. The tablets date from about 2000 to 200 B.C. Some contain solutions to construction problems such as the calculation of areas or volumes. Others relate to business or legal matters such as the computation of interest or the division of estates. Some were tables of multiplication or squares or square roots, whereas others gave the circumferences of circles of various diameters. A table of inverses was used to reduce the operation of division to the more easily performed operation of multiplication.

The tablet designated Plimpton 332 in the Plimpton Library at Columbia University contains columns of numbers that give the hypotenuse and one leg of a right triangle. The other leg can be calculated from the Pythagorean theorem. The sizes of some numbers imply that the Babylonians possessed a general method for solving the right triangle problem. The last column of this tablet gives the value of the cosecant of the acute angle opposite the longer leg of the triangle, indicating that the Babylonians used and tabulated some of the ratios of the sides of a right triangle, as we do today in trigonometry.

The Babylonians also considered problems of a theoretical nature such as circumscribing a circle about an isosceles triangle or finding the area of a circle, the latter leading to the erroneous value of 3 for π.

Other fertile sources of knowledge of ancient mathematics are the papyrus records from Egypt. One of the most notable, the Rhind Papyrus, is displayed in the British Museum. Dating from about 1650 B.C., the Rhind Papyrus contains material copied by the scribe Ahmes from earlier records. The material consists of problems involving arithmetic as well as ideas from geometry. Some problems involved no practical applications but seemed to be posed for the reader's amusement.

Early records such as these show that ancient civilizations were highly interested in mathematics, their achievement in mathematics was considerable, and the knowledge they possessed no doubt formed the basis for works of the Greeks and others who followed.

C. Number Theory versus Numerology

Just as the science of astronomy developed hand in hand with the superstition of astrology, so the early history of number theory is interwoven with numerology, the belief that numbers possessed mystical powers and were a dominant influence in human affairs. One form of numerology was gematria, whereby numbers were assigned to letters of the alphabet, for example, $a = 1$, $b = 2$, and so forth. The numerical value of words or names could then be calculated and studied, thus revealing hidden relationships or future events. A minor nobleman might ingratiate himself with the king by showing that the numerical values of their names, that is, the sum of the letter values, were equal.

Gematria could also warn the unwary of evil as in the Biblical quotation "Here is wisdom. Let him that hath understanding count the number of the beast; for it is the number of a man and his number is six hundred three score and six." To whom the passage referred has never been determined, but during the Reformation it was common to attack an enemy by writing the enemy's name in such a way that the numbers added to 666.

D. The Pythagoreans

One of the strongest influences in ancient number theory came from the Pythagoreans, a brotherhood founded by the philosopher Pythagoras who lived around 570–500 B.C. The members of this mystical order believed that integers were the key to explaining the universe. To support their belief, they had only to point to the fact, which could be verified by trial and error, that halving the length of a vibrating string created a sound an octave higher than the original tone. Further, they knew that when tones whose frequencies were in the ratios of certain whole numbers were sounded together, the result was a harmonious chord.

On such empirical evidence as this, the Pythagoreans based a theory of the universe ruled by integers. Numbers were imbued with human traits or characteristics. Odd numbers such as 1, 3, 5 were called masculine, whereas even numbers such as 2, 4, 6 were feminine. Square numbers such as $4 = 2(2)$ represented justice, the two factors being equal. The number 6 represented the soul, 7 represented health and understanding, and 8 was the number of love and friendship.

Although few today believe in numerology, we may speculate on the origins of some associations of numbers with philosophical concepts. The number 6 is the smallest number that is the sum of all of its proper divisors: $6 = 1 + 2 + 3$. This special property may have led to the number being identified with the important concept of the soul. Again, the Bible states that the earth was created in 6 days.

A few hints of numerology still remain in the ideas that certain numbers such as 7 are "lucky" and that numbers like 13 are "unlucky." Another example is the common expression "bad news comes in threes." Although such notions are now taken lightly, we must remember that ancient people took the principles of numerology seriously and that many purely mathematical questions and problems arose from such occult investigations.

E. Pythagorean Triples

The Pythagorean theorem states that the sum of the squares of the legs of a right triangle equals the square of its hypotenuse, that is, $a^2 + b^2 = c^2$, as shown in Fig. 1. This result was certainly known before the time of Pythagoras, but whether he was the first to actually prove the theorem is unknown because of the Pythagoreans' custom of ascribing all new knowledge to the Master.

Numbers whose values satisfy the Pythagorean theorem, such as 3, 4, and 5 ($3^2 + 4^2 = 9 + 16 = 25 = 5^2$), are permissible values for the sides of a right triangle. In Egypt, men known as rope stretchers made use of the 3, 4, 5 relationship to establish right angles in surveying property or constructing buildings by stretching a rope with knots marking lengths of 3, 4, and 5 units around three stakes, as in Fig. 2.

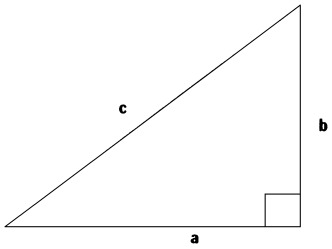


FIGURE 1 The Pythagorean theorem: $a^2 + b^2 = c^2$.

Trial and error leads to the discovery of many such Pythagorean triples, for instance, 5, 12, 13 and 8, 15, 17. It is natural to inquire whether formulas exist that will generate sets of Pythagorean triples. One method, which is attributed to Pythagoras, is as follows: Let n be any positive integer. Then $a = 2n + 1$, $b = 2n^2 + 2n$, and $c = 2n^2 + 2n + 1$ constitute a Pythagorean triple. If $n = 3$, we obtain 7, 24, 25, for which $7^2 + 24^2 = 49 + 576 = 625 = 25^2$. For $n = 4$ we have 9, 40, 41, whereby $9^2 + 40^2 = 81 + 1600 = 1681 = 41^2$. All triples generated in this way lead to triangles having a hypotenuse that is one unit longer than the larger leg.

A more general method consists of choosing integers u and v, one of which is odd and the other even, such that u and v have no common divisors other than the number 1. We then say that u and v are relatively prime. Then, letting $a = 2uv$, $b = u^2 - v^2$, and $c = u^2 + v^2$ gives the desired result. The following examples illustrate this method.

Multiplying each member of a Pythagorean triple by the same integer yields another Pythagorean triple.

u	v	a	b	c
2	1	4	3	5
4	1	8	15	17
8	5	80	39	89

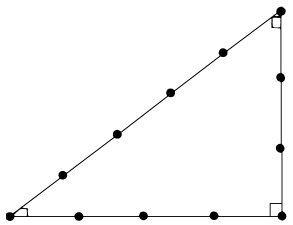


FIGURE 2 Rope stretching.

For example,

$$2(3, 4, 5) = 6, 8, 10,$$

and

$$6^2 + 8^2 = 36 + 64 = 100 = 10^2.$$

F. Prime Numbers

A subject of early investigation in number theory and one of continuing interest today is that of prime numbers. Primes are the building blocks from which all natural numbers are formed. The trial-and-error process for finding primes is a tedious and time-consuming one. It is not surprising that the search for methods or formulas for generating primes should have occupied a fundamental place in early number theory.

Eratosthenes (276–194 B.C.) devised a procedure for finding all primes less than or equal to some number N. The method, known as the Sieve of Eratosthenes, requires that the numbers from 2 to N be listed. All multiples of 2, except for 2 itself, are then crossed out, as are all multiples of 3, except for 3 itself, and so on.

2	3	~~4~~	5	~~6~~	7	~~8~~	~~9~~	~~10~~	11	~~12~~	13
~~14~~	~~15~~	~~16~~	17	~~18~~	19	~~20~~	~~21~~	~~22~~	23	~~24~~	~~25~~
~~26~~	~~27~~	~~28~~	29	~~30~~	31	~~32~~	~~33~~	~~34~~	~~35~~	~~36~~	37

After crossing out multiples of all primes whose values do not exceed $\sqrt{N}$, we have left the primes whose values do not exceed N. Clearly, any composite number N is divisible by a prime whose value is less than or equal to $\sqrt{N}$; otherwise, the product of its prime divisors would be greater than N. The sieve method can be carried out most effectively using computers. Various kinds of sieves have been applied to problems in number theory.

Euclid, who lived from about 330 to 275 B.C., is best known for the geometry contained in his great work, "The Elements." A considerable portion of this work is devoted to number theory. For example, Euclid proves that if a prime p divides the product ab, then either p divides a or p divides b, or both.

In the ninth volume of "The Elements," Euclid proved that the number of primes is infinite. To see this, suppose that only a finite number of primes exist, say $p_1, p_2, \ldots, p_n$. Then the integer $R = p_1 p_2 p_3 \ldots p_n + 1$ is either prime or composite. If R is prime, we have found a prime larger than p_n, contradicting our original assumption. If R is composite, it must be divisible by a prime p. But R is not divisible by any of the primes $p_1, p_2, p_3, \ldots, p_n$, since it leaves a remainder of 1 upon division by any of them. So p must be different from $p_1, p_2, p_3, \ldots, p_n$, again contradicting our assumption. Thus, the number of primes cannot be finite.

G. The Distribution of Primes

The knowledge that the number of primes is infinite leads to speculation about the distribution of primes throughout the set of natural numbers. In general, the density of primes is quite regular, decreasing slowly as we consider larger integers. The following table illustrates this pattern.

Range	Number of primes
2–1,000	168
1,001–2,000	135
2,001–3,000	127
3,001–4,000	120
4,001–5,000	119
⋮ ⋮	⋮
9,995,001–9,996,000	62
9,996,001–9,997,000	58
9,997,001–9,998,000	67
9,998,001–9,999,000	64
9,999,001–10,000,000	53

Despite this overall regularity in the distribution of primes, investigation of particular areas of the set of natural numbers reveals great diversity. Indeed, we can find arbitrarily long blocks of consecutive composite numbers. If $q = 2 \cdot 3 \cdot 5 \cdot \cdots \cdot p)$ is the product of all of the primes from 2 to p, then the natural numbers $q + 2$, $q + 3, q + 4, \ldots, q + p$ are all composite. But since there is no limit to the size of p, the result is established.

As an example, if we use the primes through 19, we find that 9,699,692 and the next 18 integers are all composite. Even more remarkable sequences of composite numbers are know. Thus, 370,261 is prime but the next 111 integers are composite.

H. The Fundamental Theorem of Arithmetic

Every composite number has a divisor other than 1 and the number itself. Such divisors are called proper divisors. A square number may have only one proper divisors. Thus, $49 = 7(7)$. On the other hand, a composite number may have many proper divisors, as in $36 = 2(18) = 3(12) = 4(9) = 6(6)$. Note that some of the proper divisors of 36, such as 2 and 3, are prime, whereas others, such as 4, 6, 9, 12, and 18, are themselves composite. If we continue to break the composite factors into proper divisors we obtain

$$36 = 2 \cdot 18 = 2 \cdot 2 \cdot 9 = 2 \cdot 2 \cdot 3 \cdot 3$$

$$3 \cdot 12 = 3 \cdot 3 \cdot 4 = 3 \cdot 3 \cdot 2 \cdot 2$$

$$4 \cdot 9 = 2 \cdot 2 \cdot 3 \cdot 3$$

$$6 \cdot 6 = 2 \cdot 3 \cdot 2 \cdot 3$$

Ultimately, the factorization yields $2^2 \cdot 3^2$. This illustrates the fact that if we agree to write the prime factors as powers of primes in increasing order, there is only one way to factor a given integer.

1. Fundamental Theorem of Arithmetic

Every natural number can be written uniquely as the product of primes.

Since 2 is the only even prime, for $p > 2$, consecutive primes must differ by two. The pairs of primes (3, 5), (5, 7), (11, 13), (17, 19), (29, 31), etc., are called twin primes. It has been conjectured, but never proven, that there exist an infinite number of pairs of twin primes.

Except for the triplet (3, 5, 7), not all of the numbers p, $p+2$, and $p+4$ can be prime since one of them must be divisible by 3. To see this, note that if p leaves the remainder 1 when divided by 3, $p+2$ is divisible by 3; whereas if p leaves the remainder 2, $p+4$ is divisible by 3. However, it is possible for p, $p+2$, and $p+6$ all to be prime, as in 5, 7, 11 or 17, 19, 23. Similarly, p, $p+4$ and $p+6$ may all be prime as shown by 13, 17, 19 or 37, 41, 43. It is conjectured, but as yet unproven, that there are infinitely many prime triples of each form.

Certain arithmetic sequences have been shown to contain an infinite number of primes. Thus, it is known that there are infinitely many primes of the form $4n-1$. Suppose there are only a finite number of them, say, $p_1 = 3, p_2, p_3, \ldots, p_k$. Forming $R = 4p_1p_2p_3 \cdots p_k - 1$, we see that R is not divisible by any of the primes $p_1, p_2, p_3, \ldots, p_k$. Now any integer must have one of the forms $4n, 4n+1, 4n+2$, or $4n-1$, and an odd prime must have either the form $4n-1$ or $4n+1$. The product of two numbers having the latter form also has the same form:

$$(4n_1+1)(4n_2+1) = 16n_1n_2 + 4(n_1+n_2) + 1 = 4n_3 + 1.$$

Hence if R is composite, at least one of its factors must be a prime of the form $4n-1$. But this is obviously not $p_1, p_2, p_3, \ldots, p_k$. So if R is composite, it has a factor of the form $4n-1$, which is different from the list that was assumed to include all primes of that form.

On the other hand, if R is prime, it has the form $4n-1$ and is different from $p_1, p_2, p_3, \ldots, p_k$. Either way, the assumption that the list includes all primes of the form $4n-1$ is contradicted and so the number of primes having that form cannot be finite.

In a similar fashion, we may show that there are infinitely many primes of the form $4n+1$, $6n+5$, and $8n+5$. All of these cases are included in Dirichlet's theorem, proved by Peter G. L. Dirichlet (1805–1859).

2. Dirichlet's Theorem

If a is positive and a and b are relatively prime, then there are infinitely many primes of the form $an+b$. The proof of this general result is much more difficult than the proof of specific cases.

Dirichlet's theorem stands out like an oasis in a desert wasteland. Attempts to find other functions that will generate an infinite number of primes have resulted in failure. It is not known whether such a simple sequence as n^2+1 contains an infinite number of primes.

I. Greatest Common Divisor

The basic operations of arithmetic are addition and multiplication, from which we derive the inverse operations of subtraction and division, respectively. If integers a and b are both divisible by an integer c, we say that c is a common divisor of a and b. Thus, 36 and 48 have the common divisor 2. It is easily seen that 3, 4, 6, and 12 are also common divisors of the two given numbers. The largest of these, 12, is called the greatest common divisor, or gcd, of 36 and 48, and we write $(36, 48) = 12$. Note that 2, 3, 4, and 6 all divide 12. This illustrates the theorem: If c is a common divisor of a and b, then c divides (a, b).

To show the plausibility of this theorem and to illustrate Euclid's algorithm for finding the gcd of two numbers, consider the numbers 42 and 135. We divide the larger of these two numbers by the smaller, expressing the result in the form

$$\text{dividend} = \text{quotient(divisor)} + \text{remainder}.$$

We have

$$135 = 3(42) + 9.$$

Each subsequent equation involves dividing the divisor of the previous equation by the remainder from that equation. The process stops when a zero remainder is obtained, as it must be, since the remainders form a decreasing sequence of nonnegative integers

$$\begin{aligned} 42 &= 4(9) + 6 \\ 9 &= 1(6) + 3 \\ 6 &= 2(3) + 0. \end{aligned}$$

From the last equation we see that 3 divides 6. From the next-to-the-last equation we see that 3 divides 9 since it divides the right-hand side of the equation. Similarly, 3 divides 42 and also 135. Hence, 3 is a common divisor of 42 and 135.

By rewriting the equations as

$$\begin{aligned} 135 - 3(42) &= 9 \\ 42 - 4(9) &= 6 \\ 9 - 1(6) &= 3, \end{aligned}$$

we can see that any common divisor of 135 and 42 divides 9. Also, any common divisor of 135 and 42 divides 6. Finally, we conclude that any common divisor of 135 and 42 divides 3. Hence, 3 is the gcd of 135 and 42. In general, gcd of the two numbers is always the last nonzero remainder that occurs in the algorithm.

Several results follow from Euclid's algorithm. If the product ab is divisible by c and if a and c are relatively prime, then c divides b. From this it may easily be seen that if a number is relatively prime to each of two or more numbers, it is relatively prime to their product. Also, if $(a, b) = c$, then $(ka, kb) = kc$. The latter may be verified by Euclid's algorithm. As an example, the gcd of $405 = 3(135)$ and $126 = 3(42)$ is $3(3) = 9$.

Euclid's algorithm may be used to prove the following: If c is the gcd of a and b, there exist integers x and y such that $ax + by = c$. For the preceding example, $5(135) - 16(42) = 3$. The numbers x and y can be obtained by solving for c, starting from the next-to-the-last equation and working backward to the first equation.

J. Least Common Multiple

If c is divisible by both a and b, we say that c is a common multiple of a and b. The multiples of 4 are 4, 8, 12, 16, 20, 24, etc., whereas those of 6 are 6, 12, 18, 24, etc. We find that 12 and 24 are both common multiples of 4 and 6.

The smallest of the infinite set of common multiples of a and b is called the least common multiple, or lcm, written $[a, b]$. Therefore, we have $[4, 6] = 12$.

It is easily seen that the lcm of two primes is their product. However, our previous example shows that the lcm of two numbers having a common divisor is smaller than their product. In fact, the product of two integers equals the product of their lcm and their gcd. That is, $ab = (a, b)[a, b]$.

Finding the gcd and the lcm for two numbers can be facilitated by writing each one as the product of primes. For example, $84 = 2^2(3)(7)$ and $90 = 2(3^2)(5)$. We find the gcd by taking each factor appearing in both numbers to the smallest power to which it appears in either number. So $(84, 90) = 2(3) = 6$.

We find the lcm by taking each factor appearing in either number to the largest power to which it appears in either number. So $[84, 90] = 2^2(3^2)(5)(7) = 1260$. Finally, we note that $6(1260) = 7560 = 84(90)$. The definitions of gcd and lcm can be extended to three or more numbers in a straightforward way.

K. Divisibility Rules

Before the advent of computers, finding divisors of large numbers was not an easy task. Therefore, certain divisibility rules were developed. Obviously, a number is divisible by 2 only when it is even, that is, only when it ends in 0, 2, 4, 6, or 8.

Because 100 and all higher powers of 10 are divisible by 4, a number is divisible by 4 only when it ends in 00 or when the number formed by its last two digits is divisible by 4. Only numbers ending in 0 or 5 are divisible by 5.

One hundred and all higher powers of 10 are divisible by 25. A number is divisible by 25 only when it ends in 00 or when the number formed by its last two digits is divisible by 25. Hence, a number is divisible by 25 only when it ends in 00, 25, 50, or 75.

The number 10 and all higher powers of 10 leave the remainder 1 when divided by 3. This means that the remainder from dividing a number by 3 equals the remainder from dividing the sum of the digits of that number by 3. We have $371 = 3(123) + 2$, whereas $3 + 7 + 1 = 11$ and $11 = 3(3) + 2$. Therefore, a number is divisible by 3 if and only if the sum of its digits is divisible by 3.

A similar statement holds for the divisor 9. A number is divisible by 9 if and only if the sum of its digits is divisible by 9. Thus, 12,681 is divisible by 9 since $1 + 2 + 6 + 8 + 1 = 18$ and 18 is divisible by 9. A process called casting out nines, based on the divisibility property of the number 9, can be used to check the results of addition or multiplication. For addition, the digits of each addend are added:

17	$1 + 7 = 8$		8
31	$3 + 1 = 4$		4
84	$8 + 4 = 12$	$1 + 2 = 3$	3
133			15
$1 + 3 + 3 = 7^*$			$1 + 5 = 6^*$

If the result is a two-digit number, these digits are added and so on until each addend has been reduced to a single-digit number. These resulting numbers for all of the addends are then added, the digits again being added if necessary to reduce the final result to a single-digit number.

The same process is applied to the sum, yielding a single-digit number. If these two single-digit numbers, the one from the addends and the one from the sum, are not equal, the addition is incorrect. On the other hand, if the two single-digit numbers are the same, there is no guarantee that the addition is correct. Two different errors may make these numbers the same but make the addition wrong. In particular, the error of transposing two digits will not be detected by this method. Hence, casting out nines can show that the addition process is wrong but cannot show that it is absolutely correct. The results of our example show that the original sum is incorrect.

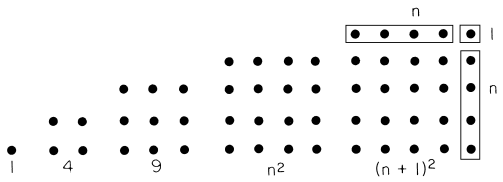


FIGURE 3 The square numbers.

For the product of two numbers the process is similar, except that the single-digit numbers of the factors are multiplied and the resulting product, reduced to a single-digit number if necessary, is compared with the single-digit number of the product. Let us check the product

$$34(28) = 942$$

$$3 + 4 = 7 \qquad 2 + 8 = 10 \qquad 9 + 4 + 2 = 15$$

$$7(10) = 70 \qquad 1 + 5 = 6^*$$

$$7 + 0 = 7^*$$

Since the two single-digit numbers are not equal, the multiplication is incorrect.

L. Figurate Numbers

It was inevitable that the Greeks would take an interest in the relationship between the two major area of mathematics, arithmetic and geometry. By associating n points with the number n and arranging the points in the shape of geometric figures, they were able to investigate relationships among the figurate or polygonal numbers. Consideration of the square numbers, Fig. 3, led to the discovery that

$$n^2 + 2n + 1 = (n + 1)^2,$$

as well as

$$1 + 3 + 5 + \cdots + 2n - 1 = n^2,$$

as shown in Fig. 4. Formulas such as these may be verified by the process of mathematical induction.

The triangular numbers are shown in Fig. 5. Consideration of the pattern in Fig. 6 leads to the relation

$$1 + 2 + 3 + \cdots + n = n(n + 1)/2.$$

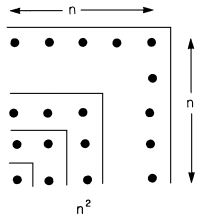


FIGURE 4 The sum of the first n odd positive integers is n^2.

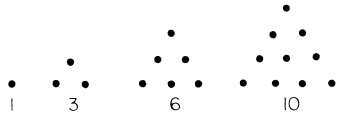


FIGURE 5 The triangular numbers.

Note that the triangular numbers arise from the pattern $1 + 2 + 3 + \cdots$, whereas the square numbers arise from $1 + 3 + 5 + 7 + \cdots$. The next logical pattern, $1 + 4 + 7 + 10 + \cdots$, gives the pentagonal numbers shown in Fig. 7.

Less obvious is the formula that gives the general expression for pentagonal numbers as

$$1 + 4 + 7 + 10 + \cdots + 3n - 2 = (3n^2 - n)/2.$$

The pattern $1 + 5 + 9 + 13 + \cdots$ yields the hexagonal numbers shown in Fig. 8. The general formula for the hexagonal numbers is

$$1 + 5 + 9 + 13 + \cdots + 4n - 3 = n(2n - 1).$$

Even more remarkable is the formula generating the nth number of the series of polygons having m sides:

$$\left(\frac{m-2}{2}\right)n^2 + \left(\frac{4-m}{2}\right)n.$$

M. Perfect Numbers

A perfect number is a natural number that is the sum of its proper divisors, such as $6 = 1 + 2 + 3$ or $28 = 1 + 2 + 4 + 7 + 14$. A general result concerning perfect numbers appears in Euclid's "The Elements."

Theorem. Let p be a prime. The number $N = 2^{p-1}(2^p - 1)$ is a perfect number when the factor $2^p - 1$ is a prime.

The first four perfect numbers generated by this formula are

$$(2^{2-1})(2^2 - 1) = 2^1(3) = 6$$

$$(2^{3-1})(2^3 - 1) = 2^2(7) = 28$$

$$(2^{5-1})(2^5 - 1) = 2^4(31) = 16(31) = 496$$

$$(2^{7-1})(2^7 - 1) = 2^6(128 - 1) = 64(127) = 8128$$

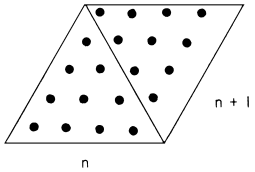


FIGURE 6 The sum of the first n positive integers is $n(n+1)/2$.

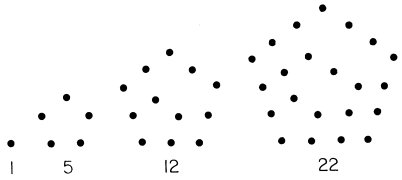


FIGURE 7 The pentagonal numbers.

However, for $p = 11$, we have $2^{11-1}(2^{11} - 1) = 2^{10}(2047)$, which is not a perfect number since $2047 = 23(89)$ is composite.

Nearly 2000 years after the appearance of this theorem, Leonhard Euler (1707–1783) proved that every even perfect number has the form given by Euclid. As a consequence, every even perfect number ends in 6 or 8. There is known to exist a perfect number that requires 54 digits to write in the base 10 system.

The existence of odd perfect numbers is as yet an unresolved question. None has ever been found and it has been shown that none can exist that is smaller than 10^{200}. This is, of course, not the same as showing that no odd perfect numbers exist.

If the sum of the proper divisors of a number is a multiple of the number itself, we call that number a multiply perfect number. For example, the divisors of 120 have a sum of 240. We call 120 a doubly perfect number or a perfect number of class 2. Other multiply perfect numbers of class 2 are 672 and 523, 776. More than 300 multiply perfect numbers have been found, including some of class 7.

Numbers for which the sum of the proper divisors is less than the number itself are called deficient numbers, the first few being 2, 3, 4, 5, 7, 8, 9, 10, and 11. Numbers for which the sum of the proper divisors is greater than the number itself are called abundant numbers, of which the first few are 12, 18, 20, 24, 30, and 36. The smallest odd abundant number is 945.

Since a prime has only the number 1 as a proper divisor, it is obvious that all primes are deficient numbers as are powers of primes. Furthermore, all divisors of a perfect number or a deficient number are deficient, whereas all multiples of a perfect number or an abundant number are abundant.

As one example of the role of abundant and deficient numbers in numerology, in the eighth century, Alcuin noted that the original Creation was accomplished in 6 days, 6 being a perfect number. However, the human race was supposed to be descended from 8 persons in the Ark, 8 being a deficient number. This implied that God's act was perfect and the latter occurrence imperfect.

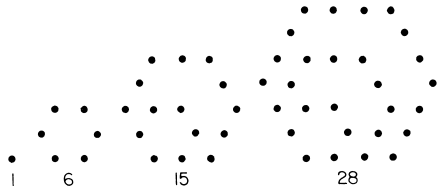


FIGURE 8 The hexagonal numbers.

Two numbers a and b are called amicable numbers if the sum of the proper divisors of a equals b and the sum of the proper divisors of b equals a. For example, the sum of the proper divisors of 220 is 284 and the sum of the proper divisors of 284 is 220. Thus, 220 and 284 form an amicable pair.

Numerologists believed that a couple could strengthen their bond of love or friendship by inscribing each of a pair of keepsakes with one of two amicable numbers, with each person retaining one of the keepsakes. If the names of the two friends could be written in such a way that the number of one name was 220 and the number of the other name was 284, the relationship would be even stronger. Other pairs of amicable numbers are 1184 and 1210 as well as 17,296 and 18,416. Nearly 400 amicable pairs have been discovered, including a pair for which each number requires 50 digits to write in the base 10 system.

N. Diophantine Equations

Diophantus lived about the third or fourth century A.D. His work, "Arithmetic," deals with the solution of certain algebraic problems in which a symbol is used to represent the unknown quantity, just as we do today. Although we do not know his dates, a problem from the "Palatine Anthology," which purports to describe the tomb of Diophantus, suggests a way of finding his age at death. The description says that God granted him the sixth part of his life for his youth, that after a twelfth part his cheeks were bearded, after a seventh part he married, a son was born in the fifth year of his marriage, the child was "half of his father," and that Diophantus grieved the remaining 4 years of his life. Interpreting "half of his father" to mean that the child's age at death was half of the age at which the father died, we have

$$X/6 + X/12 + X/7 + 5 + X/2 + 4 = X,$$

where X is Diophantus' age upon his death. Solving the equation gives $X = 84$.

In another problem, the reader is asked to find two numbers whose sum is 20 and the sum of whose squares is 208. Instead of letting one number be x and the other $20 - x$, as a modern student of algebra would do, Diophantus called the required numbers $10 + x$ and $10 - x$. Then the conditions of the problems require that $(10 + x)^2 + (10 - x)^2 = 208$, where $x = 2$ and the required numbers are 8 and 12. Diophantus' selection of $10 + x$ and $10 - x$ to represent the two numbers leads to a simpler equation to solve than does the modern approach.

Not all of the problems given by Diophantus yield integral solutions. One requires that 13, which equals the sum of the squares 4 and 9, be written as the sum of two other squares. The answer is $\frac{1}{25} + \frac{324}{25}$ or $(\frac{1}{5})^2 + (\frac{18}{5})^2$.

Neither did all of the problems have unique solutions. For example, a tax collector must collect 2 gold pieces for each man, 1 gold piece for each woman, and $\frac{1}{2}$ gold piece for each child in a certain town. If he collects 100 gold pieces from 100 residents, how many men, women, and children live in the town? The conditions lead to the equations:

$$2x + y + \tfrac{1}{2}z = 100,$$

and

$$x + y + z = 100,$$

where x, y, and z represent the number of men, women, and children, respectively. Subtracting the second equation from the first, we find

$$x - \tfrac{1}{2}z = 0 \qquad \text{or} \qquad z = 2x.$$

Thus, we know that there are twice as many children as men. (This is as far as mathematics will take us because we have two equations containing three unknowns.) We can choose to let x be any integer less than or equal to 33 and determine y and z. Some of the solutions are

x	$z = 2x$	$y = 100 - x - z$
10	20	70
20	40	40
25	50	25
33	66	1

Such a problem with multiple solutions is called an indeterminate problem. Problems whose solutions are required to be rational numbers or integers came to be known as Diophantine equations. Although Diophantus frequently gave rational as well as integral solutions to his problems, in present day terminology, the word Diophantine usually refers to problems requiring integral solutions.

O. Magic Squares

If a square is divided into n^2 smaller squares and one of the integers from 1 to n^2 is written in each of the smaller squares (using each number only once) in such a way that the sum of the integers in any row, column, or main diagonal is always the same, the result is called a magic square of order n. A few examples of magic squares are shown in Fig. 9.

Interest in magic squares dates back to about 2200 B.C. in China where the diagram of Fig. 10, called the Lo-Shu, appeared.

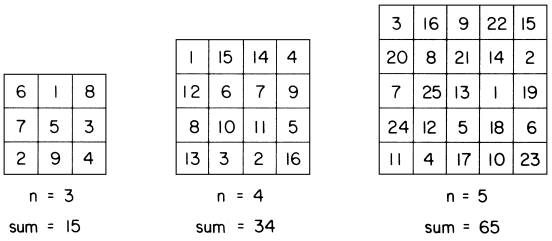


FIGURE 9 Magic squares of order 3, 4, and 5.

The sum of the integers in any row, column, or main diagonal of a magic square of order n, constructed from the integers $1, 2, 3, \ldots, n^2$, is $n(n^2 + 1)/2$. Such a magic square is said to be primitive. Many different rules exist for constructing magic squares, with some rules depending on whether n is odd or even.

It is easily seen that no magic squares exist of order 2. Also, any set of n^2 consecutive integers may be substituted for the integers in a magic square of order n as long as the original pattern for integers of increasing size is followed. This is, of course, equivalent to adding a constant to each number of a given magic square. Furthermore, the requirement that the integers be consecutive may be weakened to require sequences of integers that differ by a fixed amount.

For example, if we add 11 to each number in the magic square of order 3 given above, we have the result shown in Fig. 11, where the sum is $15 + 3(11) = 48$. If we replace the original integers by the sequence $20, 25, 30, \ldots$, we obtain the magic square of Fig. 12, for which the sum is 120. Certain rows or columns of a magic square can be interchanged without destroying the equal sum property.

While many purely mathematical properties have been discovered concerning magic squares, they were originally of interest because of their supposed power to ensure good fortune or to ward off evil influences. Magic squares engraved on precious metal were worn as amulets.

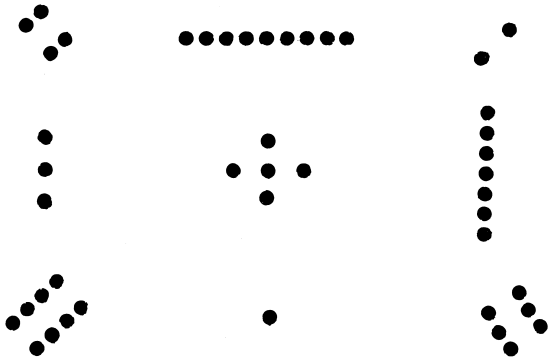


FIGURE 10 The Lo-Shu.

17	12	19
18	16	14
13	20	15

FIGURE 11 Magic square formed by the addition of a constant.

A magic square, when engraved on a plate of silver, was believed to protect the owner from the plague. The magic square of order 4, shown in Fig. 9, appears in Dürer's engraving Melancholia.

What must rank near the top of any list of incredible coincidences was discovered by T. E. Lobeck. Starting with the magic square of order 5, shown in Fig. 13, he substituted the nth digit in the decimal expansion of $\pi = 3.14159\ldots$ for the integer n in the magic square. The result appears in Fig. 14. The sum of each column also occurs as the sum of a row. Perhaps more amazing is the fact that the sum of the two main diagonals is $38 + 27 = 65$, which is the sum of the numbers in any row, column, or main diagonal of the original magic square.

II. CLASSICAL PROBLEMS AND RESULTS

Certain themes in number theory that had their origins in antiquity reappeared during the Middle Ages or the Renaissance. Indeed, some of them have continued to be the objects of research to this day. In this section, we examine some of the major results that have been achieved in number theory since the days of Greek mathematics, regardless of whether those results concern problems of ancient derivation or more recent origin.

A. Indeterminate Problems

We start with a simple problem. A barnyard contains chickens and goats. Altogether there are 72 legs in the barnyard. How many chickens and how many goats are there? The conditions result in the equation $4x + 2y = 72$, where x is the number of goats and y is the number of chickens. Obviously, there are a multitude of solutions for this equation, some of which we may discover by solving for one variable in terms of the other, $y = (72 - 4x)/2$:

45	20	55
50	40	30
25	60	35

FIGURE 12 Magic square formed by substitution of numbers in a sequence.

17	24	1	8	15
23	5	7	14	16
4	6	13	20	22
10	12	19	21	3
11	18	25	2	9

FIGURE 13 A magic square of order 5.

x	y
10	16
15	6
-5	46
$\sqrt{2}$	$-2\sqrt{2} + 36$ etc.
$\vdots$	$\vdots$

The realities of the world preclude certain solutions such as those involving negative or irrational numbers. Even with these restrictions, there are 19 integral solutions corresponding to $x = 0, 1, 2, \ldots, 18$.

Of course, the number of variables is not restricted to 2. Consider the following problem: Thirty people enter a movie theater, paying a total of \$50. If men pay \$3 each, women \$2 each, and children \$1 each, how many men, how many women, and how many children make up the party?

If we let x equal the number of men, y the number of women, and z the number of children, we have

$$x + y + z = 30$$

and

$$3x + 2y + z = 50.$$

Subtracting the first equation from the second yields

$$2x + y = 20 \qquad \text{or} \qquad y = 20 - 2x.$$

Letting x take on the values $0, 1, 2, \ldots, 10$ gives these values for y: $20, 18, 16, \ldots, 0$. Then either equation gives these values for z: $10, 11, 12, \ldots, 20$. There are eleven solutions consisting of the triples (0, 20, 10), (1, 18, 11), (2, 16, 12), . . . , (10, 0, 20).

2	4	3	6	9	24
6	5	2	7	3	23
1	9	9	4	2	25
3	8	8	6	4	29
5	3	3	1	5	17
17	29	25	24	23	

FIGURE 14 The result of substituting the digits of π into the magic square of Fig. 13.

Sometimes, additional conditions restrict the number of solutions. In this problem, one might require that the number of women be twice the number of men, yielding the unique solution (5, 10, 15). At other times there may be an infinite number of solutions or even no solution. One renowned indeterminate problem, called the cattle problem of Archimedes, results in seven equations in eight unknowns whose solution yields extremely large numbers.

Many problems of this type, called linear indeterminate problems, originated in India. The Hindu mathematician Brahmagupta (born in 598 A.D.) authored a treatise on astronomy that included chapters devoted to mathematics. This work, "Brahma-Sphuta-Siddhanta" (or "Brahma's Correct System") includes solutions of numerous linear indeterminate equations. Also found there is the second-degree indeterminate equation $nx^2 + 1 = y^2$, for which Brahmagupta gives the solution

$$x = 2t/(t^2 - n)$$
$$y = (t^2 + n)/(t^2 - n),$$

where t is any integer. Thus, if $n = 3$, we have $3x^2 + 1 = y^2$ and $x = 2t/(t^2 - 3)$ and $y = (t^2 + 3)/(t^2 - 3)$, which leads to

t	x	y
1	-1	-2
2	4	7
3	1	2
10	$\frac{20}{97}$	$\frac{103}{97}$
$\vdots$	$\vdots$	$\vdots$

He also states that the equation $nx^2 - 1 = y^2$ has no integral solutions for x and y unless n is the sum of the squares of two integers. For example, $4x^2 - 1 = y^2$ has no integral solutions, whereas $13x^2 - 1 = y^2$ has integral solutions because $13 = 2^2 + 3^2$. One solution is $x = 5$, $y = 18$.

The search for general solutions was aided by the introduction of symbols for certain quantities and operations that occur frequently within a given problem. The use of this technique is credited to Diophantus. For a long period of time, mathematicians did not differentiate between problems leading to determinate or indeterminate solutions. Today, algebra students are exposed almost exclusively to determinate problems.

The following results are useful in solving certain indeterminate problems.

Theorem. If the integers a and b are relatively prime then there exist integers x and y such that $ax + by = 1$. This follows directly from Euclid's algorithm and leads to

Theorem. There exist integers x and y satisfying the equation $ax + by = c$ if and only if the gcd of a and b divides c.

For example, in the equation $14x + 35y = 56$, we find that $(14, 35) = 7$ and 7 divides 56. Dividing, we obtain the equivalent equation $2x + 5y = 8$. We now let $2x + 5y = 1$ and find the solution $x = -2$, $y = 1$. Then $x = 8(-2) = -16$ and $y = 8(1) = 8$ are solutions of $2x + 5y = 8$, and therefore of $14x + 35y = 56$. From the particular solution x_0, y_0, we can find the general solution $x = x_0 + tb$ and $y = y_0 - ta$, where t is any integer. In our case, $x = -16 + 35t$ and $y = 8 - 14t$.

Diophantine equations involving variables to the second or higher powers can prove difficult or impossible to solve. Although many special results exist, a general method of solving any Diophantine equation or proving the nonexistence of solutions is unknown. We mention a few particular results. It is known that the equations $x^3 + y^3 = z^3$ and $x^4 + y^4 = z^4$ have no positive integral solutions.

The equation $x^4 + y^4 = u^4 + v^4$ has general solutions that we shall not list. The smallest integral solution is $133^4 + 134^4 = 158^4 + 59^4$.

It has been proven that the equation $ax^n + by^n = c$ has only finitely many solutions if $n \geq 3$. More generally, Axel Thue (1863–1922) showed that if a function of x and y,

$$f(x, y) = a_n x^n + a_{n-1} x^{n-1} y + \cdots + a_1 x y^{n-1} + a_0 y^n,$$

with a_i integers, cannot be factored into two polynomials having integral coefficients, then the equation $f(x, y) = c$ has only a finite number of solutions for $n \geq 3$.

The equation $x^2 - cy^2 = 1$ is known as Pell's equation, after John Pell (1610–1685), although he had no part in its solution. For any value of c there is the trivial solution $x = \pm 1$, $y = 0$. If c is a square number, the left-hand side is easily factored. If c is a positive nonsquare integer, then it can be proved that the equation always has a nontrivial solution. Such solutions may not necessarily be obtained readily by trial and error. The equation $x^2 - 61y^2 = 1$ has $x = 1{,}766{,}319{,}049$ and $y = 226{,}153{,}980$ as its smallest positive nontrivial solution.

The foregoing discussion is meant to illustrate the wide range of problems leading to Diophantine equations, the great variety of approaches to solving such problems and the enormous difficulty that is entailed in the solution of certain of them.

B. Congruences

The theory of congruences was begun by the work of Carl Friedrich Gauss (1777–1855), which originally appeared

in his "Disquisitiones Arithmeticae" in 1801. We say that the two integers a and b are congruent for the modulus m, and write $a \equiv b(\text{mod } m)$ if their difference $a - b$ is divisible by the integer m. For example, $21 \equiv 6 \pmod 5$, since $21 - 6 = 15$ is divisible by 5. Similarly, $19 \equiv 54(\text{mod } 7)$ since $19 - 54 = -35$ is divisible by 7. However, as is easily verified, 18 is not congruent to 11(mod 3) and we write $18 \not\equiv 11(\text{mod } 3)$.

Since any integer n leaves a remainder of $0, 1, 2, \ldots, m-1$ when it is divided by m, we see that the integers can be partitioned into m classes according to the remainder obtained upon division by m. They are called the residue classes (mod m). Thus, the residue classes (mod 5) consist of

$$\begin{array}{cccccccc} \ldots, & -10, & -5, & 0, & 5, & 10, & \ldots \\ \ldots, & -9, & -4, & 1, & 6, & 11, & \ldots \\ \ldots, & -8, & -3, & 2, & 7, & 12, & \ldots \\ \ldots, & -7, & -2, & 3, & 8, & 13, & \ldots \\ \ldots, & -6, & -1, & 4, & 9, & 14, & \ldots . \end{array}$$

Obviously, any two integers within a given residue class are congruent (mod m).

Certain properties of congruences follow immediately from the definition. Thus, if $a \equiv b(\text{mod } m)$ and $c \equiv d(\text{mod } m)$, then

(1) $(a + c) \equiv (b + d)(\text{mod } m)$,
(2) $ac \equiv bd(\text{mod } m)$,
(3) $a^n \equiv b^n(\text{mod } m)$, and
(4) $a \equiv b(\text{mod } n)$ where n is a divisor of m.

Furthermore, if $a \equiv b(\text{mod } m_1)$ and $a \equiv b(\text{mod } m_2)$, then $a \equiv b(\text{mod } M)$, where M is the lcm of m_1 and m_2. So, from the congruences $17 \equiv 377(\text{mod } 24)$ and $17 \equiv 377(\text{mod } 40)$, we conclude that $17 \equiv 377(\text{mod } 360)$, where $[24, 40] = 360$.

Finally, if $ac \equiv bc(\text{mod } m)$, then $a \equiv b(\text{mod } m/d)$, where d is the gcd of c and m. Thus, when the congruence $48 \equiv 18(\text{mod } 15)$ is written as $8(6) \equiv 3(6)(\text{mod } 15)$, we see that $(6, 15) = 3$. Hence, $8 \equiv 3(\text{mod } 5)$.

To illustrate the utility of calculating with congruences, consider the problem of finding the smallest positive remainder obtained when 15^{19} is divided by 23. We have

$$\begin{aligned} 15 &\equiv -8(\text{mod } 23) \\ 15^2 &\equiv 64(\text{mod } 23) \equiv -5(\text{mod } 23) \\ 15^3 &\equiv 40(\text{mod } 23) \equiv -6(\text{mod } 23) \\ 15^4 &\equiv 25(\text{mod } 23) \equiv 2(\text{mod } 23) \\ 15^{16} &\equiv (15^4)^4(\text{mod } 23) \equiv 2^4(\text{mod } 23) \\ &\equiv 16(\text{mod } 23). \end{aligned}$$

Then,

$$\begin{aligned} 15^{19} &\equiv 15^{16}(15^3)(\text{mod } 23) \equiv 16(-6)(\text{mod } 23) \\ &\equiv -96(\text{mod } 23) \equiv 19(\text{mod } 23). \end{aligned}$$

Hence, 15^{19} leaves a remainder of 19 when divided by 23.

C. Linear Congruences

Just as in the case of linear equations, linear congruences involve the first power of a variable. Whereas linear equations yield only a single solution, linear congruences may give an unlimited number of solutions. The linear congruence $2x \equiv 6(\text{mod } 12)$ has the solution $x = 3$ as well as any integer that is congruent to 3(mod 12).

Of greater interest, however, is the question of whether two or more noncongruent solutions exists. Trial and error will show that our previous example also has the solution $x \equiv 9(\text{mod } 12)$. Also, unlike linear equations, linear congruences may have no solution. In general, $ax \equiv b(\text{mod } m)$ has solutions only when the greatest common divisor of a and m also divides b. In this case, there are exactly (a, m) noncongruent solutions.

Again, the congruence $18x \equiv 24(\text{mod } 30)$ has $(18, 30) = 6$ solutions given by the numbers 3, 8, 13, 18, 23, and 28. Note that all solutions after the first are obtained by successively adding $\frac{30}{6} = 5$ to the previous solution. On the other hand, the congruence $15x \equiv 12(\text{mod } 35)$ has no solution because $(15, 35) = 5$ does not divide 12.

Two or more linear congruences may be solved simultaneously. By listing the solutions of the two congruences $x \equiv 5(\text{mod } 7)$ and $x \equiv 4(\text{mod } 11)$, which are $12, 19, 26, 33, \ldots$ and $15, 26, 37, \ldots$, respectively, we see that 26 is a common solution. Then $x \equiv 26(\text{mod } 7 \cdot 11)$ or $x \equiv 26(\text{mod } 77)$ is the simultaneous solution.

D. Chinese Remainder Theorem

Many ancient problems involved linear congruences. Such a problem concerned the removal of eggs from a basket 2, 3, 4, 5, and 6 at a time, whereupon 1 egg remained. However, when they were removed 7 at a time, none remained. This may be stated in terms of congruences as follows: $x \equiv 1(\text{mod } 2)$, $x \equiv 1(\text{mod } 3), \ldots, x \equiv 1(\text{mod } 6)$, $x \equiv 0(\text{mod } 7)$. These are equivalent to the two congruences $x \equiv 1(\text{mod } 60)$ and $x \equiv 0(\text{mod } 7)$, which have the solution $x \equiv 301(\text{mod } 420)$.

Problems of this type are found in the works of the Chinese mathematician Sun-Tse. When each pair of moduli is relatively prime, the solution is found in the following theorem.

Chinese Remainder Theorem. Given the simultaneous congruences $x \equiv a_i(\text{mod } m_i)$ for $i = 1, 2, 3, \ldots, n$,

where $(m_i, m_j) = 1$ for all i, j, let $M = m_1 m_2 m_3 \cdots m_n$. If for each i the congruence $b_i(M/m_i) \equiv 1 (\text{mod } M/m_i)$ is solved, then the original set of congruences has the solution

$$x \equiv [a_1 b_1 (M/m_1) + a_2 b_2 (M/m_2) + a_3 b_3 (M/m_3) + \cdots + a_n b_n (M/m_n)] (\text{mod } M).$$

We illustrate the use of the Chinese remainder theorem with the following problem. A shipwrecked sailor passes the time of day by counting the coconuts he has gathered. When he counts by threes, there are 2 coconuts left over. When he counts by fives, there are 4 left over, and when he counts by sevens, there are 5 left over. How many coconuts has the sailor gathered?

The conditions of the problem lead to the congruences $x \equiv 2 (\text{mod } 3)$, $x \equiv 4 (\text{mod } 5)$, and $x \equiv 5 (\text{mod } 7)$; whence $a_1 = 2$, $m_1 = 3$, $a_2 = 4$, $m_2 = 5$, $a_3 = 5$, and $m_3 = 7$. Then $M = 3(5)(7) = 105$ and $M/m_1 = 35$, and $M/m_2 = 21$, and $M/m_3 = 15$. The new set of congruences is $35b_1 \equiv 1 (\text{mod } 3)$, $21b_2 \equiv 1 (\text{mod } 5)$ and $15b_3 \equiv 1 (\text{mod } 7)$. The solutions of the latter are $b_1 = 2$, $b_2 = 1$, and $b_3 = 1$. Thus,

$$\begin{aligned} x &\equiv [2(2)(35) + 4(1)(21) + 5(1)(15)] (\text{mod } 105) \\ &\equiv 299 (\text{mod } 105) \equiv 89 (\text{mod } 105) \end{aligned}$$

is the solution of the original set of congruences.

Other theorems of interest that involve congruences include one due to Pierre de Fermat (1601–1665) called Fermat's Little Theorem. If p is a prime number that does not divide the number a, then $a^{p-1} \equiv 1 (\text{mod } p)$.

For instance, if $p = 5$ and $a = 8$, we have $8^{5-1} = 8^4 = 4096 \equiv 1 (\text{mod } 5)$.

Fermat's little theorem is a special case of a theorem by Euler. For a positive integer n, the number of integers from 1 to n that are relatively prime to n is indicated by $\phi(n)$. Thus, $\phi(8) = 4$ since 8 is relatively prime to 1, 3, 5, and 7. Also, $\phi(11) = 10$ because 11 is relatively prime to the integers from 1 to 10. Note that for any prime p, $\phi(p) = p - 1$. We state

Euler's Theorem. If a and m are relatively prime, then $a^{\phi(m)} \equiv 1 (\text{mod } m)$.

Thus, if $a = 9$ and $m = 8$, we have $9^4 = 6561 \equiv 1 (\text{mod } 8)$, which is easily verified.

E. Wilson's Theorem

We next consider

Wilson's Theorem. The integer p is a prime if and only if $(p-1)! \equiv -1 (\text{mod } p)$.

For example, if $p = 7$, then

$$(7-1)! = 6! = 720 \equiv -1 (\text{mod } 7).$$

However, for $p = 6$,

$$(6-1)! = 5! = 120 \not\equiv -1 (\text{mod } 6).$$

Wilson's theorem is credited to John Wilson (1741–1793), although it was first proved by Joseph Louis Lagrange (1736–1813). It is noteworthy because it gives a necessary and sufficient condition for an integer to be prime. In practice, however, it is of limited use in determining the primality of integers, because the factorial expression becomes so large that its computation quickly exceeds all reasonable time limits, even on the most modern computers.

A complementary result is the following theorem.

Theorem. If n is a composite integer different from 4, then $(n-1)! \equiv 0 (\text{mod } n)$.

The fact that 1255 is composite guarantees that $(1254)! \equiv 0 (\text{mod } 1255)$.

We now examine some results for congruences involving a variable to the second or higher power. When the highest power of the variable is 2, we call the congruence a quadratic congruence, analagous to a quadratic equation. The following theorems, which are consequences of Wilson's theorem, give explicit solutions for certain quadratic congruences.

Theorem. If p is a prime having the form $4n + 1$, then the roots of the congruence $x^2 \equiv -1 (\text{mod } p)$ are

$$x \equiv \pm \left(\frac{p-1}{2} \right)! (\text{mod } p).$$

To illustrate, let $p = 13 = 4(3) + 1$. Then the congruence $x^2 \equiv -1 (\text{mod } 13)$ has the solutions

$$x \equiv \left(\frac{13-1}{2} \right)! (\text{mod } 13),$$

or

$$x \equiv 6! (\text{mod } 13) \equiv 5 (\text{mod } 13),$$

which we illustrate by

$$5^2 = 25 \equiv -1 (\text{mod } 13)$$

and

$$x \equiv -6! (\text{mod } 13) \equiv 8 (\text{mod } 13),$$

which we illustrate by

$$8^2 = 64 \equiv -1 (\text{mod } 13).$$

Any odd prime must be of the form $4n + 1$, considered previously, or of the form $4n + 3$, for which the following holds.

Theorem. If p is a prime of the form $4n+3$, then one of the congruences

$$\left(\frac{p-1}{2}\right)! \equiv \pm 1 (\text{mod } p)$$

holds. As an example, we see that when $n=3$,

$$\left(\frac{3-1}{2}\right)! = 1! = 1 \equiv 1 (\text{mod } 3)$$

and when $n=7$,

$$\left(\frac{7-1}{2}\right)! = 3! = 6 \equiv -1 (\text{mod } 7).$$

There is a complex procedure for determining whether the positive or negative sign prevails.

A more general result is contained in the following theorem.

Theorem. If p is an odd prime that does not divide the integer a, then $x^2 \equiv a (\text{mod } p)$ has a solution when $a^{(p-1)/2} \equiv 1 (\text{mod } p)$ and has no solution when $a^{(p-1)/2} \equiv -1 (\text{mod } p)$.

Let $p=7$ and $a=8$. Then $8^{(7-1)/2} = 8^3 = 512 \equiv 1 (\text{mod } 7)$, so the congruence $x^2 \equiv 8 (\text{mod } 7)$ has a solution. One easily found solution is $x \equiv 1 (\text{mod } 7)$.

On the other hand, if $p=5$ and $a=8$, then $8^{(5-1)/2} = 8^2 = 64 \equiv -1 (\text{mod } 5)$. Hence, the congruence $x^2 \equiv 8 (\text{mod } 5)$ has no solutions.

By trial and error we find that the congruence $x^2 + 5x \equiv 0 (\text{mod } 6)$ has the four solutions $x \equiv 0 (\text{mod } 6)$, $x \equiv 1 (\text{mod } 6)$, $x \equiv 3 (\text{mod } 6)$, and $x \equiv 4 (\text{mod } 6)$. However, if the modulus is prime, a polynomial congruence of degree n cannot have more than n solutions. It may have fewer than n. For example, $x^3 \equiv 1 (\text{mod } 5)$ has only the solution $x \equiv 1 (\text{mod } 5)$.

In general, in order to solve a polynomial congruence, one has only to solve polynomial congruences having moduli that are powers of a prime. We illustrate the method by considering the congruence

$$x^3 - 2x^2 + 12 \equiv 0 (\text{mod } 44).$$

Because $44 = 2^2 \cdot 11$, we write two new congruences

$$x^3 - 2x^2 + 12 \equiv 0 (\text{mod } 2^2)$$

and

$$x^3 - 2x^2 + 12 \equiv 0 (\text{mod } 11).$$

The first has the solutions $x \equiv 0 (\text{mod } 4)$ and $x \equiv 2 (\text{mod } 4)$, whereas the second has the solutions $x \equiv 1 (\text{mod } 11)$, $x \equiv 4 (\text{mod } 11)$, and $x \equiv 8 (\text{mod } 11)$. By applying the Chinese remainder theorem to pairs of these latter congruences, each pair containing a congruence of modulus 4 and a congruence of modulus 11, we find the solutions of the original congruence to be

$$\begin{aligned} &x \equiv 4 (\text{mod } 44), && x \equiv 8 (\text{mod } 44) \\ &x \equiv 12 (\text{mod } 44), && x \equiv 26 (\text{mod } 44) \\ &x \equiv 30 (\text{mod } 44), && x \equiv 34 (\text{mod } 44). \end{aligned}$$

F. Quadratic Residues

We now examine the law of quadratic reciprocity, one of the most profound and powerful results in the theory of congruences. If the integers a and m are relatively prime, we say that a is a quadratic residue of m if the congruence $x^2 \equiv a (\text{mod } m)$ has a solution. If it has no solution, we say that a is a quadratic nonresidue of m. Since $x^2 \equiv 11 (\text{mod } 5)$ has the solution $x \equiv 1 (\text{mod } 5)$, we see that 11 is a quadratic residue of 5. However, $x^2 \equiv 13 (\text{mod } 5)$ has no solution, showing that 13 is a quadratic nonresidue of 5.

The Legendre symbol simplifies discussion of quadratic reciprocity. If p is an odd prime that does not divide a, we write $(a/p) = 1$ if a is a quadratic residue of p and $(a/p) = -1$ if a is a quadratic nonresidue of p. From our previous discussion, we have $(11/5) = 1$ and $(13/5) = -1$.

We have already mentioned Euler's criterion, which states that r is a quadratic residue of an odd prime p if $r^{(p-1)/2} \equiv 1 (\text{mod } p)$ and r is a quadratic nonresidue of p if $r^{(p-1)/2} \equiv -1 (\text{mod } p)$. Thus, from the fact that $5^{(11-1)/2} = 5^5 = 3125 \equiv 1 (\text{mod } 11)$, we conclude that 5 is a quadratic residue of 11. However, $5^{(13-1)/2} = 5^6 = 15{,}625 \equiv -1 (\text{mod } 13)$, which indicates that 5 is a quadratic nonresidue of 13. Hence, $(5/13) = -1$ while $(5/11) = 1$.

Also, $x^2 \equiv 11 (\text{mod } 7)$ has the solution $x \equiv 2 (\text{mod } 7)$, so $(11/7) = 1$. However, $x^2 \equiv 7 (\text{mod } 11)$ has no solution, whence $(7/11) = -1$.

We see that in some cases $(p/q) = (q/p)$ as in $(5/13) = (13/5) = -1$ or $(11/5) = (5/11) = 1$, but in other cases $(p/q) = -(q/p)$, for example, $(11/7) = 1 \neq (7/11) = -1$. The question arises as to when equality holds and when it does not. Euler discovered a pattern that he believed answered this question, and Gauss later gave several proofs that this pattern was indeed true in all cases. The key to the pattern is the observation that every odd prime has either the form $4n+1$ or $4n+3$. If p and q are odd primes and either one of them has the form $4n+1$, then $(p/q) = (q/p)$, as in $(5/11) = (11/5)$. However, if both of the primes have the form $4n+3$, then $(p/q) = -(q/p)$. An example of the latter is $(11/7) = -(7/11)$.

The theorem is usually stated in the more elegant form

Law of Quadratic Reciprocity. When p and q are distinct odd primes then

$$(p/q)(q/p) = (-1)^{[(p-1)/2][(q-1)/2]}.$$

Gauss gave six proofs of the law of quadratic reciprocity and more than 50 proofs have been devised.

A number of assertions by Fermat can be shown to follow from the law of quadratic reciprocity, including the following.

1. Every prime of the form $8n+1$ or $8n+3$ has the form x^2+2y^2. For example, $17=3^2+2(2^2)$ or $41=3^2+2(4)^2$.
2. Every prime $3n+1$ has the form x^2+3y^2 but no prime $3n-1$ has this form.
3. Every prime $4n+1$ is the sum of two squares in only one way.

G. Formulas for Primes

A glance at a list of the first few primes is sufficient to convince one that any formula that would yield all of the primes would be a most unusual contrivance. Even the related but less demanding task of finding a formula that would produce only primes seems most formidible.

Certain formulas furnish partial results, For example, if we replace x in the function $f(x)=x^2+x+41$ with $x=0, \pm1, \pm2, \ldots, \pm39$, or -40, the result in each case is a prime. The same is true of $g(x)=x^2-79x+1601$ for $x=0, 1, 2, \ldots, 79$. The function f is a composite of two formulas, one given by Euler in 1772, the other by Legendre in 1798.

Euler also found that $2x^2+n$ is prime for $x=0, 1, \ldots, n-1$ where n is one of the numbers 3, 5, 11, or 29. Paul Pritchard discovered the imposing expression,

$$11{,}410{,}337{,}850{,}553 + 4{,}609{,}098{,}694{,}200x,$$

which is prime for $x=0, 1, \ldots, 21$.

Similar formulas of this nature exist, but it can be shown that no polynomial function having integral coefficients can generate only prime numbers for integral values of x. One need only observe the effect of calculating $f(41)=(41)^2+41+41$ to understand why any polynomial will eventually produce a composite number.

The function f given above generates primes for 80 consecutive integers. It has been conjectured that this is the largest number of consecutive primes that a quadratic polynomial can produce. All that is known for sure is that no quadratic having the form x^2+x+c, where $c>41$, can yield primes for all values of $x=0, 1, 2, 3, \ldots, c-2$.

From algebra we know that a polynomial of degree n can be constructed to assume $n+1$ arbitrary values. Thus, the function $h(x)=-x^3/6+x^2+x/6+2$ gives the first four primes for $x=0, 1, 2$, and 3. However, a similar polynomial yielding the first 101 primes would have degree 100. The longest known arithmetic progression consisting entirely of primes is given by $223{,}092{,}870n+2{,}236{,}133{,}941$ where $n=0, 1, 2, \ldots, 15$.

W. H. Mills proved the following remarkable theorem: There exists a real number θ such that $[\theta^{3n}]$ is prime for all positive integers n. Unfortunately, there seems to be little hope of determining the value of θ, so the theorem is presently of no use for constructing primes. Similar formulas exist for expressing p_n as a function of some parameter θ. In order to determine θ to an accuracy sufficient to calculate p_n, it proves necessary to know the primes $p_1, p_2, p_3, \ldots, p_n$. Such formulas are thus equally useless for constructing new primes, unless some method can be found for determining θ without using $p_1, p_2, p_3, \ldots, p_n$. No such method has yet been found, nor has it been proved impossible for such a method to exist. The question is thus an open one at present.

The number of primes whose values do not exceed some positive integer x is symbolized by $\pi(x)$. Thus, $\pi(12)=5$, since 2, 3, 5, 7, and 11 are primes not exceeding 12. Similarly, $\pi(13)=6$.

The method of finding primes known as the sieve of Eratosthenes leads to the formula

$$\begin{aligned}\pi(N)-\pi(\sqrt{N})+1 = N &- [N/p_1] - [N/p_2] - \cdots \\ &- [N/p_k] + [N/p_1p_2] + [N/p_1p_3] + \cdots \\ &+ [N/p_{k-1}p_k] - [N/p_1p_2p_3] - \cdots,\end{aligned}$$

where $p_1, p_2, \ldots, p_k$ are the primes less than $\sqrt{N}$. The formula involves considerable calculation. To determine the number of primes less than one million, one needs to consider the primes less than 1000. Obviously, today's computers make such a formula vastly easier to apply and therefore more useful.

There are certain shortcuts that can be employed with the formula for $\pi(x)$; Before the advent of computers, Bertelsen determined that there are 50,847,478 primes that do not exceed one billion. An analogous formula for the number of pairs of twin primes the larger of which does not exceed N was given by G. J. Kostis and R. L. Page.

There exists a number pattern that generates primes in a curious but inefficient way:

```
                1   1
              1   2   1
            1   3   2   3   1
          1   4   3   5   2   5   3   4   1
  1   5   4   7   3   8   5   7   2   7   5   8   3   7   4   5   1
```

Each successive row is formed by inserting between each consecutive pair of integers their sum. Any two consecutive entries are relatively prime. It is also true that the integer n occurs $\phi(n)$ times in the nth line, where $\phi(n)$ is the number of integers less than n and relatively prime to n. Since a prime p is relatively prime to all $p-1$ integers less than itself, we see that n is a prime if and only if it appears $n-1$ times in the nth row. Since 2 appears once in row 2, it must be a prime. Similarly, 3 appears twice in row 3 and 5 appears 4 times in row 5. However, 4 appears

only twice in row 4, making 4 a composite number. The number of integers in each row increases rapidly, there being 433 of them in row 10. Thus, the pattern is of little practical value in finding primes.

H. The Prime Number Theorem

By examining tables of prime numbers and by a great amount of trial-and-error calculation, mathematicians discovered that the quantities $\pi(x)$ and $x/\ln x$ behave in a similar fashion and that their ratio approaches 1 as x increases without bound. This was confirmed upon proof of the following theorem.

Prime Number Theorem. $\lim_{x\to\infty}[\pi(x)/(x/\ln -x)] = 1$.

This was proved independently by Hadamard and De la Vallée Poussin.

Another function that approximates $\pi(x)$ better than $x/\ln x$ is the integral logarithm, defined by

$$\mathrm{Li}(x) = \int_2^x \frac{dt}{\ln t}.$$

I. Riemann's Zeta Function

Another function of great importance in the study of the distribution of primes is Riemann's zeta function: $\zeta(s) = \Sigma_{n=1}^{\infty}(1/n^s)$. For example, $\zeta(1) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots$, which may be shown to diverge and $\zeta(2) = 1 + \frac{1}{4} + \frac{1}{9} + \cdots$, which converges to $\pi^2/6$. The function converges for all $s > 1$. Its relation to prime numbers stems from an identity, which Euler discovered, that expresses the zeta function as the repeated product of a term evaluated only for primes. Many of the important functions used in studying properties of primes may be expressed as combinations of zeta functions.

While no formula exists for generating every prime, Bertrand's postulate assures us that if $n \geq 1$, there exists at least one prime p such that $n < p \leq 2n$.

J. Mersenne Primes

The study of the primality of numbers having a particular form has occupied a great deal of attention in number theory. A number of the form $M_n = 2^n - 1$ is called a Mersenne number in honor of Marin Mersenne (1588–1648). Thus, $M_2 = 3$, $M_3 = 7$, $M_4 = 15$, etc. When M_n is a prime p, it is called a Mersenne prime and the number $2^{p-1}(2^p - 1)$ is a perfect number. Furthermore, every even perfect number must be of this form. Thus, the search for even perfect numbers reduces to a search for Mersenne primes. Because $2^n - 1$ can be factored when n is composite, a Mersenne prime must have the form $M_p = 2^p - 1$. For $M_2 = 3$, the perfect number is $2(3) = 6$ and for $M_3 = 7$, the perfect number is $4(7) = 28$.

K. Fermat Numbers

Numbers of the form $2^{2^t} + 1$ are called Fermat numbers. We see that $F_0 = 3$, $F_1 = 5$, $F_2 = 17$ and $F_4 = 65{,}537$ are all primes. The next one, $F_5 = 2^{32} + 1$, is difficult to test for factors without the use of a computer. Fermat conjectured that all Fermat numbers were primes, and it was a hundred years before Euler found a counterexample. He did this by proving that any factor of a Fermat number must have the form $(2^{t+1})k + 1$. For $t = 5$, these factors are of the form $64k + 1$. He then showed that 641 is a factor of F_5.

No further primes have been discovered among the Fermat numbers and many believe that quite the opposite of Fermat's conjecture is true. Nevertheless, interest in Fermat numbers continues because of a remarkable result due to Euler. He showed that a regular polygon of N sides can be constructed with only straightedge and compass if $N = 2^k p_1 p_2 \cdots p_n$, where the p_i are Fermat primes. This unexpected connection between number theory and geometry is an example of the richness of results obtained by pondering prime numbers.

L. Twin Primes

Little is known about twin primes, not even whether an infinite number of pairs exist. By examining the first few pairs, (3,5), (5,7), (11,13), (17,19), (29,31), etc., one is led to conjecture: If p and q are twin primes both larger than 3, then they have the form $6k \pm 1$. This conjecture is readily proven. Any integer must have one of the forms $6k$, $6k \pm 1$, $6k \pm 2$, or $6k + 3$, all of which, except $6k \pm 1$, are easily seen to be composite.

The largest known pair of twin primes, discovered in 1990 by B. K. Parady, J. F. Smith, and S. Zarantonello, is

$$1706595 \times 2^{11235} \pm 1$$

M. Representation of Numbers in Certain Forms

The operation of division leads to questions of factorability of integers and to congruences, to mention only two things. So, too, does the operation of addition lead to many important questions in number theory. For example, the question of whether a given integer can be represented as the sum of an arbitrary number of squares has occupied many mathematicians over the years. For the case of two squares, we have the following theorem:

Theorem. Every prime of the form $4k + 1$ can be written as the sum of two squares.

Thus, $29 = 4(7) + 1$ has the required form and can be written $29 = 4 + 25 = 2^2 + 5^2$. A more powerful result states: An integer n can be written as the sum of two squares if and only if all of its prime factors having the form $4k + 3$ occur with even exponents. We have $5(7^2) = 245 = 49 + 196 = 7^2 + 14^2$ as the sum of two squares, whereas $7^3 = 343$ is not.

The case for three squares is settled by the following theorem:

Theorem. An integer n can be represented as the sum of three squares unless n has the form $4^e(8k + 7)$ for some integers e and k.

We see that $45 = 4 + 16 + 25 = 2^2 + 4^2 + 5^2$, whereas $60 = 4(8 + 7)$ cannot be so represented.

Lagrange proved that every positive integer is the sum of four squares. The search for the four squares that represent a certain number can be reduced to representing the prime factors of the number as the sum of four squares. An identity then gives the four squares for the original number.

The question of finding the least value of s such that every integer can be expressed as the sum of no more than s kth powers of integers is known as Waring's problem. It has long been known that every integer can be written as the sum of nine cubes. Kevin S. McCurley proved that every integer exceeding $e^{e^{13.97}}$ is a sum of seven positive integral cubes.

For fourth powers, the most that can be said is that every integer larger than $10^{10^{89}}$ can be represented as a sum of 19 fourth powers. For fifth powers, the question is unresolved. Surprisingly, for many higher powers, the problem has been solved. It is known that for $6 \le K \le 200{,}000$, every integer can be written as a sum of no more than $2^k + [(3/2)^k] - 2$ kth powers. Hence, every integer is the sum of no more than $2^6 + [(3/2)^6] - 2 = 73$ sixth powers.

We have seen that every integer can be represented as the sum of no more than nine cubes. In fact, if 23 and 239 are exempted, every integer is the sum of no more than eight cubes. Similarly, every integer larger than 454 is the sum of no more than seven cubes. One is therefore led to ask: What is the smallest number $G(k)$ of kth powers whose sum represents any sufficiently large integer? I. M. Vinogradov showed that

$$k + 1 \le G(k) \le k(3 \ln k + 11).$$

Although one of the outstanding accomplishments in number theory, this nevertheless falls short of answering the question.

From studying tables of primes, one comes to the inescapable conclusion that most integers are composite. This is true in the sense that the ratio $\pi(x)/x$ approaches zero as x approaches infinity. Nevertheless, because primes are the building blocks for all integers, research concerning their properties continues to be a major area in number theory.

We sometimes refer to numbers such as 1000 or 1,000,000 as nice round numbers because of the zeros in their base 10 representation. More generally, we consider a round number to be one having a large number of relatively small factors. Therefore, $4200 = 2^3 \cdot 3 \cdot 5^2 \cdot 7$ would be a round number but $17{,}858{,}257 = (3607)(4951)$ would not. It can be shown that the number of prime factors of an integer x is, on the average, of the order of $\ln(\ln x)$. That is, for all integers in a large interval, the preponderance of them have as the number of their prime factors a number close to $\ln(\ln x)$.

N. Factorization Methods

Many methods have been devised to aid in the factorization of large numbers. Fermat's factorization method depends on writing the number to be factored as the difference of two squares. If

$$n = x^2 - y^2 = (x + y)(x - y) = ab,$$

we may assume that n is odd and then a and b will also be odd. If we write $x^2 = n + y^2$, then $x \ge \sqrt{n}$. We examine $x^2 - n$ for various values of $x > \sqrt{n}$, and when the difference is a square, we have found the desired factorization.

For example, if $n = 26{,}781$, we see that $(164)^2 - 26{,}781 = 115$, which is not a square. Instead of considering $(165)^2 - 26{,}781$, $(166)^2 - 26{,}781$, etc., we can use the easier but equivalent method of adding $2x + 1$, $2x + 3$, etc., and check each result for a square.

x	$x^2 - 26{,}781$
165	$115 + 329 = 444$
166	$444 + 331 = 775$
167	$775 + 333 = 1108$
168	$1108 + 335 = 1443$
169	$1443 + 337 = 1780$
170	$1780 + 339 = 2119$
171	$2119 + 341 = 2460$
172	$2460 + 343 = 2803$
173	$2803 + 345 = 3148$
174	$3148 + 347 = 3495$
175	$3495 + 349 = 3844 = (62)^2$

Then

$$(175)^2 - 26{,}781 = (62)^2$$

or

$$26{,}781 = (175 + 62)(175 - 62) = 237(113).$$

Euler's factorization method depends on expressing the number to be factored as the sum of two squares in two different ways: $N = a^2 + b^2 = c^2 + d^2$. Then $a^2 - c^2 = d^2 - b^2$ or $(a+c)(a-c) = (d+b)(d-b)$, where we may assume that a and c are odd and b and d are even. Let k be the gcd of $a-c$ and $d-b$. Then k is even and $a - c = kl$ and $d - b = km$. Then $(a+c)kl = (d+b)km$ or $(a+c)l = (d+b)m$. We see that $a+c$ is divisible by m since $(l, m) = 1$. Thus $a + c = nm$, where n is even. Also, $nml = (d+b)m$ or $nl = d + b$. We see that n is the gcd of $a+c$ and $d+b$. The factorization is

$$N = [(k/2)^2 + (n/2)^2](m^2 + l^2).$$

For example,

$$1{,}000{,}009 = (235)^2 + (972)^2 = 3^2 + (1000)^2.$$

We have $a - c = 232$, $a + c = 238$, $d - b = 28$, and $b + d = 1972$. Then $k = (232, 28) = 4$, $n = (238, 1972) = 34$, $l = (a-c)/k = \frac{232}{4} = 58$, and $m = (d-b)/k = \frac{28}{4} = 7$. We find the factorization

$$1{,}000{,}009 = \left[\left(\tfrac{4}{2}\right)^2 + \left(\tfrac{34}{2}\right)^2\right](7^2 + 58^2) = 293(3413).$$

O. Fibonacci Numbers

Consider the following ancient problem. Assume that we have a pair of newborn rabbits, one male and one female. Suppose that at the end of two months the pair becomes mature. At the end of the third month, and every succeeding month, they produce a male–female pair. Finally, assume that the rules just stated apply to each new pair of rabbits that is born. Assuming no deaths, how many pairs of rabbits will there be at the end of any given month?

If we represent an immature pair of rabbits by an open circle and a mature pair by a shaded circle, we have the results shown in Fig. 15.

Notice that the number of mature pairs each month equals the total number of pairs from the previous month. Also, the number of immature pairs each month equals the number of mature pairs from the previous month. But, since the number of mature pairs from the previous month equals the total number of pairs from two months ago, we see that the total number of pairs in any given month is the sum of the total number of pairs from the previous two months.

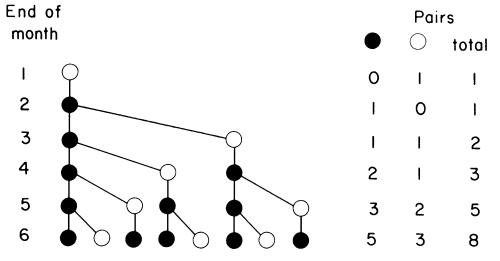


FIGURE 15 The Fibonacci sequence.

TABLE I The first 30 Fibonacci numbers

n	F_n	n	F_n	n	F_n
1	1	11	89	21	10,946
2	1	12	144	22	17,711
3	2	13	233	23	28,657
4	3	14	377	24	46,368
5	5	15	610	25	75,025
6	8	16	987	26	121,393
7	13	17	1597	27	196,418
8	21	18	2584	28	317,811
9	34	19	4181	29	514,229
10	55	20	6765	30	832,040

The sequence $1, 1, 2, 3, 5, 8, \ldots$, defined by $F_1 = 1$, $F_2 = 1$, $F_n = F_{n-1} + F_{n-2}$ for all $n \geq 3$, is called the Fibonacci sequence in honor of Leonardo of Pisa (who was known as Fibonacci, the son of Bonacci). Table I lists the first 30 terms of the Fibonacci sequence.

Many properties have been discovered for the sequence, which has been the subject of extensive study. One readily sees, for example, that every third term is divisible by 2, every fourth term is divisible by 3 and every fifth term is divisible by 5. In fact, it can be shown that every prime p divides an infinitude of Fibonacci numbers.

Several other easily discovered properties are

1. The sum of the first n Fibonacci numbers is one less than F_{n+2}.
2. $F_{n-1}F_{n+1} + 1 = F_n^2$.
3. $F_n^2 + F_{n+1}^2 = F_{2n+1}$.
4. If A, B, C, and D are four consecutive Fibonacci numbers, then $C^2 - B^2 = AD$.

The sequence is also related to patterns found in nature. For example, the seeds on the head of a sunflower lie in rows that form clockwise (CW) and counterclockwise (CCW) spirals. The numbers of spirals of each type are often consecutive Fibonacci numbers with the smaller number enumerating the CCW spirals and the larger number the CW ones. A typical sunflower will have 34 CCW and 55 CW spirals, although heads with 144 CCW and 233 CW spirals have been reported.

Other sunflower heads may have numbers of CCW and CW spirals given by consecutive Lucas numbers, named for Edouard Lucas. These numbers are 1, 3, 4, 7, 11, $18, \ldots$, where $L_1 = 1$, $L_2 = 3$, and $L_n = L_{n-1} + L_{n-2}$ for $n \geq 3$.

If we consider the ratios F_{n+1}/F_n, we have the sequence $1, 2, 1.5, 1.666, \ldots, 1.6154, 1.679, 1.6180, \ldots$.

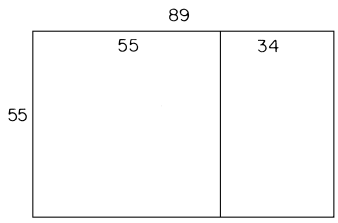


FIGURE 16 A rectangle divided in proportion to numbers in the Fibonacci sequence.

This sequence converges to the number $(1+\sqrt{5})/2 = 1.6180339\ldots$, which is the golden ratio of ancient Greece.

Another interesting relationship involves the first k Fibonacci numbers. Suppose we choose the first 11 of them: 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, and 89. We construct a rectangle having sides of 55 and 89 and then divide it into a square of side 55 and a 34-by-55 rectangle as shown in Fig. 16. Now we divide the rectangle into a square of side 34 and a 34-by-21 rectangle as in Fig. 17. We continue the process of constructing squares in the remaining rectangles until two 1-by-1 squares are constructed. Then in the 55-by-55 square, we draw a quarter circle of radius 55, in the 34-by-34 square we draw a quarter circle of radius 34 that is connected to the first quarter circle, and continue in this fashion until all of the squares have quarter circles. The resulting curve, which closely approximates the logarithmic spiral, a curve found in the shell of the chambered nautilus, is shown in Fig. 18.

III. MODERN DIRECTIONS

Research in number theory flourishes today, the results of that research appearing in numerous scholarly journals on a monthly or quarterly basis. Obviously, this article cannot begin to treat the current state of research in the field. Besides being too large in scope, the subject has become extremely abstract, with many topics which can be fully understood only by experts in the field.

High-speed computers have affected number theory by allowing the consideration of cases involving computations far beyond the capacity of the human mind. Furthermore, by performing thousands of operations a second, computers reduce the time required for certain calculations to within reasonable limits. Starting with M_{521}, the most recently discovered Mersenne primes were found with the aid of computers. Some results that were established by use of home computers have been published.

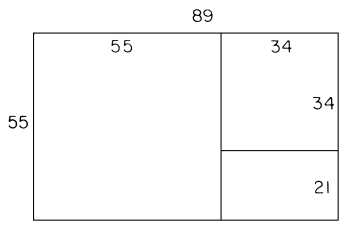


FIGURE 17 A rectangle divided in proportion to numbers in the Fibonacci sequence.

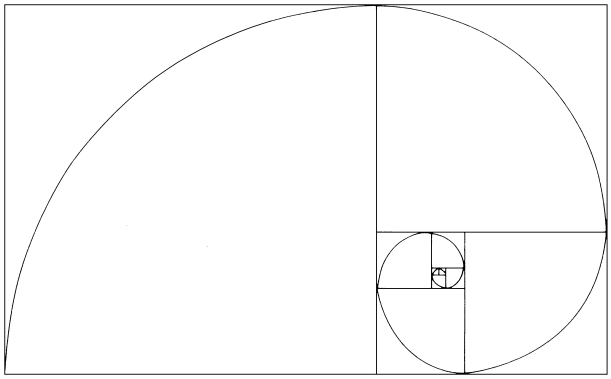


FIGURE 18 An approximation to the logarithmic spiral.

Computers have sparked renewed interest in other ancient problems. In 1974, H. J. J. teRiele announced a pair of amicable numbers, each of which has 152 digits. Using a probabilistic algorithm in conjunction with a computer, Michael O. Rabin found a pair of numbers of the order of magnitude 10^{123}, which he conjectures are twin primes.

In November, 1996, Joel Armengaud and George F. Woltman proved that $2^{1,398,269}-1$ is a Mersenne prime that would require more than 400,000 digits to write in the base 10 system and is the largest known prime.

Research in the distribution of primes continues, including the study of primes that differ by a fixed amount and gaps between primes. Sol Weintraub has found the largest known gap between primes, which consists of 682 consecutive composite numbers following the prime 61,003,096,898,749. In view of the use of computers to facilitate searches, it is safe to say that many of today's largest results will soon be surpassed.

Sometimes a long-term conjecture falls prey to a counterexample. Euler conjectured that no nth power is a sum of fewer than n nth powers; for example, no cube is the sum of fewer than 3 cubes. In 1966, L. J. Lander and T. R. Parkin found the counterexample:

$$144^5 = 27^5 + 84^5 + 110^5 + 133^5.$$

In 1921, V. Brun showed that even if the number of twin primes should be infinite, they are more thinly distributed throughout the integers than are the primes. More specifically, he showed that the sum of the reciprocals of the primes is infinite but the sum of the reciprocals of twin primes is finite.

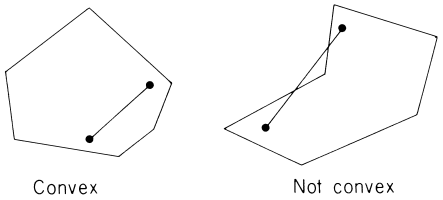


FIGURE 19 Examples of convex and nonconvex regions.

A. The Geometry of Numbers

Elementary number theory refers to those problems whose solution does not require methods from calculus. While this is still an important area in number theory, various other branches have developed in modern times. One such branch, known as the geometry of numbers, arose from a theorem by Hermann Minkowski. In its simplest form, the theorem concerns lattice points in a plane, that is, points whose coordinates are integers.

To state the theorem, we need to define a convex region in the plane, that is, a region R having the property that if any two points of R are connected by a straight line segment, all of the points of the segment will also lie in R. (See Fig. 19.)

Minkowski's theorem states that any convex region R that is symmetric about the origin and whose area is greater than 4 will contain lattice points other than the origin. (See Fig. 20.) Although seemingly self-evident, to prove the theorem requires a fair amount of work. The theorem may be extended to n dimensions and various relationships and functions defined on the lattice points of both convex and nonconvex sets. Such advanced study of lattice points constitutes the geometry of numbers.

B. Analytic Number Theory

Analytic number theory involves the use of methods from analysis or calculus, especially from the theory of complex variables, for solving problems in number theory. This branch arises from the work of Dirichlet and Georg F. B. Riemann (1826–1866), both of whom are sometimes credited with its founding.

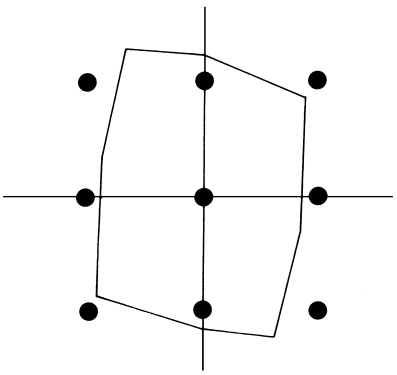


FIGURE 20 An illustration of Minkowski's theorem.

A Dirichlet series has the form $F(s) = \sum_{n=1}^{\infty}(\alpha_n/n^s)$, where α_n is an expression that can be defined for each integer n. The simplest case is $1 + 1/2^s + 1/3^s + \cdots$, which, as we have seen, is called Riemann's zeta function. Riemann's hypothesis, his only conjecture that remains unproved concerning the Riemann zeta function, states that the complex zeros of that function, that is, all zeros of the form $x + y\sqrt{-1}$, have $x = \frac{1}{2}$

This area tends to be abstract and the problems difficult, some yielding only to ingenious methods of attack. It may seem surprising that problems stated in terms of integers may require for their solution powerful methods from analysis, a branch of mathematics dealing with continuous quantities rather than the discreet set of integers. Indeed, some matters, such as the convergence of certain series, may properly be considered areas belonging to analysis.

C. Algebraic Number Theory

The branch known as algebraic number theory developed from attempts to apply concepts from number theory to sets of numbers other than the natural numbers. For example, Gauss considered numbers of the form $a + b\sqrt{-1}$, where a and b are integers. These are called the Gaussian integers and they follow a unique factorization law analogous to that of the fundamental theorem of arithmetic.

Euler, in attempting to prove Fermat's last theorem (see Section IV.B) for the case in which $n = 3$, considered numbers of the form $a + b\sqrt{-3}$, where a and b are integers. For such sets of numbers one may define primes that correspond to primes in the set of natural numbers, although some of them will be quite different from the ordinary primes.

Euler assumed, without proof, that unique factorization held for these numbers. His conclusion, although unproved, was correct. Later, Gabriel Lamé made a similar unproved assumption and announced that he had solved Fermat's last theorem. His error was quickly pointed out.

About the same time, Ernst Eduard Kummer was attempting to extend Gauss' quadratic reciprocity law to higher powers. He succeeded in showing that unique factorization does not hold for all sets of numbers. Thus, in the set of numbers $a + b\sqrt{-5}$, the number 6 has two different prime factorizations:

$$6 = 2(3) = (1 + \sqrt{-5})(1 - \sqrt{-5}).$$

An equation of the form $a_0x^n + a_1x^{n-1} + \cdots + a_n = 0$, where each a_i is rational, is said to be irreducible if the

left-hand side cannot be factored into two similar expressions. If r is a root of such an irreducible equation of degree n, then the set of all expressions that can be formed from that root by addition, subtraction, multiplication, and division by nonzero terms is called the algebraic number field of degree n generated by r.

In order to preserve the concept of unique factorization, Kummer introduced what he called ideal numbers. He was able to find a sufficient condition for Fermat's last theorem to be true and proved it for several particular cases.

Dedekind extended this concept to algebraic number fields in general by considering sets in the fields called ideals. This work led to the development of field theory, an area in which number theory overlaps another branch of modern mathematics called abstract or modern algebra.

Much research is presently conducted on Diophantine equations. A related topic is Diophantine approximation, the approximation of irrational numbers by rational numbers.

A review of current literature in number theory gives evidence of interest in old as well as new topics. Among the former are tests for divisibility, distribution of quadratic residues, Riemann's zeta function, Pythagorean triples, and the distribution of primes. Among the latter are Fortune's conjecture (see Section IV.A) and permutable primes. These are numbers such as 13 or 37 that are also prime when their digits are reversed.

IV. UNSOLVED PROBLEMS AND CONJECTURES

The natural numbers are the simplest set of numbers used in mathematics, yet the number of patterns derived from the natural numbers seems almost endless. These patterns have led mathematicians to ask questions such as: Is the pattern true for all integers? Under what conditions does the pattern hold?

A. Conjectures

When statements concerning number patterns are proved to be true, they are called theorems. Before that, they are conjectures, nothing more than educated guesses based on inductive reasoning applied to a finite number of particular cases. Conjectures, then, fall in a kind of no-man's-land between the list of facts that have been shown to be true and the host of statements that have been shown to be false. They remain there until they are proven to be true theorems or until a counterexample shows them to be false.

Some conjectures enjoy long lives before they are disproved. For example, Fermat's conjecture that numbers of the form $2^{2^t}+1$ are always prime survived for a hundred years before it died at the hands of Euler. Many other patterns are disposed of almost as quickly as they appear. As an example, consider some patterns which seem to generate primes. For $p_1^2+(p_1+1)p_2$ we have

$$2^2+3(3)=13$$
$$3^2+4(5)=29$$
$$5^2+6(7)=67$$
$$7^2+8(11)=137$$
$$11^2+12(13)=277,$$

all of which are prime. However, $13^2+14(17)=407=11(37)$.

Again, consider $p_1^{p_2}+p_1+p_2$:

$$2^3+2+3=13$$
$$3^5+3+5=251$$
$$5^7+5+7=78{,}137,$$

all of which are prime. Numbers in this sequence grow in size rapidly. Thus, $7^{11}+7+11=1{,}977{,}326{,}761$, which is harder to check for primality by use of a pocket calculator than its predecessors. Of course, it is easily handled by a modern computer. Even this is not necessary, however, since the next term $11^{13}+11+13=11^{13}+24$ is clearly divisible by 5.

The numbers 31, 331, 3331, and 33,331 are all primes. In fact, the pattern continues to yield primes until we reach 333,333,331, which is divisible by 17.

Another interesting pattern is

$$3!-2!+1!=5$$
$$4!-3!+2!-1!=19$$
$$5!-4!+3!-2!+1!=101$$
$$6!-5!+4!-3!+2!-1!=619$$
$$7!-6!+5!-4!+3!-2!+1!=4421$$
$$8!-7!+6!-5!+4!-3!+2!-1!=35{,}899,$$

which yields the primes listed. Unfortunately, the next step gives 326,981, which is divisible by 79.

The following pattern gives primes for many steps:

$$41+2=43$$
$$43+4=47$$
$$47+6=53$$
$$53+8=61$$
$$61+10=71$$
$$71+12=83.$$

In fact, it will give primes for another 33 steps before the composite number $1681 = 41^2$ appears. The pattern is made up of numbers obtained from the formula $x^2 + x + 41$, which we have previously discussed.

Finally, consider the pattern of integers in which each succeeding row is obtained by inserting the number of the row, n, between each pair of integers from row $n-1$ whose sum is n:

n														Number of terms
1							1							2
2						1	2	1						3
3					1	3	2	3	1					5
4				1	4	3	2	3	4	1				7
5		1	5	4	3	5	2	5	3	4	5	1		11
6	1	6	5	4	3	5	2	5	3	4	5	6	1	13

This pattern fails for row 10, which contains 33 terms. If we count the number of digits in each row, the tenth row contains 37 digits, a prime number. However, the 11th row contains 57 digits, a composite number.

A recent conjecture is due to Reo F. Fortune who examined the pattern

$2 + 1 = 3$	$5 - 2 = 3$
$2(3) + 1 = 7$	$11 - 6 = 5$
$2(3)(5) + 1 = 31$	$37 - 30 = 7$
$2(3)(5)(7) + 1 = 211$	$223 - 210 = 13$
$2(3)(5)(7)(11) + 1 = 2311$	$2333 - 2310 = 23$
$2(3)(5)(7)(11)(13) + 1 = 30{,}031$	$30{,}047 - 30{,}030 = 17$

The first five sums in the left-hand column are primes, but $30{,}031 = 59(509)$. However, for each sum if we find the next larger prime and subtract from it the product of consecutive primes given in that row, the result is prime. Fortune's conjecture is that this pattern always gives primes. Many feel that the conjecture is true but proving it appears to be a difficult task.

The question of whether there exist an infinite number of Mersenne primes has been unsolved for approximately 300 years, as has the companion question of the existence of an infinite number of even perfect numbers. To date, only 28 Mersenne primes are known, so the question is far from resolved.

Catalan's Conjecture, due to Eugène Charles Catalan (1814–1894), states that 8 and 9 are the only positive consecutive integral powers of integers.

In general, this suggests that $x^m - y^n = 1$, where x and y are integers greater than 0 and m and n are integers greater than 1, has as its only solution $x = 3$, $y = 2$, $m = 2$, and $n = 3$. Since m and n vary, as well as x and y, the equation above is a Diophantine equation that is not in polynomial form.

In 1974, Robert Tijdeman proved that there exists a constant k with the property that all powers of integers which equal consecutive integers are less than k. Thus, we know that there can be at most a finite number of such pairs of consecutive powers, although the work of calculating k seems too formidible to allow a definite value to be determined at present.

In 1876, Catalan also examined the sequence $P_0 = 2$ and $P_{n+1} = 2^{P_n} - 1$. Thus,

$$P_1 = 2^{P_0} - 1 = 2^2 - 1 = 3$$
$$P_2 = 2^{P_1} - 1 = 2^3 - 1 = 7$$
$$P_3 = 2^{P_2} - 1 = 2^7 - 1 = 127$$
$$P_4 = 2^{P_3} - 1 = 2^{127} - 1.$$

He speculated that P_n is prime for $n = 1, 2, 3$, and 4, all of which were subsequently verified. P_5 seems to be undecidable since it has approximately 10^{38} digits.

B. Fermat's Last Theorem

One of the most famous of all problems in number theory, unsolved for over 350 years, goes by the misnomer of Fermat's last theorem. In the margin of a copy of Diophantus' "Arithmetic," opposite a problem concerning writing a square as the sum of two squares, Fermat wrote that it is impossible to write a cube as the sum of two cubes, a fourth power as the sum of two fourth powers, and so forth. In other words, he claimed that the equation $x^n + y^n = z^n$ cannot be solved when $n > 2$.

He also claimed to "have discovered a truly marvelous demonstration of this proposition that this margin is too narrow to contain." In view of Fermat's wrong guess concerning the primality of all Fermat numbers, one must be skeptical of his claim, or at least wish he had had a supply of blank paper at hand.

Only one proof concerning number theory is known to be due to Fermat, that being found in another margin of the same book. This theorem showed that the area of a Pythagorean triangle having integral sides cannot be a square integer. This theorem leads to the proof of Fermat's last theorem for the case $n = 4$; that is, $x^4 + y^4 = z^4$ has no solutions.

Fermat claimed to be able to prove the conjecture for $n = 3$, but published no proof. Euler gave such a proof nearly 100 years later, but it contained some faulty reasoning that fortunately could be corrected.

Work on the problem progressed slowly as other mathematicians proved the conjecture true for $n = 5$, $n = 7$, and other particular values of n. It can be shown that if the conjecture is true for some integer k, then it is true for any multiple of k. Hence, it suffices to consider only odd prime powers. Partial results included showing that if the theorem is true for some value of $n > 2$, then n must exceed 4,000,000. It was also shown that any solution led to

numbers that are inconceivably exceeding the capacity of even the most modern computers.

Furthermore, it was known that $x^p + y^p = z^p$ has no solution in integers that are relatively prime to p if p is an odd prime and $q = 2p + 1$ is also prime. In 1983 Gerd Faltings proved that the formula contained in Fermat's last theorem has only finitely many rational solutions when $n > 2$. From time to time, several reputable mathematicians, as well as countless amateurs, gave purported proofs that Fermat's last theorem was true for all $n > 2$. Unfortunately, errors were discovered in each of the proofs presented before 1993.

In June, 1993, Andrew J. Wiles of Princeton University gave three lectures at Cambridge University, which culminated in a proof of Fermat's last theorem. Wiles had been interested in the problem since the age of 10 when he vowed to find a proof. His interest in the problem led him to choose mathematics as a career.

His teachers and professors advised him that he would be wasting his time pursuing such a difficult and uncertain goal. At Cambridge, his advisor guided him to the field of elliptic curves, which are curves having cubic equations and which can be used for calculating the perimeter of ellipses.

Wiles' concern for the theorem never abated, even while he did research on elliptic curves. In 1986, his determination was strengthened by the work of other mathematicians who postulated and then proved a connection between Fermat's last theorem and elliptic functions. He devoted himself completely to solving the problem and seven years later he was ready to present his proof to fellow mathematicians.

The reaction to his proof was astonishment and widespread acclaim. Verifying his proof was a slow process for two reasons: (1) his proof was 200 pages long and (2) it was estimated that elliptic curves were understood by approximately only one tenth of one percent of all professional mathematicians.

Minor errors in his proof were found and easily corrected. Then a catastrophe occurred; a flaw in the proof was discovered that could not be easily overcome. Wiles and his former student, Richard Taylor, worked tirelessly to save the proof. They finally decided to employ a different type of elliptic curv that would eliminate the error.

After 14 months Wiles began to belive that his proof would suffer the same fate as all previous ones, failure. Then, one of those events occurred that sound more like fiction than fact. They were close to giving up but on September 19, 1994 they found a way to eliminate the error and save the proof.

In June, 1997 Wiles was awarded the Wolfskehl Prize, which had been unclaimed for 89 years, for proving Fermat's last theorem. Had he not studied elliptic curves at Cambridge, Wiles might not have been prepared to do the work which led to his successful proof. Thus, his advisor unwittingly helped Wiles achieve his childhood goal.

C. Goldbach's Conjectures

Nearly 250 years ago, Christian Goldbach, in correspondence with Euler, posed two conjectures:

1. Every even number greater than or equal to 6 is the sum of two odd primes.
2. Every odd number greater than or equal to 9 is the sum of three odd primes.

The second conjecture is actually a consequence of the first, and in 1937 Vinogradoff proved that any odd number that is sufficiently large is the sum of three odd primes. How large the number has to be is not known. It has been established that if N is an even integer larger than $e^{e^{16,038}}$, then N is the sum of no more than four primes.

It is also known that every sufficiently large even integer is the sum of a prime and an integer having no more than two distinct prime factors. Complete resolution of Goldbach's conjectures awaits future results.

The question of the existence of odd perfect numbers is still unresolved. Most mathematicians are inclined to believe that perfect numbers must be even, although an odd one may be detected someday by methods as yet unimagined. It is known that any such integer must exceed 10^{50} and must have at least eight prime factors.

Number theory is the oldest branch of mathematics and concerns the simplest set of numbers, the integers. Because some of its problems can be stated in easily understood terms, it probably has attracted more amateurs than any other branch of mathematics. Although many of its problems today are stated by means of abstract technical definitions not easily mastered by the lay person, nevertheless the subject that has commanded widespread attention over the past 3000 years should remain a vital area of human learning for at least that far into the future.

SEE ALSO THE FOLLOWING ARTICLES

COMPUTER ALGORITHMS • LINEAR SYSTEMS OF EQUATIONS • NUMBER THEORY, ALGEBRAIC AND ANALYTIC • SET THEORY

BIBLIOGRAPHY

Boston, N., and Greenwood, M. L. (1995). "Quadratics representing primes," *Am. Math. Month.* **102**(7), 595–599.

Bunt, L. N. H., Jones, P. S., and Bedient, J. D. (1976). "The Historical Roots of Elementary Mathematics," Prentice-Hall, Englewood Cliffs, New Jersey.

Crandall, R. E. (1997). "The challenge of large numbers," *Sci. Am.* **276**(2), 74–78.
Dudley, U. (1969). "Elementary Number Theory," Freeman, San Francisco, California.
Edwards, H. M. (1978). Fermat's last theorem. *Sci. Am.* **239**(4), 104–122.
Gardner, M. (1979). *Sci. Am.* **241**(3), 22–32.
Gardner, M. (1980). Mathematical games. *Sci. Am.* **243**(6), 18–28.
Hardy, G. H., and Wright, E. M. (1960). "The Theory of Numbers," 4th ed. Oxford Univ. Press (Clarendon), London and New York.
Garrison, B. (1981). Consecutive integers for which $n^2 + 1$ is composite. *Pacific J. Math.* **97**(l), 93–96.
Kostis, G. J., and Page, R. L. (1964). A formula concerning twin primes. *Math. Mag.* **37**(3), 153–154.
McCoy, N. H. (1965). "The Theory of Numbers," Macmillan, New York.
McCurley, K. S., An effective seven cube theorem, *J. Number Theory* **19**(2), 176–183.
Newman, J. R. (ed.) (1956). "The world of Mathematics," Simon & Schuster, New York.
Pomerance, C. (1980/1981). Recent developments in primality testing. *Math. Intelligencer.* **3**(3), 97–105.
Ribenboim, P. (1994). "Prime number records," *College Math. J.* **25**(4), 280–290.
Rouse Ball, W. W. (1960). "A Short Account of the History of Mathematics," Dover, New York.
Singh, S., and Ribet, K. A. (1997). "Fermat's last stand," *Sci. Am.* **277**(5), 68–73.
Weintraub, S. (1982). A prime gap of 682 and a prime arithmetic sequence. *BIT* **22**(4), 538.



Numerical Analysis

John N. Shoosmith
NASA, Langley Research Center, retired

I. Numerical Analysis as a Subject Area
II. Finite-Precision Numerical Operations
III. Algebraic Equations in a Single Variable
IV. Systems of Algebraic Equations
V. Matrix Eigenproblems
VI. Numerical Representation of Functions
VII. Differentiation and Integration
VIII. Differential Equations
IX. Recent Developments

GLOSSARY

Algorithm A numerical algorithm is a precise, step-by-step description of the implementation of a numerical method.

Condition A problem is ill-conditioned if a small change in the data results in a large change in the solution. Conversely, it is well-conditioned if the solution is relatively insensitive to changes in the data. The condition number is a measure of condition. It is small for a well-conditioned problem and large otherwise.

Convergence A numerical method is said to converge if the solution generated by the method approaches a solution of the problem to which it is applied—either as an iteration count becomes large, or as a discretization parameter approaches zero. The order of convergence is a measure of the rate of convergence. For example, in the case of an iteration method, the order of convergence is p if, after a sufficient number of cycles, the error after any cycle is comparable to the error of the previous cycle raised to the power p.

Data By input data for a problem, we mean numbers that are required to be provided in order for the solution process to start or continue. Output data are numbers generated by the solution process.

Error Difference between the approximation to a quantity and its true value. Relative error is the error divided by the true value. An error bound is a positive number that is known to be larger than the magnitude of an error.

Iteration Repeated process, where the input to each cycle is determined from the output of preceding cycles. The input to at least the first cycle is given in order to start the process. Criteria for stopping the iteration must be provided.

Method A numerical method is a procedure used to solve a mathematical problem through the use of numbers.

Parallel algorithm Algorithm designed to use more than one processor at the same time. The computations are organized into tasks that can be assigned to multiple processors.

Roundoff error Error introduced because of the limit to the precision (number of places) to which numbers can be represented in any finite computation.

Stability A numerical method is unstable if, when it is applied to a well-conditioned problem, a small change in the data results in a large change in the numerical solution. It is stable if the change in the solution remains small.

Truncation error Error introduced because of the limit to the number of numbers that can be used in any finite computation. For example, a variable may be represented exactly by an infinite mathematical series, but only the most significant terms can be retained. The discarded terms contribute to the truncation error in the computed solution.

Vector algorithm Algorithm designed to make efficient use of a vector processor. In a vector processor a single instruction causes the same operation to be performed on one or more sequences of numbers (vectors) in an assembly-line fashion.

NUMERICAL ANALYSIS is the study of the solution of mathematical problems through the manipulation of numbers. The solution processes are usually referred to as numerical methods, and the required sequences of numerical and logical operations, when precisely set down, are called algorithms. The solutions thus obtained are usually not exactly correct, but approximately so. In the context of science and engineering, the problems of concern are either the result of reducing the analysis of a physical situation to mathematical equations that cannot be simplified further by usual mathematical means, or the physical laws and constraints governing a system under study are "modeled" by mathematical equations—in which case, the numerical solution can be said to simulate the behavior of the system. Although numerical methods have been used for centuries (Johannes Kepler used one to determine the orbit of Mars in 1607), the development of digital computers, with their tremendous capacity for carrying out arithmetic and logical operations on numbers, has made numerical methods practical in a great variety of scientific and technological applications. Today, almost all numerical methods are carried out on computers. From an understanding of number representations and manipulations, the numerical analyst must devise the methods and design the algorithms to solve a variety of problems. In doing so, he or she must be concerned with the source and propagation of errors, and whether or not the method "converges" to a close approximation of the solution. Also, in spite of the great speed of modern computers, the matter of convergence rate and computational efficiency are of paramount importance and may determine whether or not the method is practical. A recent development in computing is the use of collections of computers to solve individual problems (parallel processing), and this requires the consideration of problem partitioning and algorithm development to take advantage of the particular computer architectures involved. It is in this area that numerical analysis and computer science are particularly closely related.

I. NUMERICAL ANALYSIS AS A SUBJECT AREA

A. The Numerical Approach to Problem Solution

The means by which physical situations and processes are described, analyzed, designed, and simulated is through mathematics. Natural laws are stated in terms of mathematical equations, and the behavior of systems that obey those laws is described by their solutions.

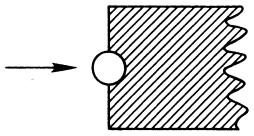
Unfortunately, the mathematics of many of the processes we would like to study quickly becomes intractable when approached by conventional means. For example, analytical solutions to most nonlinear systems of equations simply cannot be found. The best that can be done, in the traditional sense, is to attempt a series expansion of the solution.

Today, there is another approach: the problem statement and variables of interest can be approximated numerically. Analysis and problem solution can then be performed through numerical computation with the aid of high-speed digital computers. To be sure, something is lost when it becomes necessary to resort to numerical methods, because characteristics of the solution that are immediately apparent from inspection of analytical expressions may be obscured in listings of numbers; however, a numerical solution is certainly better than no solution, and sometimes its nature can be revealed by repeating the process with small changes in the data. Also, computer-generated graphs or images can often provide a sufficiently accurate visual interpretation of the solution to aid in its understanding.

The numerical approach is illustrated for a very simple problem in Fig. 1. The problem is posed in physical terms in (a), and its mathematical formulation is given in (b). In this situation we know the solution, but in more complex cases, of course, we may not.

The next step is to select or develop a numerical method. Here, we have chosen the Euler method for the solution of initial value problems involving ordinary differential equations, again because of its simplicity. Numerical methods can be thought of as operators that accept numbers as input (in this case the initial velocity

V_0, the problem parameters D and M, and the discretization parameter h) and produce other numbers as output (the successive values of time and velocity).



A ball of mass M kg, moving with velocity V_0 m/sec, strikes a block of paraffin. If the drag force per unit of velocity is D kg/sec, what is the subsequent velocity of the ball as a function of time?

(a)

$$dV/dt = -(D/M)V$$
$$V(0) = V_0$$

Note that for this problem we know the solution to be

$$V(t) = V_0 e^{-(D/M)t}$$

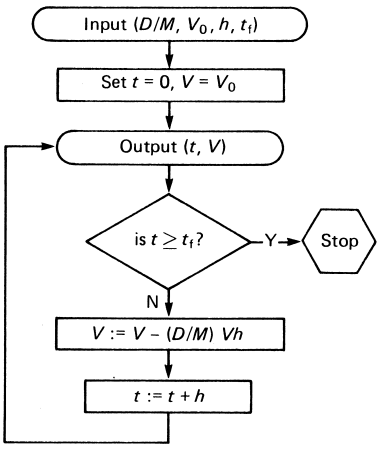
(b)

The Euler method for the solution of initial-value problems in ordinary differential equations. Given V_0

$$V_{i+1} = V_i + (dV/dt)_i h$$
$$i = 0, 1, 2, \ldots$$

where h is a selected interval length.

(c)



(d)

FIGURE 1 The numerical approach to problem solution. (a) Physical situation. (b) Mathematical formulation. (c) Numerical method. (d) Algorithm.

The final stage is to produce an algorithm, a step-by-step implementation of the method. Algorithms are thought of rather like flow charts and are usually described in an unambiguous way by means of an algorithmic or even a computer-programming language. Algorithms are recipes that could conceivably be followed by a person with pencil and paper; however, it is usual to convert them to computer programs, which can then be executed on a suitable computer.

B. Errors, Their Sources, and Propagation

We use the example of Fig. 1 to illustrate the source of errors in the numerical solution of a problem.

If we carry out the first step of the Euler method for this example, using symbols V_0, D, and M for the data, and the symbol h for the time increment, we arrive at

$$V(h) \approx V_1 = V_0[1 - (D/M)h].$$

On the other hand, we know from the series expansion for e^x, carried out to two terms with remainder, that the solution at $t = h$ is

$$V(h) = V_0 e^{-(D/M)h}$$
$$= V_0\left[1 - \left(\frac{D}{M}\right)h + \left(\frac{D}{M}\right)^2 e^{-(D/M)z}\frac{h^2}{2}\right],$$

where z appearing in the final (remainder) term has a value between 0 and h. Comparing the approximate and true values of $v(h)$, we see that the difference is

$$\varepsilon(h) = V_0(D/M)^2 e^{(-D/M)z}h^2/2,$$

where $\varepsilon(h)$ is the truncation error committed by the Euler method in carrying out the first step. Since we know that $e^{(-D/M)x}$ can be no more than 1 for any non-negative value of x, we can say that $\varepsilon(h)$ is bounded by a constant, independent of h, times h^2; that is,

$$\varepsilon(h) < Kh^2.$$

This tells us that if, for example, we halve the interval, the maximum possible truncation error for the first step will be divided by four.

We are normally interested in computing the solution out to some initially specified time, say t_f, which will require $n = t_f/h$ steps. The accumulated truncation error, which we will designate by $\varepsilon(h, t_f)$, is bounded by n times the bound for one step, thus

$$\varepsilon(h, t_f) < nKh^2 = t_f Kh.$$

This brings us to the concept of convergence. We see that as the step size h is made smaller, the answer produced by the Euler method (in the absence of any arithmetic error) for the velocity at time t_f becomes more accurate.

In fact, in the limit as h approaches zero, the answer is exact, since the error bound is then zero. Of course, it does not make sense to use a zero interval size, but the point is that we can make the error as small as we wish by selecting h sufficiently small. In this case, the truncation error in computing $V(t_f)$ is bounded in direct proportion to the interval size h, and we say that the method converges with order h.

Turning now to the computation of the solution as described by the algorithm, we first must specify values for the parameters of the problem and the initial data. Because, in any practical computing device, the number of digits allocated to a number is limited, it will probably be necessary to chop or round these numbers before they are stored. The error committed by doing this is called inherent roundoff. Also, during the course of the computation, arithmetic operations are performed that produce results with more digits than the operands, and these results must be chopped or rounded before they are stored. This error is called arithmetic roundoff. A particularly serious consequence of roundoff occurs when two numbers of nearly the same value are rounded before they are subtracted, since this can result in the loss of significant information.

It is possible to minimize the effect of roundoff error by judicious design of the algorithm. For example, the computation of the roots of a quadratic equation

$$ax^2 + bx + c = 0,$$

by the quadratic formula,

$$r = (-b \pm \sqrt{b^2 - 4ac})/2a,$$

when b is large relative to a and c, can produce a poor approximation to the smallest in magnitude root. To avoid this, it is possible to use the mathematically equivalent formula,

$$r = 2c/(-b \mp \sqrt{b^2 - 4ac})$$

for that root.

Another example is in the calculation of expressions of the type

$$\sum_{i=1}^{n} a_i b_i.$$

The results of each multiplication can be saved and added together in extended precision. (If the computer hardware cannot do this, then it can be done with the use of a special program, treating the lower and upper halves of the multiplication results as separate variables.) Then the accumulated sum may be rounded at the end of the calculation.

In order to assess the effect of an error as it is propagated through the computation, we consider an algorithm as a mapping F from a set of input numbers $X \equiv \{x_1, x_2, \ldots, x_n\}$ to a set of results $Y \equiv \{y_1, y_2, \ldots, y_m\}$ and write

$$y_i = F(x_1, x_2, \ldots, x_n), \quad i = 1, 2, \ldots, m.$$

The error propagation formula is simply a first order difference expansion of this equation, Thus,

$$\Delta y_i \approx \left(\frac{\partial F_i}{\partial x_1}\right)\Delta x_1 + \left(\frac{\partial F_1}{\partial x_2}\right)\Delta x_2 + \cdots + \left(\frac{\partial F_i}{\partial x_n}\right)\Delta x_n, \quad i = 1, 2, \ldots, m.$$

The effect on the ith output variable of an error in the jth input variable can be estimated by knowing the appropriate partial derivative. Of course, in most cases this will not be known directly; however, it may be possible to find an experimental estimate by observing the change in the output variables as the algorithm is executed with small, incremental changes in the input data, one variable at a time.

Arithmetic operations that introduce roundoff error are often treated as equivalent, exact operations with erroneous operands. By working backward, the effect of arithmetic roundoff can be reduced to an equivalent error in the input, which can then be translated to output error by the use of the error propagation formula.

C. Condition and Numerical Stability

In solving a problem numerically, it is inevitable that errors will be committed, both because of the use of a discrete method (truncation) and because of finite precision calculations (roundoff). There are two effects that can cause such errors to grow to serious proportions, one of which has to do with the problem itself, and the other due to the method which is employed.

Some problems are inherently ill-conditioned, which means that a small change in the data results in a large change in the solution. A simple example is the system of two linear equations representing the intersection of two straight lines that are nearly parallel, such as

$$y = x + 1, \quad y = 1.01x.$$

These equations have the solution (100, 101); however, if the coefficient of x is changed to $1.001x$, less than a 1% change, the solution is then (1000, 1001), which is a 900% change in x. It is important that the numerical analyst be aware that a problem is ill-conditioned. It may be possible to reformulate it to avoid the ill conditioning, or if not, to use a higher-order method and increased precision arithmetic in order to reduce the introduction of numerical error as much as possible.

Numerical instability is a condition that is due to the method employed. A simple example is the solution of a differential equation problem

$$y' = f(x, y), \quad y(0) = y_0$$

by the "midpoint method,"

$$y_{i+1} = y_{i-1} + 2hf(x_i, y_i), \quad y_1 = y_0 + hf(x_0, y_0).$$

In this method, a small error can grow as x increases, even when the problem itself is well-conditioned, because of the existence of a parasitic solution. This is explained more fully in the section on differential equations. It is a challenge to the numerical analyst to recognize instability and to provide an alternate, stable method.

II. FINITE-PRECISION NUMERICAL OPERATIONS

A. Number Representation

In our standard, positional, decimal notation, a real number is represented as

$$\pm d_n d_{n-1} \cdots d_0 . d_{-1} d_{-2} \cdots,$$

where each d is a digit, that is, a symbol representing zero or one of the first nine natural numbers, and the subscript is an index specifying the position of the digit in the number. The $\pm$ indicates that the sign may be either $+$ or $-$. The period after d_0 is called the decimal point, and it separates the integral part of the number (on the left) from the fractional part. The leading digit d_n is nonzero except when the integral part is zero.

The value of the number, which we will designate by x, is then given by

$$x = \pm\left(\sum_{i=0}^{n} d_i \times 10^i + \sum_{j=1}^{\infty} d_{-j} \times 10^{-j}\right),$$

the first sum being the value of the integral part and the second being that of the fraction.

The number of distinct symbols used to represent a single digit in the decimal system is 10 (probably because we have 10 fingers), and we say that the base of the decimal system is 10. Electronic devices depend on sensing either the presence or absence of a signal, just two possible states, which we can represent with 0 and 1. Thus, for electronic computers we need to consider the binary representation, which uses the base 2. The binary representation of a real number is

$$\pm b_n b_{n-1} \cdots b_0 . b_{-1} b_{-2} \cdots,$$

where each b (bit) is either 0 or 1. The value of this number is

$$x = \pm\left(\sum_{i=0}^{n} b_i \times 2^i + \sum_{j=1}^{\infty} b_{-j} \times 2^{-j}\right).$$

In general, for the integer base $B > 1$, a real number is represented by

$$\pm g_n g_{n-1} \cdots g_0 . g_{-1} g_{-2} \cdots,$$

where the g's are symbols for zero and the first $B - 1$ positive integers; its value is given by

$$x = \pm\left(\sum_{i=0}^{n} g_i \times B^i + \sum_{j=1}^{\infty} g_{-j} \times B^{-j}\right).$$

Commonly used bases, in addition to 10 and 2, are 8 (octal), which uses the symbols 0 through 7, and 16 (hexadecimal), which uses 0 through 9, then A, B, C, D, E, and F to represent the integer values from 10 to 15. These bases are useful because conversion to and from binary is easily accomplished through the grouping of bits three and four at a time, respectively. Table I gives representations of the first 16 positive integers for five different bases.

Arithmetic is carried out in any base system by using the same procedure as in decimal arithmetic, keeping in mind, however, that "carries" and "borrows" are dependent on the value of the base. Figure 2 gives examples of arithmetic in binary, octal, and hexadecimal.

Conversion from one base to another is accomplished by a slightly different procedure, depending on the base in which the arithmetic is to be performed. Examples are given, in Fig. 3, of conversion between octal and decimal. If the arithmetic is to be done in the base notation *from*

TABLE I The First Sixteen Natural Numbers Written in Some Different Base Notations

Decimal (base 10)	Binary (base 2)	Ternary (base 3)	Octal (base 8)	Hexadecimal (base 16)
1	1	1	1	1
2	10	2	2	2
3	11	10	3	3
4	100	11	4	4
5	101	12	5	5
6	110	20	6	6
7	111	21	7	7
8	1000	22	10	8
9	1001	100	11	9
10	1010	101	12	A
11	1011	102	13	B
12	1100	110	14	C
13	1101	111	15	D
14	1110	112	16	E
15	1111	120	17	F
16	10000	121	20	10

Binary

```
  1110        1110        1110          111
+  111      −  111      ×    11     10)1110
 10101         111         1110         10
                          1110          11
                        101010          10
                                        10
                                         0
```

Octal

```
  256         256         256           35
 +173        −173      ×    4      6)256
 451           63        1270         22
                                       36
                                       36
                                        0
```

Hexadecimal

```
  5AC         5AC         5AC          16B
 +1F6        −1F6      ×    4      4)5AC
 7A2           63        16BO         4
                                      1A
                                      18
                                       2C
                                       2C
                                        0
```

FIGURE 2 Arithmetic in different base notations.

which we are converting, then the procedure is to separate the integer and fractional parts. The integer part is converted by successive divisions by the base to which we are converting, as illustrated in Fig. 3 in Example 1, while the fractional part is converted by successive multiplications as illustrated in Example 2. On the other hand if the arithmetic is to be performed in the base notation to *which* we are converting, the procedure is to use the value formula

$$x = \pm\left(\sum_{i=0}^{n} g_i B^i + \sum_{j=1}^{\infty} g_{-j} B^{-j}\right)$$

as illustrated in Example 3.

B. Computer-Representable Numbers

The memory of a computer is composed of cells, called words, each consisting of a number of bits. Most computers use 8, 16, 32, or 64 bits per word. It is usual to store a single number in each word (although multiple and fractional precision modes are possible with some loss of processing speed). There are several ways in which to do this; however, if the word length is m bits, it is possible to store only 2^m different numbers in a given word. The bit positions in a word are numbered, starting with zero, from the least significant (which we can think of as the right end) bit to the most significant (left end) bit [Fig. 4(a)]. The most significant bit is usually reserved for the sign of the number, for example, 0 for a positive number and 1 for negative.

It is most efficient, in terms of the use of storage space and processing speed, to store numbers in binary form; however, it should be mentioned that it is possible to represent decimal numbers by allocating groups of four bits to each digit (binary-coded decimal). This is done in calculators and in applications where the input and output of data dominates computation [Fig. 4(b)].

In the "binary integer" mode, integers are stored, right-justified, in the computer word just as they are written in binary [Fig. 4(c)]. All of the integers from -2^{m-1} to $+2^{m-1}$ can thus be represented in a word of length m bits. Counting and indexing operations use this mode.

For more general computation, modern computers use a mode referred to as binary floating point, in which numbers are represented in the form

$$x = \pm M \times 2^t,$$

where M, the "significand," is in the range $0 \le M < 2$ and t, the "exponent," is an integer. If $M \ge 1$, the number is said to be "normalized," otherwise it is "denormalized." Thus the decimal number 358.625 (101100110.101 in binary) is represented in normalized binary floating point form as

$$+1.40087890625 \times 2^8.$$

Expressed in binary, the significand is 1.01100110101 and the exponent is 1000.

A national standard for binary floating point floating point arithmetic has been developed under the auspices of the Institute for Electrical and Electronic Engineers-(IEEE Standard 754-1985). In this standard, a single format number is stored in a 32-bit computer word with the fractional part of the significand, the "fraction," stored in binary in the least significant 23 bits. The sign is stored as 0 for a positive number and 1 for a negative number in the most significant bit. In order to accommodate negative exponents, the exponent is biased by the addition of 127 (1111111 in binary) and stored in binary in the 8 bits between the sign and fraction. Provided that the biased exponent is not 0 or 255, the number is assumed to be normalized and the integer part of the significand is implicitly taken as 1. The "precision" of a single format normalized binary floating point number is 24 bits. The single format representation of 358.625 is given in Fig. 4(d).

The number zero is stored as a 0 in every bit position, except possibly the most significant (sign). Positive zero

EXAMPLE 1. Conversion of the decimal integer 358 to octal by successive division by 8 in decimal arithmetic.

```
8)358
8) 44    R 6        Answer: 546 (base 8)
8)  5    R 4
    0    R 5
```

EXAMPLE 2. Conversion of the decimal fraction .769 to octal by successive multiplication by 8 in decimal arithmetic.

```
  .769
×   8
(6).152
×   8              Answer: .611 ... (base 8)
(1).216
×   8
(1).728
```

EXAMPLE 3. Conversion of the octal number 546.612 to decimal by expansion in decimal arithmetic.

$$
\begin{aligned}
5 \times 8^{2} &= 320 \\
+4 \times 8^{1} &= 32 \\
+6 \times 8^{0} &= 6 \qquad \text{Answer: 385.769 ... (base 10)} \\
+6 \times 8^{-1} &= .75 \\
+1 \times 8^{-2} &= .0156 \\
+2 \times 8^{-3} &= .0039 \\
&\overline{358.7695}
\end{aligned}
$$

FIGURE 3 Conversion of base.

and negative zero are numerically equivalent but the distinction can be useful; for example, division of a positive number by negative zero produces negative infinity.

The smallest positive normalized single format number is 2^{-126}. If any operation produces a positive number that is smaller than this, the biased exponent is set to zero and the significand is shifted to the right so that the number is no longer normalized and the precision is less than 24 bits. Similarly, the largest negative normalized number is -2^{-126} and the production of a larger negative number results in a denormalized form. The generation of a denormalized number signals an underflow condition.

A biased exponent of 255 with a zero fraction indicates plus or minus infinity (e.g., the result of dividing a nonzero number by a signed zero). A biased exponent of 255 with a nonzero fraction is defined as a NaN (Not a Number). A NaN will be produced when the magnitude of the result of an operation is equal to or larger than 2^{128} (overflow condition).

The IEEE standard also defines a double format number that is stored in a 64-bit word. In the double format mode the fraction length is 52 bits, the bias is 1022, and the biased exponent length is 11 bits. The precision of a double format normalized number is 53 bits.

In order to reduce the effect of roundoff error, the floating point processing units of modern computers generally contain a number of registers that are longer than either the single or double format standard. Register to register computations take place in extended precision (also defined by the IEEE standard) and only when a result is to be stored in the computer's memory does it get shortened to single or double precision.

C. Roundoff Error

Exact arithmetic operations typically produce results that contain a greater number of significant positions than the operands. For example, the sum of 8.5 and 9.2 is 17.7 (one more significant digit than the addends) or the product of 1.234 and 1.432 is 1.767088 (the fractional part has twice as many significant digits). In a binary floating point processing unit, the exact result of an arithmetic operation on single format operands may be contained in an extended precision register; however, if the result is to be stored in single format, some loss of precision is inevitable. The shortening of a number to a lower precision is accomplished by "rounding."

There are two choices in rounding. The first is to merely discard the portion of the significand that follows the least significant bit of the shortened format; and the second is to add one (with appropriate carries to more significant positions) to the least significant bit after discarding the same portion. The default rounding mode in the IEEE standard is "round-to-nearest." This means that the choice is made

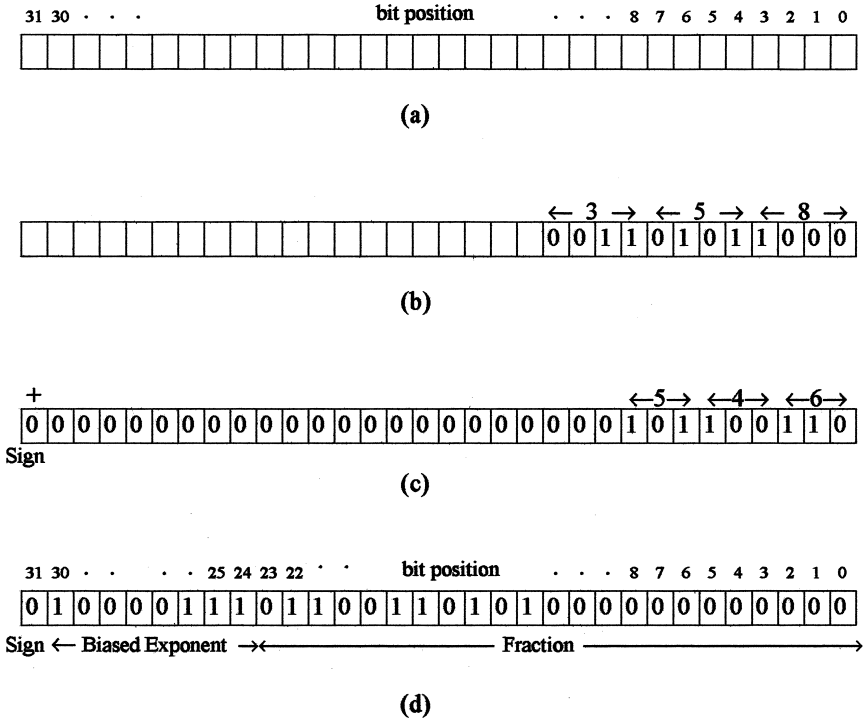


FIGURE 4 Representation of numbers in a 32 bit computer word. (a) Bit numbering convention. (b) Binary-coded decimal representation of 358. (c) Binary integer representation of +358 (+546 octal). (d) Single format floating point binary representation of +358.625.

so that the rounded number is nearest to the infinitely precise result. If both choices are equally near, then the even shortened significand is chosen.

The IEEE standard also provides for user specified "directed" rounding modes in order to reproduce results that may have been obtained on nonstandard implementations. Round-towards-zero (also referred to as chopping) always uses the first choice. Round-towards-positive-infinity uses the first choice if the number is negative and the second if it is positive. Round-towards-negative-infinity uses the first choice if the number is positive and the second if it is negative.

A bound for the relative error (magnitude of the maximum possible error divided by the exact number) is 2^{-p} for round-to-nearest and 2^{-p+1} for directed rounding, where p is the precision in bits (e.g., $p = 24$ for single format). This bound is referred to as the machine unit and designated by μ.

The generation of error in sequences of floating point operations is determined by backward error analysis. In this approach, each rounded result is supposed to be obtained by an exact calculation with numbers that are initially in error. It is possible to work backward in this manner to establish a bound on the error in the initial data that would lead to the same final result with exact arithmetic. An estimate of the accumulated error can then be obtained from the error propagation formula. Bounds for the error that have been obtained in this manner for some common sequences of calculations are shown in Table II.

The rate of convergence of iterative methods can be strongly affected by roundoff error; thus, the use of multiple extended precision registers in floating point units, in which sequences such as those shown in Table II can be executed prior to rounding to single or double format, can significantly improve the performance of the computer.

TABLE II Bounds for Errors in Some Common Floating Point Calculations

$F(x_1, x_2, \ldots, x_n)$	Absolute error bound[a]
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} \lvert(n+1-i)x_i\rvert \quad (1.06\mu)$
$\prod_{i=1}^{n} x_i$	$(n-1)\prod_{i=1}^{n} \lvert x_i\rvert \quad (1.06\mu)$
$\sum_{i=1}^{n} x_i y_i$	$\sum_{i=1}^{n} \lvert(n+2-i)x_i y_i\rvert \quad (1.06\mu)$
$\sum_{i=1}^{n} a_i x^i$	$\sum_{i=1}^{n} \lvert(2i+1)a_i x^i\rvert \quad (1.06\mu)$

[a] Note: μ is the machine unit; 2^{1-p} for chopping, 2^{-p} for rounding, where p is the significand length.

III. ALGEBRAIC EQUATIONS IN A SINGLE VARIABLE

A. Background

By an algebraic equation in a single independent variable x, we will mean an equation that can be put in the form

$$f(x) = 0,$$

where f is a single-valued function of x, containing no derivatives nor integrals with respect to x. Examples are

$$x - \sin x = 10$$
$$x^3 = 2$$
$$e^x + \ln x - 3 = 0.$$

For purposes of this discussion, we will assume that f is continuous and differentiable in x. Clearly, by adding x to both sides, we have the equivalent equation

$$F(x) = f(x) + x = x.$$

A commonly occurring problem is to solve the equation in one of these two forms, by which we mean to find a value of x for which the equation is true. In the first case a solution is called a root of $f(x)$, and in the second it is called a fixed point of $F(x)$.

In the case where $f(x)$ is linear, that is, of the form

$$f(x) = ax + b, \quad a \neq 0,$$

a solution exists, is unique, and can be determined immediately for given values of a and b. If $f(x)$ is not linear, the situation is a great deal more complicated, because there may not be a solution or there may be many solutions. In order to get some understanding of the nature of the problem, it is usually advantageous to construct a graph of $f(x)$ versus x and observe if and where this graph crosses the x axis, or alternatively to graph $F(x)$ and observe if and where it crosses the line $y = x$. This is illustrated in Fig. 5.

Another complication for a nonlinear equation is that (with certain specific exceptions) it is not possible to obtain a solution in a finite number of operations. The best that can be done is to approximate a solution by making an initial numerical estimate and then to methodically refine it by obtaining a succession of (hopefully, ever closer) estimates, until there is little change between successive values. This process is referred to as iteration.

To make this more precise, suppose a solution is x^*, the initial estiamte is x_0, and successive iterations produce the infinite sequence

$$x_1, x_2, x_3, \ldots, x_i, \ldots,$$

then the absolute error of the ith iterate is $|x_i - x^*|$. The iteration is said to converge to x^* if

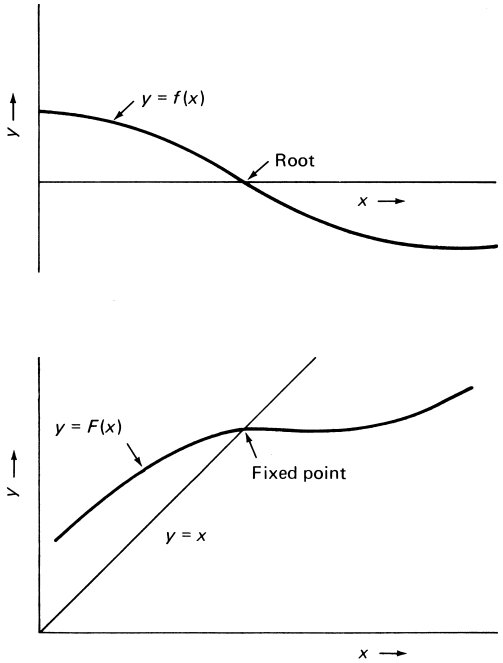


FIGURE 5 The graphical determination of a root or fixed point.

$$\lim_{i \to \infty} |x_i - x^*| = 0.$$

If $f(x)$ is continuous and differentiable, a sufficient condition for convergence is that there exists a constant C such that

$$\frac{|x_{i+1} - x^*|}{|x_i - x^*|} \leq C < 1$$

for all i greater than some threshold number. If this equation holds, the convergence is said to be of order 1, or linear. More generally, if it can be shown that there is a number $p \geq 1$ and a constant C such that

$$\frac{|x_{i+1} - x^*|}{|x_i - x^*|^p} \leq C < 1$$

for all i greater than some threshold, then the iteration converges with order p. In particular, if $p = 2$, the convergence is said to be quadratic. The higher the order of convergence, the fewer the number of iterations that should be required to home in on the solution; however, whether or not a particular iteration will converge and how rapidly it will do so depend on a number of factors, such as how close the initial estimate is to the desired solution and how close solutions are to one another. The situation of multiple, identical solutions can be particularly troublesome.

As a practical matter, a test for convergence is usually made at the end of each iteration. If

$$\frac{|x_{i+1} - x_i|}{|x_i|} < \varepsilon_1$$

for a prescribed $\varepsilon_1 > 0$, usually of the order of the machine unit μ, the iteration is stopped and the solution is taken as x_{i+1}. In the event that x_i is close to zero, however, the test must be modified to the absolute form, and the iteration is stopped when

$$|x_{i+1} - x_i| < \varepsilon_2,$$

where, again, ε_2 is a sufficiently small positive number. Finally, in the root-finding case, a termination test can be in the form

$$|f(x_{i+1})| < \varepsilon_3.$$

B. Iteration Methods

The simplest possible iteration method is the ancient method of repeated substitution, which can most readily be applied to an algebraic equation in the (fixed-point) form

$$x = F(x).$$

Starting with an initial estimate of x_0, the iteration formula is

$$x_{i+1} = F(x_i).$$

It can be shown that the method converges to x^* if

$$|F'(x)| < 1$$

for $(x^* - |x_0 - x^*|) \le x \le (x^* + |x_0 - x^*|)$; the order of convergence is linear if $F'(x^*) \neq 0$.

Bracketing methods are applied to equations in the (root) form,

$$f(x) = 0.$$

They rely on first finding two values of x, say a and b, such that $f(a)$ and $f(b)$ have opposite sign. Thus, if $f(x)$ is continuous, a root lies somewhere between a and b. Two methods for successively narrowing the interval containing the root are the bisection method and the method of false position. These methods are illustrated in Fig. 6. Their advantage is that once an interval containing a root has been found, convergence is guaranteed. On the negative side, it is not always easy to find two values that bracket a root, and the order of convergence of both of these methods is only linear. In most cases the method of false position converges more rapidly than the method of bisection; however, there are situations where it is much slower. A common approach is to start with the method of false position and then to revert to the method of bisection if convergence is not achieved after a prescribed number of iterations.

Extrapolation methods use information at previous iterations to extrapolate to a new estimate of a root. Newton's method uses the value and derivative of $f(x)$ at $x = x_i$ to extrapolate linearly to x_{i+1}; the secant method uses the latest two values of $f(x)$ to extrapolate linearly; and Muller's method uses the latest three values to perform a quadratic extrapolation. The order of convergence of these methods (to simple roots) is 2, 1.62, and 1.84, respectively, but the major disadvantage is that initial estimates must be relatively close in order to achieve convergence.

Also, Newton's method requires the computation of the function and its derivative at each step. Newton's method and the secant method are illustrated in Fig. 7.

C. Roots of Polynomials

The particular case where $f(x)$ is a polynomial of degree n, that is, to find the roots of

$$p_n(x) = a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_n;$$
$$a_0 \neq 0$$

is a frequently occurring problem. Any of the methods of the previous section can, of course, be used for this case; however, it is sometimes possible to take advantage of the special properties of polynomials.

To start with, $p(x)$ can be evaluated for $x = \alpha$ by the following recurrence formula:

$$b_0 = a_0$$
$$b_i = \alpha b_{i-1} + a_i \qquad \text{for} \quad i = 1, 2, \ldots, n.$$

This is called the method of synthetic division, because the numbers $b_0, b_1, \ldots, b_{n-1}$ are the coefficients of the polynomial of degree $n-1$ that is the result of dividing $p(x)$ by the binomial $(x - \alpha)$. Then $p(\alpha) = b_n$ is the remainder for this division. Also, once a root x_1 has been found, that is, $p(x_1) = 0$, then the b's of the final iteration define the reduced polynomial which has for its roots the remaining $n-1$ roots of $p(x)$, so that additional roots can be sought, starting from the reduced polynomial.

The term $p'(x)$ can be evaluated for $x = \alpha$ by repeating the synthetic division process for the first $n-1$ b's; thus,

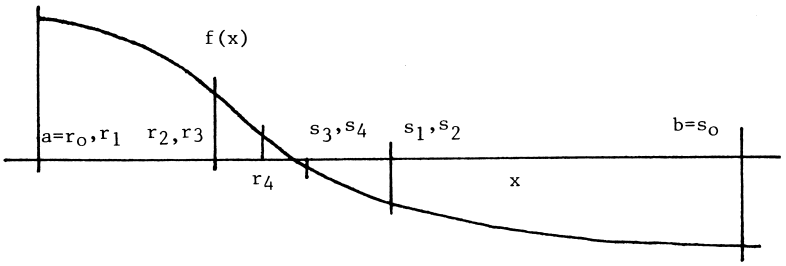
$$c_0 = b_0$$
$$c_i = \alpha c_{i-1} + b_i \qquad \text{for} \quad i = 1, 2, \ldots, n-1,$$

and

$$p'(\alpha) = c_{n-1}.$$

For this reason, Newton's method is easy to apply.

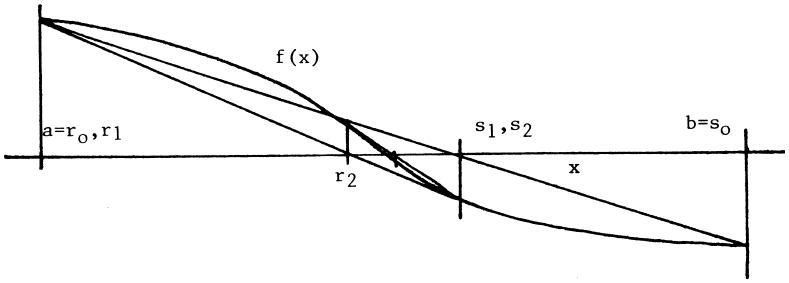
A disadvantage of the methods considered so far, except Muller's method, is that they cannot directly find the complex roots that are usually needed in the case of polynomial equations. Muller's method and Cauchy's method, which bears the same relationship to Muller's method as Newton's method does to the secant method [that is, it



Assume sgn$[f(a)f(b)] = -1$. Set $i = 0$, $r_0 = a$, $s_0 = b$.

→ Let $x_i = (r_i + s_i)/2$.
If $|r_i - s_i| < \varepsilon$, the root is $x \approx x_i$ (Stop).
If sgn$[f(r_i)f(x_i)] = -1$, set $r_{i+1} = r_i$, $s_{i+1} = x_i$.
Otherwise set $r_{i+1} = x_i$, $s_{i+1} = s_i$.
— Set $i = i + 1$.

(a)



The same as the bisection method except that

$$x_i = [s_i f(r_i) - r_i f(s_i)]/[(f(r_i) - f(s_i)]$$

(b)

FIGURE 6 Bracketing method. (a) Bisection method. (b) False position method.

uses the first and second derivative of $f(x)$ at an estimate of a root to extrapolate quadratically to the next estimate], are capable of finding complex roots and are thus more useful for the polynomial case.

A method designed for polynomial equations is due to Lin and Bairstow (not described in detail here because of its complexity). It uses synthetic division of $p(x)$ by a quadratic and iterates to reduce the linear remainder to zero. The quadratic factor thus determined may have complex conjugate roots, and the reduced polynomial has degree 2 less than $p(x)$.

The properties of polynomials can also be invoked to determine approximate locations of roots, which are often necessary in order to improve the chance of convergence of the methods discussed so far. Particularly useful in this regard is the theory of Sturm sequences, which can be used to determine the number of real roots in a prescribed interval.

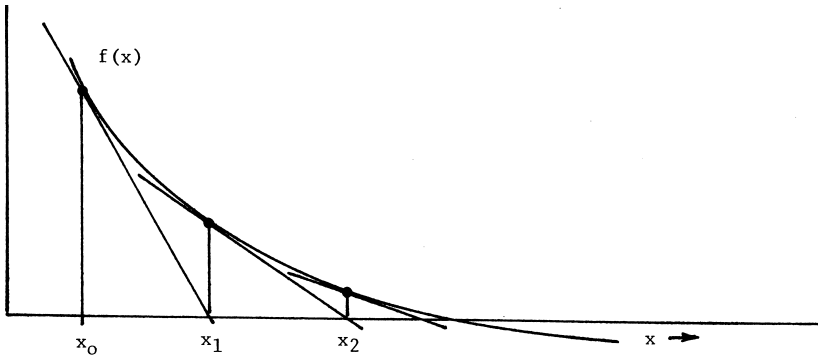
Finally, for every polynomial, a matrix can be found that has the same characteristic polynomial; thus, the problem of finding roots of a polynomial can be expressed as the problem of finding the eigenvalues of a matrix. Methods exist for finding all or selected numbers of eigenvalues of a matrix that do not depend for convergence on close estimates to start with, and are, therefore, often preferred over the methods discussed in this section. Matrix eigenvalue methods are discussed in Section V.

IV. SYSTEMS OF ALGEBRAIC EQUATIONS

A. Systems of Linear Equations

If $x_1, x_2, \ldots, x_n$ represent n variables, and $a_1, a_2, \ldots, a_n$ are given numbers (called coefficients), then the expression

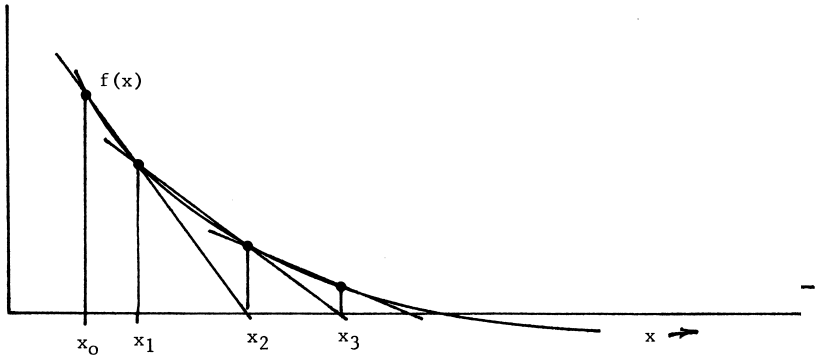
$$a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n$$



Given x_0,

$x_{i+1} = x_i - f(x_i)/f'(x_i)$; $i = 0, 1, 2, \ldots$

(a)



Given x_0 and x_1,

$x_{i+1} = x_i - (x_i - x_{i-1})f(x_i)/[f(x_i) - f(x_{i-1})]$

$i = 1, 2, 3, \ldots$

(b)

FIGURE 7 Extrapolation methods. (a) Newton's method. (b) Secant method.

is called a linear combination of the x's. The equation

$$a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n = b$$

is a linear equation. A system of m linear equations is written in the form

$$\begin{aligned}
a_{1,1}x_1 + a_{1,2}x_2 + a_{1,3}x_3 + \cdots + a_{1,n}x_n &= b_1 \\
a_{2,1}x_1 + a_{2,2}x_2 + a_{2,3}x_3 + \cdots + a_{2,n}x_n &= b_2 \\
a_{3,1}x_1 + a_{3,2}x_2 + a_{3,3}x_3 + \cdots + a_{3,n}x_n &= b_3 \\
&\vdots \\
a_{m,1}x_1 + a_{m,2}x_2 + a_{m,3}x_3 + \cdots + a_{m,n}x_n &= b_m,
\end{aligned}$$

where the subscript of the b's and the first subscript of the a's designate the equation, while the subscript of the x's and the second subscript of the a's designate the variable. We define the $m \times n$ matrix A of coefficients as the m-row by n-column array

$$A = \begin{bmatrix}
a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n} \\
a_{2,1} & a_{2,2} & a_{2,3} & \cdots & a_{2,n} \\
a_{3,1} & a_{3,2} & a_{3,23} & \cdots & a_{3,n} \\
\vdots & & & & \\
a_{m,1} & a_{m,2} & a_{m,3} & \cdots & a_{m,n}
\end{bmatrix}$$

and we define the n-vector x and the m-vector b as the columnar arrays

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ . \\ . \\ . \\ x_n \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ . \\ . \\ . \\ b_m \end{bmatrix}.$$

The system of linear equations can now be written concisely as

$$Ax = b,$$

where the multiplication of A times x (the order of these two factors is important) is defined to form an m-vector with the ith component being the inner product of the ith row of A with the vector x

$$a_{i,1}x_1 + a_{i,2}x_2 + a_{i,3}x_3 + \cdots + a_{i,n}x_n.$$

The characteristics of a linear system are determined by the properties of the matrix A and are addressed in the subject of linear algebra. Except where we discuss overdetermined systems at the end of the next section, we will assume that A is a square matrix ($m = n$, the order of A), and that there is a unique solution, that is, a unique set of values for the variables $x_1, x_2, \ldots, x_n$ that satisfies the system of equations. Under these circumstances the solution can always be obtained by direct methods in a finite number of arithmetic operations. When n is very large or when A is sparse (has a large proportion of zero elements), it is sometimes more efficient to use an iterative method, following the same general approach as discussed in the previous section on the solution of algebraic equations in a single variable.

B. Direct Methods

If A happens to be lower triangular, that is, has the structure

$$A = \begin{bmatrix} a_{1,1} & 0 & 0 & \cdots & 0 \\ a_{2,1} & a_{2,2} & 0 & \cdots & 0 \\ a_{3,1} & a_{3,2} & a_{3,3} & \cdots & 0 \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ a_{n,1} & a_{n,2} & a_{n,3} & \cdots & a_{n,n} \end{bmatrix}$$

then it is straightforward to obtain a solution by the process of forward substitution

$$x_1 = b_1/a_{1,1}$$

$$x_i = \left(b_i - \sum_{j=1}^{i-1} a_{i,j}x_j\right)/a_{i,i}, \qquad i = 2, 3, \ldots, n.$$

Similarly, if A is upper triangular, which can be visualized by reversing the order of the rows in the above matrix, then the solution can be obtained by back substitution

$$x_n = b_n/a_{n,n}$$

$$x_i = \left(b_i - \sum_{j=i+1}^{n} a_{i,j}x_j\right)/a_{i,i},$$

$$i = n-1, n-2, \ldots, 1.$$

(Note that under our assumption that there is a unique solution, it can be shown that the diagonal elements must be nonzero so that the division can always be carried out.) The number of additions and multiplications required for either process is of the order of $\frac{1}{2}n^2$, and the number of divisions is n.

The objective of direct methods for an arbitrary nonsingular matrix is to transform A into a triangular matrix. This may be done by a process that is equivalent to premultiplying A by a sequence of $n-1$ matrices. The product of two matrices is a matrix, each column is the product of the first matrix and the corresponding column of the second. Let A be designated as $A^{(0)}$; then the sequence of multiplications can be written recursively as

$$A^{(k)} = M^{(k)}A^{(k-1)}, \qquad k = 1, 2, 3, \ldots, n-1,$$

where $M^{(k)}$ is an $n \times n$ matrix with its diagonal elements equal to 1 and with 0 in every other position except, possibly, below the diagonal in the kth column, where the elements are

$$m_{i,k}^{(k)} = -a_{i,k}^{(k-1)} \big/ a_{k,k}^{(k-1)}.$$

Now $A^{(k)}$ has zeros below the diagonal through the kth column, and the final result $A^{(n-1)}$ is an upper-triangular matrix that we will denote by U. Now let L be the lower-triangular matrix that has diagonal elements equal to 1 and

$$l_{i,j} = -m_{i,j}^{(j)}$$

for all other elements with $i > j$. Multiplying L by the same sequence of matrices $M^{(k)}$, it is found that the result is the identity matrix. Thus L is the inverse of the product of the M matrices, and we have established that

$$A = LU.$$

Finding L and U is referred to as LU factorization. An example for a given 4×4 matrix is shown in Fig. 8.

The solution of the linear system $Ax = LUx = b$ can now be found by solving the two triangular systems,

$$Ly = b, \qquad Ux = y.$$

The number of additions and multiplications required for computing the elements of L and U is of the order of $\frac{1}{3}n^3$; thus for large n there is considerable savings in

$$\text{Let } A = A^{(0)} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{bmatrix}$$

$$A^{(1)} = M^{(1)}A^{(0)}$$

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 6 & 24 & 60 \\ 0 & 14 & 78 & 252 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{bmatrix}$$

$$A^{(2)} = M^{(2)}A^{(1)}$$

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 36 & 168 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -3 & 1 & 0 \\ 0 & -7 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 6 & 24 & 60 \\ 0 & 14 & 78 & 252 \end{bmatrix}$$

$$A^{(3)} = M^{(3)}A^{(2)}$$

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 0 & 24 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -6 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 36 & 168 \end{bmatrix}$$

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 1 & 7 & 6 & 1 \end{bmatrix} \qquad U = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 0 & 24 \end{bmatrix}$$

FIGURE 8 *LU* factorization of a 4 × 4 matrix.

performing the LU factorization once, then solving the triangular systems only for each new b for which the solution may be needed.

In solving an arbitrary $n \times n$ system it can occur that, at the kth stage, the divisor $a_{k,k}^{(k-1)}$ is zero, in which case, the solution can proceed no further. Also, if not identically zero, it may be very small, which can cause relatively large arithmetic roundoff errors to be introduced. To avoid these difficulties the technique of partial pivoting is employed, in which at each stage the largest element in the column headed by the pivotal element, $a_{k,k}^{(k-1)}$ is found, and the entire row in which this element occurs is exchanged with the kth row. Partial pivoting does not change the solution since it merely changes the order in which the equations of the system appear; however, the order of elements in the vector b must be adjusted accordingly.

Pivoting is not necessary when the matrix A is diagonally dominant: $|a_{i,i}| \geq \sum_j |a_{i,j}|$ for all i, or when it is symmetric and *positive definite*, meaning that for any n-vector x that is not identically zero, the quadratic form $x^{\mathrm{T}}\mathrm{Ax}$ is positive. If A is symmetric and pivoting is not necessary, a symmetric version of LU factorization is possible in which only the elements of $A^{(k)}$ above the diagonal are computed and the elements of the $M^{(k)}$ are not saved, thereby reducing the storage required by half and the number of operations to the order of $\frac{1}{6}n^3$. In this case, after the matrix U has been computed, the solution is found by solving

$$U^T y = b, \qquad Ux = Dy,$$

where D is the diagonal of U. This is often referred to as the (root-free) Cholesky method.

Systems that have a banded structure (the matrix A has zero elements wherever the difference between the indices i and j is greater than a fixed number less than $n-1$) retain their band width during LU factorization if pivoting is not required; thus advantages in storage and computation can be realized when working with them. Matrices that are sparse within a band of some nonzero elements, however, suffer from "zerofill," which means that no advantage can be taken of sparsity with direct methods, except that by judicious exchanging of equations and order of variables, the bandwidth or profile of the matrix can sometimes be reduced, thereby reducing the number of operations required.

Before considering the error in the solution to a system of equations (either linear or nonlinear), we need to have a means to measure a vector or matrix. Just in measuring

a scalar quantity by its magnitude, we measure a vector by its "norm," if x is a vector, we write $\|x\|$ to represent the norm of x. The most commonly used norm in numerical analysis is the "infinity norm," which is defined as: $\|x\|_\infty = \max_{1\le i\le n} |x_i|$. Other norms are the 2-norm: $\|x\|_2 = \sum_{i=1}^{n} x_i^2$, and the 1-norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$; however, these are more extensive to compute. The norm of a matrix A, that corresponds to a given vector norm is defined by: $\|A\| = \max_{\|x\|=1} \|Ax\|$.

The error introduced into the solution due to accumulated arithmetic roundoff depends on the condition of the system: that is, the degree to which small errors in the data of the problem are magnified during the computation—even assuming these computations are exact. A measure of the condition of a linear system is given by the condition number, which can be thought of as an error magnification factor; the larger the condition number, the larger will be the relative error in the solution. The condition number is defined to be the product of the norm of the matrix A with the norm of the inverse of A: $K(A) = \|A\|\,\|A^{-1}\|$. This requires the computation of a matrix inverse, which is more work than needed to solve the linear system in the first place; nevertheless, mathematical subroutines that are used to solve linear systems by direct means often give estimates of condition number.

Direct methods for the solution of linear systems of equations are exact in the sense that no error is introduced by the method; however, erroneous coefficients and right-hand sides of the equations will result in an error in the solution. Suppose that we have

$$(A + E)(x + h) = b + e,$$

where E is the matrix of errors in the coefficients and e is the vector of errors in the right hand sides, then we can show that the norm of the error h in the solution is bounded by

$$\frac{\|h\|}{\|x\|} \le \frac{K(A)}{1 - K(A)(\|E\|/\|A\|)}\left(\frac{\|e\|}{\|b\|} + \frac{\|E\|}{\|A\|}\right).$$

This inequality tells us that if A is well-conditioned, then small changes in the coefficients and right-hand sides, introduce correspondingly small changes in the solution. On the other hand, if A is ill-conditioned, small changes in the data may produce a relatively large error.

Assuming that the coefficients and right-hand side of the equations are exact, the only error that is introduced is due to the roundoff of intermediate results. Backward error analysis is used to bound the effect of roundoff errors in terms of equivalent errors in the coefficients; thus it can be shown that

$$\|E\| \le f(n)\mu \max_{i,j,k} \left|A_{i,j}^{(k)}\right|,$$

where μ is the machine unit. In practice, it has been found that when pivoting is employed, $f(n)$ is of order n, the number of equations in the system.

Before leaving this section we mention the problem of obtaining the "least squares" solution to an overdetermined system of linear equations: that is, the situation when there are more equations than unknowns. Such systems arise, for example, in smoothing observed data, as is discussed later. In this situation the system of equations is represented by

$$Ax = b,$$

where now, A is an m-row by n-column matrix with $m > n$, x is an n-vector, and b is an m-vector. There is, in general, no vector x that will satisfy this equation; however, we seek an x that minimizes the sum of the squares of the differences between the corresponding elements of Ax and b. It can be shown that such an x satisfies the equation

$$A^{\mathrm{T}}Ax = A^T b,$$

which is now an $n \times n$ system. Applying the LU factorization method to this new system is not recommended, however, because a more efficient and stable method of accomplishing the same purpose is to factor A into an $m \times n$ matrix, say Q, which has orthogonal columns (this means that $Q^{\mathrm{T}}Q = D$, an $n \times n$ diagonal matrix) and an $n \times n$ upper-triangular matrix R. This can be accomplished by premultiplying A by $(n-1)$ Householder matrices (see Section V). Thus we can obtain

$$Q^{\mathrm{T}}A = R,$$

or equivalently,

$$A = QR,$$

from which it can be established that the solution of

$$Rx = D^{-1}Q^{\mathrm{T}}b$$

is the desired least-squares solution.

C. Iteration Methods

These methods are based on the idea of splitting the matrix of a linear system into parts, which are then associated with either the left- or right-hand side of an iteration equation. For example, let

$$A = M + N.$$

Then the linear system

$$Ax = (M + N)x = b$$

can be expressed as

$$Mx = -Nx + b,$$

from which, providing M is nonsingular, we can obtain the matrix iteration equation

$$x^{(k+1)} = -M^{-1}Nx^{(k)} + M^{-1}b.$$

It can be shown that if the spectral radius (the magnitude of the largest eigenvalue) of the iteration matrix $M^{-1}N$ is less than 1, the iteration will converge for any nonzero starting estimate $x^{(0)}$.

We may split A into

$$A = E + D + F,$$

where E is the lower-triangular matrix containing the elements $e_{i,j} = a_{i,j}$ where $i > j$ and 0 otherwise, D is the diagonal matrix containing the elements $d_{i,i} = a_{i,i}$ for all i, and F is the upper-triangular matrix with elements $f_{i,j} = a_{i,j}$ when $i < j$ and 0.

The Jacobi method takes $M = D$ and $N = E + F$; thus,

$$x^{(k+1)} = -D^{-1}(E + F)x^{(k)} + D^{-1}b.$$

In terms of the individual elements of the vector x, this method can be expressed as

$$x_i^{(k+1)} = \frac{b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)}}{a_{i,i}}, \qquad i = 1, 2, \ldots, n.$$

Thus, each element in turn is updated in terms of elements of the previous iteration only. The Jacobi method is easy to implement but is usually rather slow to converge.

The Gausss–Seidel method uses $M = E + D$ and $N = F$, so that one iteration is equivalent to solving the triangular system

$$(E + D)x^{(k+1)} = -Fx^{(k)} + b.$$

In this case, the equation for the ith element of x is

$$x_i^{(k+1)} = \frac{b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)}}{a_{i,i}}, \qquad i = 1, 2, \ldots, n,$$

so that in updating each element, advantage is taken of the most recent value of previous elements. This method is a little more difficult to implement but usually converges more rapidly than the jacobi method.

Substantial improvement in the rate of convergence can often be achieved by introducing a relaxation factor ω, and using the iteration equation

$$(D + \omega E)x^{(k+1)} = [(1 - \omega)D - \omega F]x^{(k)} + \omega b.$$

If ω is greater than 1, this method is referred to as successive over-relaxation (SOR). In terms of the individual elements it can be expressed as the two-step process

$$r_i^{(k)} = \frac{b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)}}{a_{i,i}}$$

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + r_i^{(k)}, \qquad i = 1, 2, \ldots, n,$$

where ω is usually chosen to be a number between 1 and 2; however, it is often necessary to experiment to find a value that is close to optimum.

For systems in which the matrix A is symmetric, positive definite, and relatively sparse, a class of methods based on the minimization of the quadratic functional

$$\tfrac{1}{2}x^{\mathrm{T}}Ax - b^{\mathrm{T}}x$$

with respect to the vector x has been shown to be highly efficient. The most promising of these is the "conjugate gradient" method. Starting with an initial guess x^0, the previous functional is minimized in the direction of the residual $b - Ax^0$ (this is a scalar minimization problem). Thereafter, an iterative sequence of scalar minimization is performed in directions that are mutually orthogonal with respect to the inner product $x^{\mathrm{T}}Ay$ (i.e., any two directions represented by the vectors x and y are such that $x^{\mathrm{T}}Ay$ is zero). These directions are termed "conjugate" directions. It can be shown that such a series of iterations will converge in exactly n steps; however, in practice, fewer than n iterations are required to obtain a sufficiently accurate result. The basic algorithm is given next, in which we use the inner product $\langle x, y\rangle = x^{\mathrm{T}}y$, and $r^k \equiv b - Ax^k$ is the residual at the kth step.

Choose x^0, and set $p^0 = r^0$.

For $k = 0, 1, \ldots,$

$$\alpha_k = -\langle r^k, p^k\rangle / \langle p^k, Ap^k\rangle$$
$$x^{k+1} = x^k - \alpha_k p^k$$
$$r^{k+1} = r^k + \alpha_k Ap^k$$

if

$$\|r^{k+1}\| < \varepsilon,$$

stop; otherwise

$$\beta_k = \langle r^{k+1}, r^{k+1}\rangle / \langle r^k, r^k\rangle$$
$$p^{k+1} = r^{k+1} + \beta_k p^k$$

In practice, the conjugate gradient method is used in conjunction with some form of preconditioning of the matrix A. This is still an active area of research.

D. Systems of Nonlinear Equations

The general form of a system of n nonlinear equations in n unknown variables is

$$
\begin{aligned}
f_1(x_1, x_2, x_3, \ldots, x_n) &= 0 \\
f_2(x_1, x_2, x_3, \ldots, x_n) &= 0 \\
&\vdots \\
f_n(x_1, x_2, x_3, \ldots, x_n) &= 0
\end{aligned}
$$

The iteration methods discussed in the last section can also be used for the nonlinear system. For example, the Jacobi method consists of making an initial estimate; then for the $(k+1)$st iteration, the ith individual equation is solved for $x_i^{(k+1)}$ in terms of the kth iteration values of all other variables. In the Gauss–Seidel method, the ith equation is solved for $x_i^{(k+1)}$ in terms of the already calculated values of $x^{(k+1)}$ from the current iteration and the rest of the variables from the previous iteration. An analogous situation exists for the successive over-relaxation method. The solution of each individual nonlinear equation is itself an iterative process, as discussed in Section III; thus for large systems there is a large amount of work. Also, convergence will only occur if the initial guess is sufficiently close to the solution.

The Newton method can be used if the partial derivatives of each equation with respect to each variable can be readily computed. This method is given by the matrix iteration equation

$$x^{(k+1)} = x^{(k)} - (J^{(k)})^{-1} f^{(k)},$$

where $x^{(k)}$ is the vector of kth iteration values of the n variables, $J^{(k)}$ is the"Jacobian" matrix of partial derivatives

$$
J^{(k)} = \begin{bmatrix}
\partial f_1/\partial x_1 & \partial f_1/\partial x_2 & \cdots & \partial f_1/\partial x_n \\
\partial f_2/\partial x_1 & \partial f_2/\partial x_2 & \cdots & \partial f_2/\partial x_n \\
\vdots & & & \vdots \\
\partial f_n/\partial x_1 & \partial f_n/\partial x_2 & \cdots & \partial f_n/\partial x_n
\end{bmatrix}
$$

evaluated at the kth iteration values of the variables, and $f(x^{(k)})$ is the vector of function values, again evaluated at the kth iteration. Note that the inverse of J is not explicitly computed, but that the system of equations

$$J^{(k)} y = f\left(x^{(k)}\right)$$

is solved. The Newton method converges more rapidly than those given above, but at the expense of computing derivatives. A number of variations of this method have been devised that approximate and rapidly update J^{-1}.

V. MATRIX EIGENPROBLEMS

A. Background

The multiplication of an $n \times n$ matrix A times an n-vector x results in a second n-vector,

$$Ax = y.$$

Thus, the operation can be thought of as a vector transformation. In general, both the length (the square root of the sum of the squares of the elements) and the orientation (relative magnitude of the elements) of the vector change. For a given matrix A, however, there are some vectors that, when premultiplied by A, do not change in orientation but only in length. That is,

$$Ax = \lambda x,$$

where λ is a scalar multiplier. Another way of stating this is

$$(A - \lambda I)x = 0.$$

This is a system of linear equations that has a nontrivial solution (not all components zero) only if $(A - \lambda I)$ is a singular matrix, that is, the determinant $\det(A - \lambda I)$ is zero. It is readily shown that $\det(A - \lambda I)$ is a polynomial of degree n in λ that has exactly n roots (allowing for multiple roots). The values of these roots are called the eigenvalues of A. For each eigenvalue λ_i, $i = 1, 2, \ldots, n$, there is a nontrivial vector $x^{(i)}$ that satisfies the equation

$$Ax^{(i)} = \lambda_i x^{(i)}$$

and any such vector is called an eigenvector corresponding to λ_i. Eigenvectors are not unique; in fact, any scalar multiple of an eigenvector is also an eigenvector corresponding to the same eigenvalue. Eigenvectors that correspond to distinct eigenvalues, however, are linearly independent.

The problems of finding eigenvalues or eigenvectors of a matrix are called eigenproblems. There are also "generalized eigenproblems," which are to find values λ or vectors x that satisfy

$$Ax = \lambda Bx$$

for any second matrix B.

Clearly, one approach for finding eigenvalues is to seek the roots of the polynomial $\det(A - \lambda I)$; however, this can be excessively time-consuming and unstable (meaning that a small error in the calculation can lead to a large error in a computed root). This approach is used only for selected eigenvalues when the matrix is symmetric and tridiagonal, in which case the polynomial can be evaluated without computing the coefficients.

The choice of a method for a specific eigenproblem depends on the properties of the matrix, whether just eigenvalues or both values and vectors are required, and whether

all or selected values are sought. The eigenvalues of symmetric matrices are real, whereas those of nonsymmetric matrices may be complex. Less computation is required for symmetric matrices, and the process is inherently more stable. Advantage can be taken of a banded structure. Finally, if the elements of A are complex, either the problem must be converted to a larger real problem, or complex arithmetic must be employed.

Most practical methods for computing eigenvalues depend on finding a similarity transformation,

$$B = C^{-1}AC,$$

where C is a nonsingular matrix that transforms A into a matrix B, whose eigenvalues can be more readily found. A similarity transformation preserves eigenvalues; that is, the eigenvalues of B are the same as those of A. In particular, if B is either diagonal or triangular, its eigenvalues are the diagonal elements.

B. Symmetric Matrices

There are two general classes of methods for finding the eigenvalues of a symmetric matrix. The first (Jacobi methods) attempts to reduce the matrix to diagonal form by iterating with similarity transformations. The second begins by reducing the matrix to tridiagonal form with a finite number of similarity transformations and then follows an iterative procedure.

The premultiplication of a two-dimensional vector by a matrix of the form

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

has the effect of rotating the vector counterclockwise through the angle θ. This matrix is therefore called a rotation matrix. By a judicious choice of θ, we can rotate the vector into a vertical orientation; that is, with its second element zero. In particular, if we consider the symmetric matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$$

and select $\theta = \tan^{-1}\frac{-a_{12}}{a_{11}}$, the result of premultiplying A by the rotation matrix is a matrix with zero in the lower left corner

$$\frac{1}{\sqrt{a_{11}^2 + a_{12}^2}} \begin{bmatrix} a_{11} & a_{12} \\ -a_{12} & a_{11} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$$

$$= \frac{1}{\sqrt{a_{11}^2 + a_{12}^2}} \begin{bmatrix} a_{11}^2 + a_{12}^2 & a_{11}a_{12} + a_{12}a_{22} \\ 0 & a_{11}a_{22} - a_{12}^2 \end{bmatrix}.$$

In order to preserve the eigenvalues of A, we must now complete the similarity transformation by multiplying on the right by the inverse of the rotation matrix (which has the same elements except that the signs of the off-diagonal elements are reversed). This results in the matrix

$$A^{(1)} = \frac{1}{a_{11}^2 + a_{12}^2} \times \begin{bmatrix} a_{11}^3 + 2a_{11}a_{12}^2 + a_{12}^2a_{22} & a_{11}a_{12}a_{22} - a_{12}^3 \\ a_{11}a_{12}a_{22} - a_{12}^3 & a_{11}^2a_{22} - a_{11}a_{12}^2 \end{bmatrix}.$$

The off-diagonal elements of $A^{(1)}$ are smaller in magnitude than those of A. Thus it is closer to a diagonal matrix. An iteration in which this process is repeated will converge to a diagonal matrix and the eigenvalues of A will appear on the diagonal.

For a symmetric matrix of arbitrary order, the rotation matrix required to annihilate the element of A appearing below the diagonal in the ith row and jth column is obtained by replacing the (i, i) and (j, j) elements of the identity matrix by $a_{ii}/\sqrt{a_{ii}^2 + a_{ij}^2}$, the (i, j) element by $-a_{ij}/\sqrt{a_{ii}^2 + a_{ij}^2}$, and the (j, i) element by $a_{ij}/\sqrt{a_{ii}^2 + a_{ij}^2}$. The similarity transformation involving this rotation will affect only the elements of A in these four positions. For a single iteration, a similarity transformation must be performed for each position below the diagonal until the magnitude of the element in that position falls below a predetermined threshold.

On a serial computer, the amount of work required to perform the Jacobi method and its variations exceeds that of the tridiagonalization methods described next; however, the parallelism of multiprocessor computers can be exploited more readily by the Jacobi methods. Furthermore, it has been found that the effect of roundoff error is less severe in these methods. For these reasons, they have been receiving greater attention in recent years.

The first step in the tridiagonilization method is to reduce a symmetric matrix to a tridiagonal matrix having the same eigenvalues. Although all of the off-diagonal elements of a matrix cannot be annihilated by similarity transformations in a finite number of operations, all of the elements outside of the diagonal and the adjacent super-and sub-diagonals can be. The use of rotation matrices, as above, to do this is called the "Givens" method of tridiagonalization; however, a more efficient method that annihilates all of the elements in a column below the tridiagonal in a single similarity transformation can be accomplished through the use of "Householder" matrices that have the form

$$P = (I - 2ww^{\mathrm{T}}).$$

In two dimensions, premultiplying a vector by a Householder matrix has the effect of reflecting the vector

about the line defined by w, thus these matrices are also called reflection matrices.

The vector w is chosen to have unit length, $(w^T w = 1)$, which implies that $P^{-1} = P$. If $A^{(0)}$ is taken to be A, the reduction to tridiagonal from takes $n-2$ operations.

$$A^{(k)} = P^{(k)} A^{(k-1)} P^{(k)}, \quad k = 1, 2, \ldots, n-2,$$

where the w in the matrix $P^{(k)}$ is chosen to introduce zero elements in $A^{(k)}$ below the subdiagonal in column k and to the right of the superdiagonal in row k. Then $A^{(n-2)}$ is a tridiagonal matrix, which we will refer to below as $T^{(0)}$ and which has the same eigenvalues as A.

The second step depends on whether most or a few of the eigenvalues are required. If more than, say, 40% of the eigenvalues are required, then it is usual to proceed with a variant of the QR method. The kth iteration of this method consists of first factoring the matrix $T^{(k)}$ into an orthogonal matrix $Q^{(k)}$ (a matrix whose transpose is its inverse) and an upper triangular matrix $R^{(k)}$, and then multiplying these two matrices in reverse order. Thus,

$$T^{(k)} = Q^{(k)} R^{(k)}$$
$$T^{(k+1)} = R^{(k)} Q^{(k)} = Q^{(k)\mathrm{T}} T^{(k)} Q^{(k)}$$

The factorization is usually accomplished through the Gram–Schmidt orthogonalization process (or a modification thereof) on the columns of $T^{(k)}$. Each step of the method is a similarity transformation of the matrix of the previous step, thus the sequence of matrices $T^{(0)}, T^{(1)}, T^{(2)}, \ldots$ has the same eigenvalues. The sequence will converge to a diagonal matrix with elements that are those eigenvalues. The convergence can be accelerated by judicious selection of a scalar quantity σ_k, called an origin shift, which is subtracted from the diagonal elements at each step. The algorithm then becomes

$$T^{(k)} - \sigma_k I = Q^{(k)} R^{(k)}$$
$$T^{(k+1)} = R^{(k)} Q^{(k)} + \sigma_k I,$$

where k takes on values 0, 1, 2, ... Once a number sufficiently close to zero has appeared in the $t_{n,n-1}$ position, the algorithm can be applied to the $(n-1) \times (n-1)$ upper left submatrix and the origin shift adjusted accordingly, and so on until all subdiagonal elements have been reduced to zero.

Sometimes this method is implemented as the QL algorithm, in which the second factor is made to be a lower-triangular matrix instead of an upper-triangular one. The QL matrix has an advantage in convergence when, as often happens in the formulation of matrices from physical situations, the smaller elements are at the top.

The transformation matrices for all steps in the process can be accumulated to obtain the eigenvectors of A, if they are needed.

If only a few eigenvalues are required, they may be found more efficiently by computing the corresponding roots of the characteristics polynomial of the tridiagonal matrix obtained after the first stage of the above method. This is done iteratively without computing the coefficients of the polynomial, but estimates of root locations are obtained from the theory of Sturm sequences of polynomials.

The problem of computing the eigenvector corresponding to a specific eigenvalue may be solved very effectively through the process of inverse iteration: an estimate of the eigenvalue is subtracted from the diagonal elements of A to form the matrix $(A - \lambda I)$, and then the solutions to the sequence of linear systems

$$(A - \lambda I) x^{(k+1)} = x^{(k)}, \qquad k = 0, 1, 2, \ldots,$$

are found successively. (The LU factorization of $A - \lambda I$ need be performed only once.) This iteration is relatively insensitive to the first estimate $x^{(0)}$ and will converge to the required eigenvector.

C. Nonsymmetric Matrices

The problem of finding the eigenvalues of a nonsymmetric matrix is inherently less stable than the symmetric case; furthermore, some or all of the eigenvalues and eigenvectors may be complex.

The first stage is to "balance" the matrix by applying similarity tranformations in order to make the maximum element in corresponding rows and columns approximately equal. Next, the matrix is reduced to "Hessenburg form" by applying stablilized elementary similarity transformations. An upper Hessenburg matrix has the form

$$H = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots & a_{2,n} \\ 0 & a_{3,2} & a_{3,3} & \cdots & a_{3,n} \\ 0 & 0 & a_{4,3} & \cdots & a_{4,n} \\ \cdot & & & \cdots & \cdot \\ \cdot & & & \cdots & \cdot \\ \cdot & & & \cdots & \cdot \\ 0 & 0 & 0 & \cdots a_{n,n-1} & a_{n,n} \end{bmatrix}$$

The elementary matrices are similar in form to those used in the LU factorization (Gaussian elimination) process, except that the pivotal element is taken to be one element below the diagonal. As in LU factorization for general matrices, the rows are interchanged to maximize the pivotal element; however, a difference is that for each elementary matrix multiplying A on the left, its inverse is multiplied on the right, and each row interchange is accompanied by the corresponding column interchange, in order to preserve the eigenvalues.

Since real nonsymmetric matrices may have complex conjugate pairs of eigenvalues, an application of the QR algorithm with origin shift as described for symmetric matrices will involve computing with complex numbers. To avoid this, the double-shift QR algorithm is usually applied to the Hessenburg matrix $H^{(0)}$ obtained in the first stage. The kth iteration of this method is

$$(H^{(k)} - \sigma_{k,1} I)(H^{(k)} - \sigma_{\sigma_{k,2}} I) = Q^{(k)} R^{(k)}$$
$$H^{(k+1)} = Q^{(k)T} H^{(k)} Q^{(k)}$$

The shifts $\sigma_{k,1}$ and $\sigma_{k,2}$ are taken to be the eigenvalues of the lowest 2×2 submatrix of $H^{(k)}$; however, when convergence of this submatrix is ascertained, the method is applied to the remaining upper submatrix in the same way. Eventually the sequence $H^{(k)}$, $k = 0, 1, 2, \ldots$, will converge to a triangular matrix, except that there will be 2×2 blocks on the diagonal for each pair of complex eigenvalues.

As before, the transformations may be accumulated if all of the eigenvectors are required. If the eigenvector associated with a specific eigenvalue is required, inverse iteration may be used (albeit with complex arithmetic if the eigenvalue is complex).

VI. NUMERICAL REPRESENTATION OF FUNCTIONS

A. Polynomial Representation and Interpolation

The exact representation of an arbitrary continuous function would require an infinite number of infinite-precision numbers. If a function is to be numerically represented for computer manipulation, it must therefore be approximated in some way.

We will consider an arbitrary function $f(x)$ in the single independent variable x, which can take on values between a and b. The general approach is to let $f(x)$ be approximated by a finite linear combination of basis functions

$$f(x) \approx \sum_{i=0}^{m} a_i \phi_i(x),$$

where the a_i are fixed coefficients and the $\phi_i(x)$ are, usually, simple functions defined for $a \leq x \leq b$. For example, if $\phi_i(x) = (x - x_0)^i$, $i = 0, 1, 2, \ldots m$, successive differentiation and evaluation at $x = x_0$ yields the familiar Taylor's series, truncated after the $(m+1)$-st term

$$f(x) \approx f(x_0) + (x - x_0) f'(x_0) + \left[(x - x_0)^2/2\right] f''(x_0) + \cdots + \left[(x - x_0)^m/m!\right] f^{(m)}(x_0)$$

If, instead, we identify $m+1$ distinct values of $x(x_0, x_1, x_2, \ldots, x_{m-1}, x_m)$; take $\phi_0(x) = 1$; set

$$\phi_i(x) = (x - x_0)(x - x_1)(x - x_2) \cdots (x - x_{i-1})$$

for $i = 1, 2 \ldots m$; and solve for the coefficients a_i, $i = 0, 1, 2, \ldots, m$, in turn, we arrive at the Newton general interpolation formula

$$f(x) \approx f[x_0] + f[x_0, x_1]\phi_1(x) + f[x_0, x_1, x_2]\phi_2(x) + \cdots + f[x_0, x_1, \ldots, x_m]\phi_m(x),$$

where the $f[x_0, x_1, x_2, \ldots x_i]$ for $i = 0, 1, 2, \ldots, m$ are called divided differences and may be calculated recursively, as illustrated in Table III.

The right-hand sides of both of these examples are polynomials of degree m. In the first case, the polynomial takes on the value of $f(x)$ and all of its derivatives up to $f^{(m)}(x)$

TABLE III Divided Differences

x	Zeroth order	First order	Second order	Third order
x_0	$f[x_0] = f(x_0)$			
		$f[x_0, x_1] = \dfrac{f[x_1] - f[x_0]}{(x_1 - x_0)}$		
x_1	$f[x_1] = f(x_1)$		$f[x_0, x_1, x_2] = \dfrac{f[x_1, x_2] - f[x_0, x_1]}{(x_2 - x_0)}$	
		$f[x_1, x_2] = \dfrac{f[x_2] - f[x_1]}{(x_2 - x_1)}$		$f[x_0, x_1, x_2, x_3] = \dfrac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{(x_3 - x_0)}$
x_2	$f[x_2] = f[x_2]$		$f[x_1, x_2, x_3] = \dfrac{f[x_2, x_3] - f[x_1, x_2]}{(x_3 - x_1)}$	
		$f[x_2, x_3] = \dfrac{f[x_3] - f[x_2]}{(x_3 - x_2)}$		$f[x_1, x_2, x_3, x_4] = \dfrac{f[x_2, x_3, x_4] - f[x_1, x_2, x_3]}{(x_4 - x_1)}$
x_3	$f[x_3] = f(x_3)$		$f[x_2, x_3, x_4] = \dfrac{f[x_3, x_4] - f[x_2, x_3]}{(x_4 - x_2)}$	
		$f[x_3, x_4] = \dfrac{f[x_4] - f[x_3]}{(x_4 - x_3)}$		
x_4	$f[x_4] = f(x_4)$			

when $x = x_0$, whereas in the second case, the polynomial takes on the value of $f(x)$ at $m+1$ different values of x. It is possible, by repeating the occurrence of the factor $(x - x_i)$ the proper number of times for any of the x_i in the construction of the basis functions of the second example, to provide for the polynomial to take on the value of $f(x_i)$ and any number of derivatives of $f(x)$ at x_i.

Polynomials constructed in this way both represent $f(x)$ numerically and can be used to interpolate $f(x)$: that is, to estimate a value of $f(x)$ (by evaluation of the polynomial) between exactly known values. In fact, as implied by the first example they may be used to extrapolate $f(x)$, meaning to estimate a value outside the interval containing all of the known values of the function.

Provided that $f(x)$ can be differentiated $m+1$ times, the error in estimating $f(x)$ for, say, $x = x^*$ can be shown to be no greater in magnitude than

$$(M/(m+1)!)\phi_{m+1}(x^*),$$

where M is the maximum value of $|f^{(m+1)}(x)|$ in the interval containing x^* and all of the x_i. This error may be small for x^* close to the center of the interval containing all the x_i, and, for this reason polynomial representation of a function using known data *local* to the point of interpolation is generally satisfactory; however, as the interpolation point approaches the extremes of the interval containing the known data, the error can become quite large. Extrapolation using high-degree polynomial representation of a function can be highly inaccurate and is not recommended.

B. Piecewise Polynomials and Splines

A commonly used approach to overcome the problem of representing a given function globally, where the use of a single high-degree polynomial will produce large errors away from the central region, is to divide the domain of the function into segments. The function is represented by a different polynomials in each segment, and these polynomials are constrained to at least have the same value at the segment boundaries. The points where the polynomials meet are called knots.

In the case of interpolating between distinct values x_i where the function $f(x_i)$ is known, it is usual to make the knots coincident with the points $(x_i, f(x_i))$. The simplest example is to construct a piecewise linear function by joining successive points with straight lines. Such a function is continuous over its length, but, of course, may not have a continuous derivative at the knots. If the derivatives $f'(x_i)$ are known, then in each interval defined by $x_{i-1} \le x \le x_i$, the four coefficients of a cubic polynomial

$$h_i(x) = a_i + b_i x + c_i x^2 + d_i x^3$$

can be computed to satisfy

$$\begin{aligned} h_i(x_{i-1}) &= f(x_{i-1}) \\ h_i(x_i) &= f(x_i) \\ h_i'(x_{i-1}) &= f'(x_{i-1}) \\ h_i'(x_i) &= f'(x_i). \end{aligned}$$

This is referred to as cubic Hermite interpolation and results in a piecewise cubic polynomial that is continuous and has a continuous first derivative. This concept can be extended if derivatives higher than the first are known; the interpolating polynomial is always of degree $2p+1$, where p is the highest derivative known.

The cubic spline is a very elegant mathematical function, named for the draftsman's spline curve that it models. If the points $(x_i, f(x_i))$ are given for $i = 0, 1, 2, \ldots, m$, then the $4m$ coefficients required to define a cubic polynomial s_i in each of the m intervals are computed from the $4m$ equations

$$\begin{aligned} s_i(x_{i-1}) &= f(x_{i-1}), && i = 1, 2, \ldots, m \\ s_i(x_i) &= f(x_i), && i = 1, 2, \ldots, m \\ s_{i+1}'(x_i) &= s_i'(x_i), && i = 1, 2, \ldots, m-1 \\ s_{i+1}''(x_i) &= s_i''(x_i), && i = 1, 2, \ldots, m-1 \\ s''(x_0) &= s''(x_m) = 0. \end{aligned}$$

The last two equations, which are the end conditions for the *natural* cubic spline, may be replaced by other conditions, such as $s'(x_0) = f'(x_0)$, $s'(x_m) = f'(x_m)$. The resulting function is continuous and has continuous first and second derivatives. This concept can be extended to other polynomial degrees, and the spline is continuous in all derivatives up to 1 less than the degree; however, for an equal number of conditions at each end, the degree must be odd. Figure 9(a) illustrates interpolation with a natural cubic spline function.

As in the case of polynomials, there is always a representation for an arbitrary piecewise polynomial $g(x)$ in the form

$$g(x) = \sum_{j=1}^{n} a_j \phi_j(x).$$

There are always as many basis functions $\phi_i(x)$ as there are matching conditions plus end conditions for the entire curve, for example, $n = m+3$ in the case of the cubic spline. There is at least one basis function associated with each knot, and they have the desirable property of local support, meaning that they are zero everywhere except close to the knot with which they are associated. Figure 9(b) illustrates the basis functions for a natural cubic spline having equally spaced knots.

The error associated with interpolation using piecewise polynomials is dependent on the degree of the polynomial and the spacing between the knots. If h is the maximum

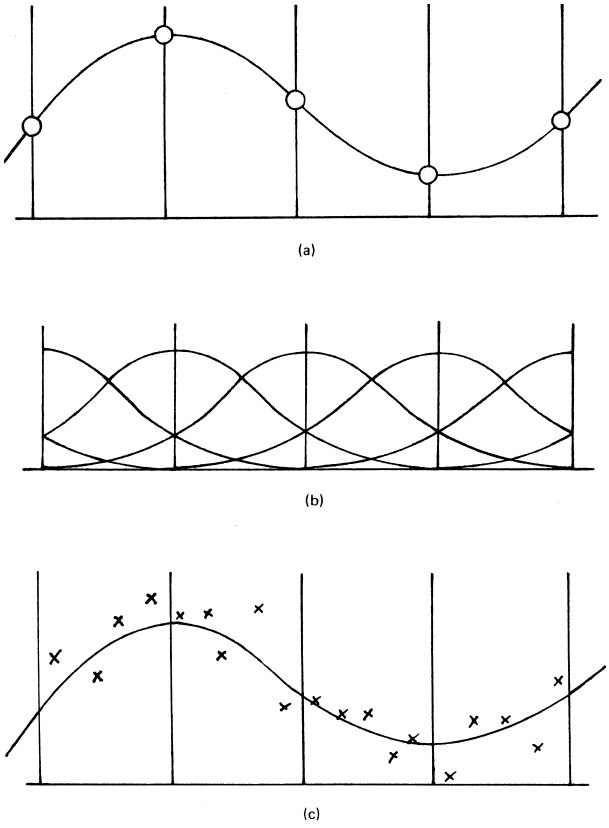


FIGURE 9 Natural cubic spline function. (a) Interpolation. (b) Basis functions. (c) Approximation.

spacing between knots and p is the degree of the polynomial, the error for an arbitrary x between x_0 and x_m is bounded by

$$|f(x) - g(x)| < Kh^{p+1},$$

where K is a constant. The error is thus said to be of order h^{p+1} and written $O(h^{p+1})$.

Even using unmodified piecewise polynomials often results in numerical representations with undesirable oscillations, or in regions where, for example, the numerical curve is convex where $f(x)$ is concave. In an attempt to overcome these problems, special functions, such as rational polynomials (ratios of polynomials) or splines with tension factors (a generalization of the spline already described that contains a parameter whose effect is to flatten the curve between the knots in the same manner as applying longitudinal tension in a mechanical spline), have been introduced in recent years. The general principle, however, is the same.

C. Approximation

Frequently, the function $f(x)$ that is to be represented numerically describes a physical situation and is known in terms of observed data x_i, $f(x_i)$, for $i = 1, 2, \ldots, m$. If this data is voluminous, the interpolating polynomials described in the last two sections require more storage than the raw data. Furthermore, any observational error in the data will tend to introduce (sometimes extreme) oscillations. It is therefore often desirable to approximate such functions with polynomials of lower degree than indicated by the number of data points, or with piecewise polynomials with widely spaced knots compared to the data intervals, but which do not necessarily take on the exact given values of $f(x_i)$. The usual technique is to attempt to find a function with these characteristics, say $g(x)$, that minimizes the sum (over all i) of the squares of the residuals,

$$r_i = f(x_i) - g(x_i).$$

A useful first step for this purpose is to determine basis functions ϕ_j for the approximating function

$$g(x) = \sum_{j=1}^{n} \alpha_j \phi_j(x).$$

In the case of piecewise polynomials, this implies choosing the knots, which are more widely spaced than the data, and not necessarily coincident with any data point. The n unknown coefficients α_j are then found by solving the overdetermined system of equations

$$\sum_{j=1}^{n} \alpha_j \phi_j(x_i) = f(x_i), \qquad i = 1, 2, \ldots, m,$$

by the least-squares method described in Section IV.B. The concept is illustrated in Fig. 9(c) for a natural cubic spline function.

In practice, numerical representation, interpolation, and approximation generally require experimentation with the degree of polynomials, placement of knots (in the case of piecewise functions), and tension factors (in the case of splines under tension), in order to obtain a fair curve. Interactive techniques using computer-generated graphics are particularly useful for this purpose.

The concepts discussed in this section carry over into multidimensional regions for the representation of functions of more than one independent variable or the parametrically defined surfaces of solid objects.

VII. DIFFERENTIATION AND INTEGRATION

A. Differentiation and Finite Differences

Once a numerical representation $g(x)$ of a function $f(x)$ has been found, it is relatively simple to differentiate it in order to obtain an approximation to the derivative of $f(x)$, that is,

$$f'(x) \approx g'(x) = \sum_{i=0}^{m} a_i \phi_i'(x).$$

Care must be exercised in doing this, however, because the error in approximating the derivative can be very much larger than the error in approximating the function itself. This can be seen by differentiating the error $f(x) - g(x)$ in any of the cases previously considered, from which it is found that the bound on the error in the derivative is one order lower than the corresponding error in the function. For example, if

$$|f(x) - g(x)| < Kh^p,$$

then

$$|f'(x) - g'(x)| < K'h^{p-1},$$

where K' is, in general, a different constant than K. When h is small, the second bound can be much larger than the first.

Nevertheless, some very useful formulas can be obtained from the Newton general interpolation formula introduced in Section VI.A. Differentiation with respect to x produces a very complicated formula, but considerable simplification can be achieved by taking the points x_i to be equally spaced along the x axis (say a distance h units apart) and by evaluating the result at one of these points. The final result is an expression for the derivative of $f(x)$ at one point in terms of h and values of $f(x)$ evaluated at multiples of h units away from that point. Such expressions, and similar ones for higher derivatives, are called finite-difference formulas.

The order in which the points are taken is actually arbitrary; thus, for $x = x_0$, the natural left-to-right ordering produces a forward-difference formula, and the reverse ordering produces a backward-difference formula. If the number of points is odd and $x = x_0$, two difference formulas can be obtained by alternating $x_1, x_2 \ldots$ to the left and right of x_0, starting first to the left and then first to the right. Averaging these two formulas results in a symmetric central-difference formula.

If the error term of the Newton interpolation formula is carried along in the derivation of these difference formulas, an estimate of the error can be obtained in each case. The power of h that appears in the error term determines the order of accuracy of the corresponding formula.

Table IV gives the coefficients and error estimate for second-, fourth-, and sixth-order accurate central-difference approximations of $f'(x_0)$ and $f''(x_0)$; and second- and fourth-order accurate approximations to $f^{(3)}(x_0)$ and $f^{(4)}(x_0)$. For example, a second-order accurate finite-difference approximation to the first derivative of $f(x)$ at $x = x_0$ is

$$\left[-\tfrac{1}{2} f(x_0 - h) + \tfrac{1}{2} f(x_0 + h) \right] / h.$$

Note that z, appearing in the error term, is some point in the domain of the corresponding finite-difference formula.

B. Integration

The interpolation formulas introduced in Sections VI.A and VI.B can also be used as starting points to develop formulas for numerical integration (sometimes called quadrature). Given a function $f(x)$ and values for lower and upper limits of integration a and b, the objective is to obtain approximations to

$$I = \int_a^b f(x)\, dx.$$

The domain of integration is partitioned into n intervals by the points $x_0 = a, x_1, x_2, \ldots, x_n = b$, where the ith interval is defined by $x_{i-1} \le x \le x_i$. The simplest method, which corresponds to $m = 0$ in the Newton interpolation formula, is to approximate $f(x)$ by the constant value $f(x_{i-1})$ in the ith interval. Thus

$$I_i = \int_{x_{i-1}}^{x_i} f(x)\, dx \approx (x_i - x_{i-1}) f(x_{i-1})$$

and

$$I = \sum_{i=1}^{n} I_i \approx \sum_{i=1}^{n} (x_i - x_{i-1}) f(x_{i-1}).$$

This method is referred to as rectangle integration and is illustrated graphically in Fig. 10(a). If all of the intervals are of equal length h, it can be established by integrating the error term in the interpolation formula that the error in I is of order h [bounded in absolute value by Kh, where K is a constant depending only on $f(x)$].

A more accurate method is derived in similar fashion by taking $m = 1$ in the Newton interpolation formula. In this case it is found that

$$I_i \approx (x_i - x_{i-1})[f(x_{i-1}) + f(x_i)]/2$$

and

$$I \approx [(x_i - x_0) f(x_0) + \sum_{i=1}^{n-1} (x_{i+1} - x_{i-1}) f(x_i) + (x_n - x_{n-1}) f(x_n)]/2.$$

This method is known as trapezoidal integration and is illustrated in Fig. 10(b). The error for equal intervals of length h is of order h^2.

Carrying this concept one step further and assuming that $f(x)$ may also be evaluated at the midpoint of each interval, a method called Simpson rule integration, which is of order h^4 for equal intervals of length h, can be derived from a symmetric version of the interpolation formula with $m = 2$. Here

TABLE IV Coefficients of Central-Difference Approximations and Error Estimates for Derivatives

First derivative								
Order	$x-3h$	$x-2h$	$x-h$	x	$x+h$	$x+2h$	$x+3h$	Error
2			$-\frac{1}{2}$	0	$\frac{1}{2}$			$-\frac{1}{6}f^{(3)}(z)h^2$
4		$\frac{1}{12}$	$-\frac{2}{3}$	0	$\frac{2}{3}$	$-\frac{1}{12}$		$\frac{1}{30}f^{(5)}(z)h^4$
6	$-\frac{1}{60}$	$\frac{3}{20}$	$-\frac{3}{4}$	0	$\frac{3}{4}$	$-\frac{3}{20}$	$\frac{1}{60}$	$-\frac{1}{140}f^{(7)}(z)h^6$
Second derivative								
Order	$x-3h$	$x-2h$	$x-h$	x	$x+h$	$x+2h$	$x+3h$	Error
2			1	-2	1			$-\frac{1}{12}f^{(4)}(z)h^2$
4		$-\frac{1}{12}$	$\frac{4}{3}$	$-\frac{5}{2}$	$\frac{4}{3}$	$-\frac{1}{12}$		$\frac{1}{40}f^{(6)}(z)h^4$
6	$\frac{1}{90}$	$-\frac{3}{20}$	$\frac{3}{2}$	$-\frac{49}{18}$	$\frac{3}{2}$	$-\frac{3}{20}$	$\frac{1}{90}$	$-\frac{1}{560}f^{(8)}(z)h^6$
Third derivative								
Order	$x-3h$	$x-2h$	$x-h$	x	$x+h$	$x+2h$	$x+3h$	Error
2		$-\frac{1}{2}$	1	0	-1	$\frac{1}{2}$		$-\frac{1}{4}f^{(5)}(z)h^2$
4	$\frac{1}{8}$	-1	$\frac{13}{8}$	0	$-\frac{13}{8}$	1	$-\frac{1}{8}$	$\frac{7}{120}f^{(7)}(z)h^4$
Fourth derivative								
Order	$x-3h$	$x-2h$	$x-h$	x	$x+h$	$x+2h$	$x+3h$	Error
2		1	-4	6	-4	1		$-\frac{1}{6}f^{(6)}(z)h^2$
4	$-\frac{1}{6}$	2	$-\frac{13}{2}$	$\frac{28}{3}$	$-\frac{13}{2}$	2	$-\frac{1}{6}$	$\frac{7}{240}f^{(8)}(z)h^4$

$$I_i \approx (x_i - x_{i-1})[f(x_{i-1}) + 4f((x_{i-1}+x_i)/2 + f(x_i)]/6$$

and

$$I \approx \left[(x_1 - x_0)f(x_0) + \sum_{i=1}^{n}(x_{i+1} - x_{i-1})f(x_i) + 4\sum_{i=1}^{n-1}(x_i - x_{i-1})f((x_{i-1}+x_i)/2) + (x_n - x_{n-1})f(x_n)\right] \Big/ 6$$

This is illustrated in Fig. 10(c).

In the situation where $f(x)$ can be evaluated for any given x, the choice of the partition points x_i in the above methods can be optimized by adaptive procedures. One such method, called adaptive Romberg integration, uses the trapezoidal method on a single interval equal to the whole range of integration to form, say $T_{0,0}$, and then with two intervals for the first and second half of the range to form, $T_{0,1}$. These two results are combined into

$$T_{1,1} = (4T_{0,1} - T_{0,0})/3,$$

which has the effect of eliminating the principal part of the error. The difference between the two results is also compared with a predetermined tolerance, and if it is sufficiently small $T_{1,1}$ is taken as the final result. If not, each of the two intervals is again halved and the procedure is repeated for both intervals. The pattern is continued with successive halving of the intervals, except that whenever two successive results for a particular interval are sufficiently close, that interval is no longer reduced and the current value of the integral over that interval is held constant

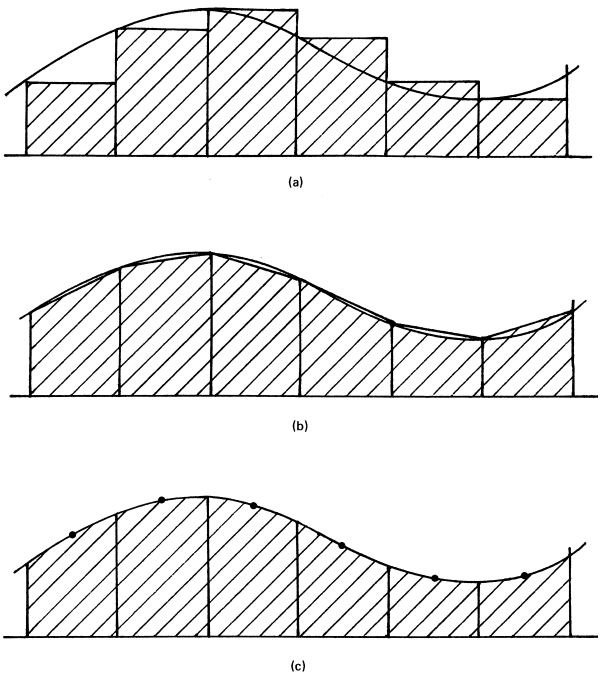


FIGURE 10 Numerical integration. (a) Rectangle method. (b) Trapezoidal method. (c) Simpson's method.

until the final result is obtained. At each stage of mesh refinement, a combination of results can be used to further reduce the error in a manner similar (but not necessarily identical) to $T_{1,1}$.

In situations where $f(x)$ cannot be evaluated but is tabulated for given values of x, integration of the spline interpolation formulas or of the approximation formulas obtained in the previous section may be used.

It should be noted that, in contrast to numerical differentiation, the order of a numerical integration method is greater than that of the interpolation or approximation formula from which it is derived. This is a reflection of the fact that integration is a well-conditioned process, whereas differentiation is ill-conditioned.

VIII. DIFFERENTIAL EQUATIONS

A. Ordinary Differential Equations—Initial-Value Problems

A first-order ordinary differential equation (ODE) in the independent variable x is usually written as

$$y'(x) = f(x, y(x)),$$

where f is a given function of the variables x and y, and y is an unknown function of x. Such an equation is to hold for a domain of x, say $a \le x \le b$. There are, in general, an infinite number of functions $y(x)$ that satisfy this equation, but under some rather mild restrictions on f, there is a unique solution that also satisfies the initial condition.

$$y(a) = y_0.$$

In realistic situations, it is usually impossible to find an analytic expression for the solution; thus a numerical approximation is often sought.

The general approach is to compute approximations to $y(x)$ at monotone increasing, discrete values of x, that is, $x_1, x_2, \ldots, x_i, \ldots$, where $x_{i+1} - x_i = h_i$ is a positive number (the step size). In what follows, we denote the computed approximate value of $y(x_i)$ by y_i.

There are two broad classes of methods for the numerical solution of the above problem: single-step methods compute y_{i+1} using only the previously computed y_i, whereas multiple-step methods use several previously computed values, say $y_{i-k}, y_{i-k+1}, \ldots, y_i$.

As a starting point for the discussion of single-step methods, we recall the Taylor's series expansion of $y(x_{i+1})$ with respect to $y(x_i)$:

$$y(x_{i+1}) = y(x_i) + h_i y'(x_i) + h_i^2 y''(x_i)/2 + \cdots.$$

By truncating this series after a specific number of terms and by differentiating the original differential equation the required number of times to obtain derivatives of $y(x)$ higher than 1 (remembering to use the chain rule in differentiating $f(x, y)$, we can obtain formulas of the form

$$y_{i+1} = y_i + h_i(f)_i + h_i^2(f_x + f_y f)_i/2 + \cdots.$$

Here the subscripts x and y denote differentiation and the subscript i following the parentheses denotes evaluation using data obtained at the previous step.

The particularly simple Euler method is obtained in this way by truncating after the second term on the right:

$$y_{i+1} = y_i + h_i f(x_i, y_i)$$

If we assume for the moment that the data at $x = x_i$ is exact, then the terms in the series that are discarded represent the error introduced in y_{i+1} in integrating from x_i to x_{i+1}. This is called the local truncation error. In the case of the Euler method, the local truncation error is bounded in absolute value by a constant times h_i^2. The formula is applied successively in stepping to the right from y_0; thus the error accumulates as the solution is developed. It can be shown that for a uniform interval length h the accumulated truncation error at a given value of x is bounded by a constant times h; thus the Euler method is said to be of order h, or first-order accurate. This method is illustrated in Fig. 11(a) for the problem $y' = -y, y(0) = 1$, using several different values of h.

In most situations it is not possible to compute the derivatives of $f(x, y)$ that are required for higher order

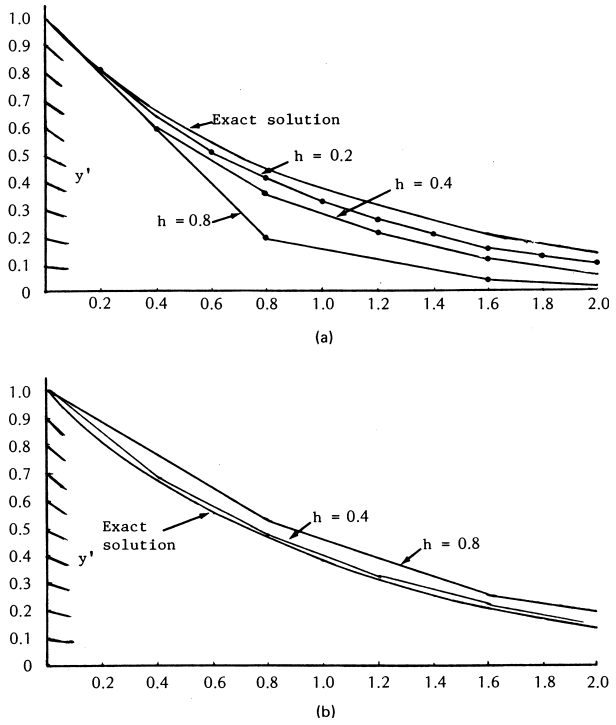


FIGURE 11 Numerical solution of an ordinary differential equation initial-value problem: $y' = -y$; $y_0 = 1$. (The solution is $y = e^{-x}$.) (a) Euler method. (b) Heun method.

versions of the Taylor method; however, the popular Runge–Kutta methods can be shown to be equivalent. For example, a second-order Runge–Kutta method, sometimes called the Heun method or the trapezoidal method, is given by

$$y_{i+1} = y_i + h_i(k_1 + k_2)/2,$$

where

$$k_1 = f(x_i, y_i)$$

and

$$k_2 = f(x_{i+1}, y_i + h_i k_1).$$

The Heun method is illustrated in Fig. 11(b) for the same problem as for the Euler method, showing the improved accuracy. The commonly used fourth-order Runge–Kutta method is given by

$$y_{i+1} = y_i + h_i(k_1 + 2k_2 + 2k_3 + k_4)/6,$$

where

$$k_1 = f(x_i, y_i)$$
$$k_2 = f(x_i + h_i/2, y_i + h_i k_1/2)$$
$$k_3 = f(x_i + h_i/2, y_i + h_i k_2/2)$$
$$k_4 = f(x_i, y_i + h_i k_3)$$

Single-step methods have the advantage that the step size h_i can be easily changed at each step so that, when the solution is changing rapidly with respect to x, the step size can be diminished; alternatively, when the solution is changing slowly, it may be increased. It is possible to automate this during the solution process by estimating the error at a given step and then to halve, double, or retain the current step size depending on the magnitude of this estimate.

Most of the computation that is needed to numerically solve an initial-value ODE problem is in the evaluation of $f(x, y)$ for different values of the variables x and y. Because single-step methods require many evaluations of this function per step, they may be prohibitively slow for complex problems. The multistep methods require fewer evaluations of $f(x, y)$ per step for the same order of truncation. They are based on the approach of integrating the differential equation from x_i to x_{i+1}, where $f(x, y)$ is replaced by a polynomial $p(x)$ of degree (say) k, which interpolates $f(x, y)$ at the $k + 1$ previously computed points (x_j, y_j), $j = i - k, 1 - k + 1, \ldots, i$. Thus,

$$y_{i+1} - y_i = \int_{x_i}^{x_{i+1}} y'\,dx \approx \int_{x_i}^{x_{i+1}} p(x)\,dx.$$

In the case where $k = 3$ and the interval between points is a uniform length h, we can derive in this way, the fourth-order Adams–Bashforth formula

$$y_{i+1} = y_i + h(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3})/24.$$

Although more efficient in terms of the number of evaluations of $f(x, y)$, multistep methods do suffer from some disadvantages.

1. Special start-up procedures are required to obtain the needed, previously computed values of y_i.
2. Complicated restart procedures are required whenever it is necessary to change the interval size.
3. These methods tend to be numerically unstable, meaning that small, initial errors tend to grow as x increases.

To help alleviate the third problem—which is explained a little more completely in the next section—an initial multistep formula is used to predict a value for y_{i+1}, and this is followed up with a compatible corrector formula. A suitable corrector for the Adams–Bashforth method is the Adams–Moulton formula,

$$y_{i+1} = y_i + h(9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2})/24$$

where f_{i+1} is computed using the predicted y_{i+1}.

Before ending this section we mention that methods for the solution of a single first-order differential

equation carry over to *systems* of first-order differential equations

$$\begin{aligned} y_1' &= f_1(x, y_1, y_2, \ldots, y_n) \\ y_2' &= f_2(x, y_1, y_2, \ldots, y_n) \\ &\vdots \\ y_n' &= f_n(x, y_1, y_2, \ldots, y_n) \end{aligned}$$

Here, the subscript refers to one of the n dependent variables, rather than evaluation at a discrete value of x. We can express this system in exactly the same way as a single equation if we merely take y to represent the vector of dependent variables.

Finally, higher-order differential equations (containing second and higher derivatives) can be reduced to a system of first-order differential equations by treating each derivative of each dependent variable as a separate variable in a first order system. Thus the equation

$$y'' + ay' + by + c = 0$$

can be written as

$$y_1' = -ay_1 - by_2 - c, \quad y_2' = y_1$$

B. Numerical Instability and Stiffness

Numerical instability is a phenomenon due to the numerical method used to solve a problem—not the problem itself. It may be illustrated in the context of initial-value problems for differential equations by attempting to solve

$$y' = -y; \quad y(0) = 1$$

using the simple two-step (midpoint) method with constant h:

$$y_{i+1} = y_{i-1} + 2hf(x_i, y_i)$$

Substituting $-y$ for $f(x, y)$, we have

$$y_{i+1} = y_{i-1} - 2hy_i$$

which is a second-order difference equation. Since the exact solution to the problem is known to be $y(x) = e^{-x}$, we can take $y(0) = 1$ and $y(h) = e^{-h}$ for the initial data corresponding to $i = 0$ and $i = 1$ in the difference equation. The solution to the difference equation can now be determined using standard techniques to be

$$\begin{aligned} y_i = {} & A(-h + \sqrt{(h^2 + 1)})^i \\ & + B(-h - \sqrt{(h^2 + 1)})^i \end{aligned}$$

where

$$A = (1 + (e^{-h} + h)/\sqrt{(h^2 + 1)})/2$$

and

$$B = (1 - (e^{-h} + h)/\sqrt{(h^2 + 1)})/2$$

If we fix a value of $x = ih$ and substitute x/h for i in the expression for the solution to the difference equation, and then take the limit as h goes to 0, we see that A goes to 1, B goes to 0, and the solution becomes

$$\lim_{h \to 0} y_{x/h} = \lim_{h \to 0} (1 - h)^{x/h} = e^x$$

which is consistent with the exact solution. However, for any *finite* h, B is not 0 and the second term contains a factor greater than 1 that is raised to the ith power. The second term is called a parasitic solution, which, as x (and therefore i) increases, will eventually dominate, generating alternate positive and negative numbers of increasing magnitude. This phenomenon is illustrated in Fig. 12(a) for the example of the midpoint method applied to the same problem as the Euler method in Fig. 11(a). Note that for a step size of 0.4 the midpoint method diverges from the solution, whereas the Euler method, having a lower order of accuracy, does not. A possible cure for this type of instability is to apply, at each step, a correction based on an implicit formula of the same order as the original method. By implicit we mean depending on y_{i+1}, which, of course, is the value we wish

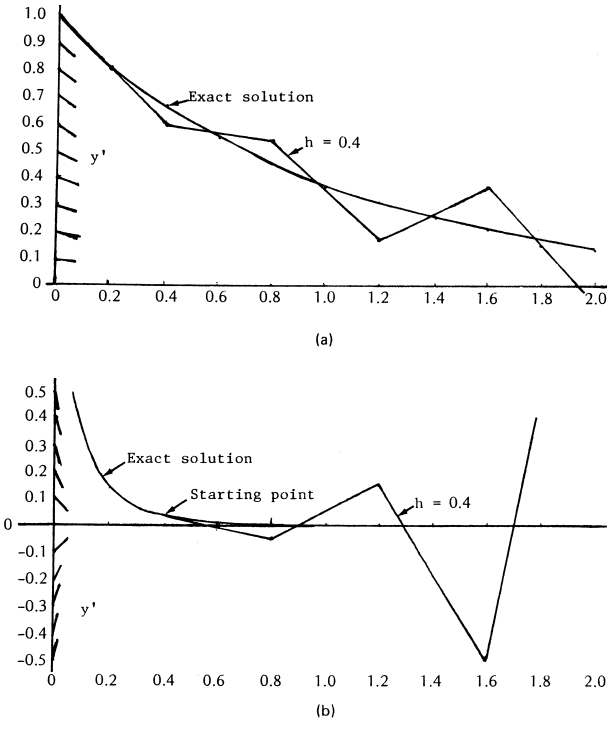


FIGURE 12 Instability and stiffness in the numerical solution of an initial-value ordinary differential equation problem: (a) $y' = -y$; $y_0 = y(0) = 1$ by midpoint method: illustrates instability; (b) $y' = -10y$; $y(0) = 1 : y_0 = y(0.4) = e^{-0.4}$, Euler method: illustrates stiffness.

to correct. As an example, the second-order trapezoidal method

$$y_{i+1} = y_i + h_{i+1}(f_i + f_{i+1})/2$$

may be used, where f_{i+1} is computed from the value of y_{i+1} obtained from the method previously given.

In contrast to numerical instability, the problem of *stiffness* is associated with the problem to be solved and is due to the function $f(x, y)$ changing greatly in response to small changes in y. The usual example given is the problem

$$y' = -Ky, \quad y(0) = 1,$$

where K is a large number, say 1000. If a standard numerical method is used to solve this problem, a very small step size is needed at the start to track the rapidly changing solution $y = e^{-Kx}$. As x grows large, however, y changes slowly with respect to x, which suggests that the interval size could be enlarged. Unfortunately, though, a small error in computing y produces a large error in y', which in turn produces a large error in the next value of y unless the interval is very small. This phenomenon is illustrated for the Euler method in Fig. 12(b) for $K = 10$, starting at $x = 0.4$ and using a step size of 0.4. In this example, h would have to remain less than 0.2 in order to maintain stability, even though once x is greater than 0.4 there is little change in the solution. The usual approach to overcome this problem is, for each step, to follow a prediction formula with an implicit correction formula, solving for y_{i+1} with a high-order iteration method (e.g., Newton's method).

It should be mentioned here that mathematical subroutine packages, which are available for most scientific computer systems, contain carefully designed routines that are tailored for a variety of problem types. The problems of instability and stiffness illustrate pitfalls in using arbitrary methods for the solution of initial-value problems in ordinary differential equations without due care; however, with proper selection, the available mathematical subroutines can usually be relied on for near-optimum solutions.

C. Ordinary Differential Equations—Boundary-Value Problems

A necessary condition for a unique solution to an ordinary differential equation is that the solution satisfy some prescribed auxiliary conditions. The number of such conditions must equal the order of the equation. In the case of initial value problems, these conditions are the initial values of the solution and all of its derivatives up to an order 1 less than the order of the equation. It is also possible, in the case of equations of second order or higher, to prescribe values of the solution or its derivatives at more than one value of the independent variable. Such problems are referred to as boundary-value problems. For example, a second-order, two-point boundary-value problem is

$$f(x, y, y', y'') = 0 \qquad \text{for} \quad a \le x \le b$$

and

$$y(a) = \alpha, \quad y(b) = \beta,$$

where α and β are known values.

There are several distinctly different approaches to obtaining numerical solutions to boundary-value problems, and the choice may depend on a number of factors, including experience with similar problems.

As stated at the end of Section VIII.A, it is possible to formulate an ordinary differential equation of second order or higher as a system of first-order equations. If this is done for a boundary-value problem, not all of the initial conditions will be known; however, if estimates are made for the missing initial conditions, the solution can proceed as an initial-value problem, and its value (or value of its derivatives) at the later boundaries can be compared with known data.

Methods based on this approach are known as shooting methods and are relatively easy to set up for computer solution, since good subroutines are usually available for initial-value problems.

If the differential equation is linear, which is true if all derivatives appear only as multiples of constants or functions of x, then it is possible to make a precise correction to the initial data after the first trial; thus, a final solution can be obtained with two initial-value solutions. If the equation is not linear, the initial-value problem must be repeated, with successive estimates of the unknown initial conditions being made in the manner of the secant method, until all boundary conditions are met.

A second approach is to approximate the derivatives that appear in the differential equation with finite differences. The domain of the problem is first partitioned into, say, m intervals by selecting the points $x_0 = a, x_1, x_2, \ldots, x_i, \ldots, x_m = b$. Since the differential equation must hold at each of the interior points ($i = 1, 2, \ldots, m-1$), we have $m-1$ equations involving discrete values of x, y, and derivatives of y. Finite-difference approximations to the derivatives involve the value of y at neighboring points; for example,

$$y_i' \approx (y_{i+1} - y_{i-1})/2h$$

and

$$y_i'' \approx (y_{i-1} - 2y_i + y_{i+1})/h^2$$

are second-order approximations to $y'(x)$ and $y''(x)$ at the point defined by the index i when the partition intervals are a uniform length h. Since y_0 and y_m are determined by

the boundary data, we are left with a system of $m-1$ algebraic equations in the $m-1$ unknown values of y. These equations will be linear or nonlinear depending on the linearity of the differential equation and will be banded. The solution to the algebraic equations is a discrete representation to the solution to the differential equation, which is of the same order of accuracy with respect to the interval length as is the finite-difference approximation used for the derivatives (assuming the exact solution of the algebraic system).

Yet another approach will be illustrated for the solution of the linear two-point boundary-value problem

$$Ly(x) \equiv y''(x) + p(x)y'(x) + q(x)y(x) = r(x)$$
$$y(a) = y(b) = 0.$$

The symbol L is defined in the above expression to represent the linear operation applied to $y(x)$, in order to simplify the notation. The solution $y(x)$ is represented in terms of, say, m predefined basis functions

$$y(x) \approx \sum_{j=1}^{m} a_j \phi_j(x),$$

where the $\phi_j(x)$ satisfy the boundary conditions and the a_j are coefficients to be determined.

There are several ways in which this may be accomplished. *Collocation* is to substitute the approximation for $y(x)$ and its derivatives into the differential equation and evaluate at m interior points to obtain a system of m algebraic equations for the as. *Least squares* is to do the same thing for more than m interior points and then to solve the resulting over-determined system by a least squares technique.

In order to discuss the Galerkin method, we must first give two definitions.

1. The "inner product" of two functions $g(x)$ and $h(x)$ on $a \le x \le b$ is denoted by $\langle g(x), h(x)\rangle$ and is defined by

$$\langle g(x), h(x)\rangle \equiv \int_a^b g(x)h(x)\, dx.$$

2. The functions $g(x)$ and $h(x)$ are orthogonal on $a \le x \le b$ if

$$\langle g(x), h(x)\rangle = 0.$$

The Galerkin method seeks to find the coefficients a_j in the above approximation to $y(x)$ by setting the residual

$$R(x) \equiv L\sum_{j=1}^{m} a_j \phi_j(x) - r(x)$$

orthogonal to each of the basis functions. Thus,

$$\langle \phi_i(x), R(x)\rangle = 0 \qquad i = 1, 2, \ldots, m.$$

This leads to the system of linear equations

$$[L_{i,j}]\{a_j\} = \{r_i\},$$

where the elements of the matrix $[L_{i,j}]$ and vector $\{r_i\}$ are

$$L_{i,j} = \langle \phi_i(x), L\phi_j(x)\rangle$$

and

$$r_i = \langle \phi_i(x), r(x)\rangle,$$

respectively.

There are a number of choices for the selection of the basis functions. They must be linearly independent in order that the linear system for the solution of the coefficients be nonsingular. It is also desirable, but not essential, that they be mutually orthogonal. Frequent choices are trigonometric functions or orthogonal polynomials such as Tchebycheff polynomials. If the domain of the problem is partitioned, as described for the finite-difference methods, then it is possible to construct basis functions for piecewise polynomials having local support—that is, being zero everywhere except on a single interval or its immediate neighbors (for example the basis functions for the cubic spline). This choice is the prototype for the finite element methods and has the advantage that the resulting linear systems have a banded structure.

In order to solve a nonlinear problem using one of the above methods, it is necessary to linearize the problem by expanding about an initial estimate of the solution. For example, for the problem given at the beginning of this section, assume a solution estimate $y^{(0)}(x)$ and expand $f(x, y, y', y'')$ about this estimate:

$$f(x, y, y', y'') \approx f(x, y^{(0)}, y^{(0)}, y^{(0)''}) + (\partial f/\partial y)^{(0)}\Delta y + (\partial f/\partial y')^{(0)}\Delta y' + (\partial f/\partial y'')^{(0)}\Delta y''$$

The superscript (0) implies substitution of $y^{(0)}(x)$ for $y(x)$. When set to zero, the right-hand side is a linear equation in the increment $y(x)$, which may be solved by one of the above techniques using the boundary conditions $y(a) = y(b) = 0$. The solution may then be added to $y^{(0)}$ to obtain an improved estimate, and the process is repeated until two successive estimates are sufficiently close.

D. Partial Differential Equations

The numerical solution of problems involving differential equations in two or more independent variables is a highly complex procedure, generally requiring extensive analysis and the customizing of approaches to suit the problem at hand. The procedure breaks down into several interdependent subtasks.

First, a description of the domain of the problem must be given in mathematical or numerical terms. This often

involves problems in geometry modeling, such as solid-object representation, or wire-frame or patch representation of surfaces.

Second, a computational mesh must be inscribed in the domain. It is best that the mesh conform to the boundary surfaces; it needs to be fine where the variables of the problem are known to change rapidly, and coarse where they do not, and finally, the mesh surfaces should be as close to orthogonal as possible. It is often desirable to set up a transformation from a physical mesh that is not rectangular to a uniform rectangular mesh, but if this is done, the governing differential equations and boundary conditions must be transformed to the new domain.

The next step is to discretize the governing equations and boundary conditions.

If the problem is elliptic, the techniques discussed under two-point boundary-value problems for ordinary equations will apply, as modified for multidimensional domains. The most popular technique is to approximate derivatives by finite differences, which leads to finite-difference equation for each interior grid point (there may be many thousands of these). Finite-element methods, based on multidimensional basis functions defined over small subdomains, are also used quite frequently in conjunction with collocation, least squares, or Galerkin's method to arrive at an algebraic system of equations. Finally, orthogonal basis functions defined over the entire domain are sometimes used in conjunction with Galerkin's method (examples are the so-called spectral methods).

If the problem is parabolic, the domain will be open in one dimension (usually time), and the solution will be given on an initial surface or volume. In this case, the solution is developed in increments of time. An "explicit" finite difference method defines the solution at a mesh point at time $t + \Delta t$ (where Δt is the time increment) in terms of neighboring mesh points at time t; however, such methods are unstable except for very small increments of time. An "implicit" finite difference method, on the other hand, defines the solution at a mesh point at time $t + \Delta t$ in terms of neighboring points at both t and $t + \Delta t$. Although more stable, the implicit methods require the solution of systems of equations at each time step. A popular method called alternating direction implicit (ADI) defines the solution at a point at time $t + \Delta t$ in terms of neighboring points at time t and points along a line (or plane for three-dimensional problems) at time $t + \Delta t$, then for the next time step uses the same procedure except the line or plane is orthogonal to the first; the solution continues as the name suggests.

If the problem is hyperbolic, the domain will again be open in one dimension. The initial solution and its derivative with respect to the open variable will be given. In this case, the direction of the (real) characteristic lines can be inferred from the initial data, and the solution can be computed incrementally along these lines.

It should be mentioned that in all cases higher-order methods can be devised through the use of high-order difference formulas and techniques analogous to those used for high-order initial-value problems for ordinary differential equations; however, difficulties may arise near the boundaries and in incorporating boundary data in a consistent manner.

The final step is to decide how to solve the system of algebraic equations that arises from the discretization of the problem (other than in the purely explicit situation). The most difficult case is the elliptic one. If the governing equation is linear, then so will be the algebraic system, and it may be possible to use Gaussian elimination, although the size and sparsity of the system will probably dictate an iterative method. Unfortunately, in most realistic situations the governing equations will be nonlinear. As in the ordinary differential equation case, it may be possible to linearize about an estimate of the solution, then solve a sequence of linear correction problems. Alternatively, a somewhat artificial technique that is frequently adopted, is to add a time derivative term to the elliptic equation, then solve the resulting parabolic equation, using something like the ADI method with a solution estimate for an initial condition.

IX. RECENT DEVELOPMENTS

The recent increase in speed, memory capacity, and parallelism in computers has influenced advances in numerical analysis. These factors have permitted attempts to solve very large and complex problems, such as the simulation of flow fields in irregular geometric domains, and have led to the development of algorithms to take advantage of parallel computers and vector processors.

A. Large, Sparse Matrices

Very large PDE problems tend to produce large matrices that have relatively few nonzero entries. This has led to the development of encoding techniques for storing only the nonzero entries of large, sparse matrices in the memory of computers and to special schemes for manipulating them (for example, to multiply a sparse matrix times a vector, or to multiply two such matrices). Direct methods for the solution of linear systems of equations, $Ax = b$, such as Gaussian elimination or Cholesky decomposition, progressively generate nonzero numbers in positions where there were previously zeros—resulting in a prohibitive increase in storage and number of computations for large, sparse systems. To minimize this effect, methods have been devised that reorganize the equations. SPARSPAK, a package of computer codes for working with large, sparse

matrices, has been written by investigators at the University of Waterloo, Canada.

B. Iterative Methods

Iterative methods for the solution of linear systems, as discussed in Section IV.C, have the advantage that they preserve sparsity; thus, they are the methods of choice for many large problems. Much progress has been made with the preconditioned conjugate gradient method. If the system of linear equations to be solved is $Ax = b$, then by preconditioning we mean finding a nonsingular symmetric matrix, C, such that the condition number of $\hat{A} = C^{-1}AC^{-1}$ is smaller than the condition number of A. The related system $\hat{A}\hat{x} = \hat{b}$, where $\hat{x} = Cx$ and $\hat{b} = Cb$, is solved by the conjugate gradient method. Careful preconditioning will substantially improve the convergence rate. The physics of the problem may assist in determining the preconditioning matrix, but there are now software packages (e.g., the Yale University PCGPAK) that contain quite robust, general purpose codes that provide preconditioning based upon analysis of the matrix A.

For the extraction of eigenvalues of large, sparse matrices, the Householder method (Section V.B) for tridiagonalizing a symmetric matrix also suffers from the problem of nonzero fill. This has led to renewed interest in the Lanczos method, which is closely related to the conjugate gradient method for solving linear systems. The Lanczos method is based on the idea of successively generating a sequence of orthonormal vectors that form the columns of a matrix Q, such that $AQ = QT$, where A is the symmetric $n \times n$ matrix whose eigenvalues are sought, and T is a symmetric, $m \times m, m \leq n$, tridiagonal matrix. Even though T has a smaller order than A, its eigenvalues will be among those of A. At any stage, the eigenvalues of T may be found by use of the QR algorithm (Section V.B). The eigenvalues of A will appear in descending order of magnitude as additional vectors of Q are found. Arithmetic roundoff error can be a problem in this method; thus, a number of varients that selectively reorthogonalize the generated vectors have been devised. It is considered an iterative method, since the procedure is normally terminated before all n possible orthogonal vectors have been found. For both the preconditioned conjugate method and the Lanczos method, variants are possible that work with some nonsymetric problems.

For elliptic PDE problems, an iterative method known as the multigrid method has become very popular. The domain of the problem is discretized through the construction of a grid which is then subdivided in successively finer scales. The PDE is discretized on the finest grid, resulting in a linear system to be solved. An iterative method such as Gauss–Seidel is used, but after a few iterations the solution is "injected" onto the next coursest grid and the procedure is repeated. After the first (coursest) grid is reached in this manner, the solution is interpolated back to a finer grid and the cycle is repeated. The advantage of this procedure is that the low frequency components of the error in the solution are damped out as the high frequency components are. The overall solution scheme is highly efficient.

C. Parallel and Vector Algorithms

Increased net computation rate is achieved through parallel operation of processors (parallel computers) and in assembly line techniques for performing arithmetic operations on arrays of numbers (vector processors). Unfortunately, at the present time, different computer architectures place different emphasis on these approaches. Some computers use a small number of powerful vector processors while others contain a large number of relatively simple processors. Generally speaking, the algorithms discussed in this article are equally applicable to different architectures; however, the implementation must be carefully thought out to take maximum advantage of the particular features. One class of algorithms, particularly suited to parallel computers, is that of domain decomposition. Here the domain of a problem is divided into subdomains and the governing differential equations are solved simultaneously on each domain by different processors. The critical problem in these methods is that of matching data at the interface between the subdomains. A common approach is to provide for a region of overlap at the interface.

D. New Approximation Methods

As mentioned in Sections VIII.C and D, an approach frequently used for representing a function $y(x)$ in a computer (particularly if it is the solution of a differential equation) is to define a finite set of basis functions, $\{\phi_j(x)\}_{j=1}^{N}$, and to set

$$\hat{y}(x) = \sum_{j=1}^{N} a_j \phi_j(x),$$

where the a_j are to be determined. Basis functions that have received much attention in recent years are the spectral functions such as trigonometric functions or the Chebychev polynomials. They have the advantage of exponential accuracy and efficient computation (the latter through use of Fast Fourier Transform approaches). Although spectral methods present some difficulties when applied to problems with irregular boundaries or discontinuous solutions (e.g., shocks), considerable progress has been made in their application to a wide variety of practical problems.

A new form of basis function, referred to as a wavelet and characterized by the Morlet wavelet

$$\phi_{a,b}(x) = e^{i\omega_0 x} e^{-b(x-a)^2},$$

is being actively investigated. This function is globally defined, yet its value is insignificant except in a small region centered on $x = a$. A family of such functions can be generated by assigning different values to the parameters a and b (shift and dilation) in order to match the characteristics of $y(x)$. Methods based on wavelet basis functions appear to have the advantages of the spectral methods while still having the flexibility to model rapid spatial changes.

SEE ALSO THE FOLLOWING ARTICLES

APPROXIMATIONS AND EXPANSIONS • COMPUTER ALGORITHMS • DIFFERENTIAL EQUATIONS, ORDINARY • LINEAR SYSTEMS OF EQUATIONS • WAVELETS

BIBLIOGRAPHY

Burden, Richard L., and Faires, J. Douglas (1993). "Numerical Analysis," 5th ed. Prindel, Weber & Schmidt, Boston.

Ciarlet, P. G., and Lions, J. L., eds. (1999). "Handbook of Numerical Analysis," vols. I–VI. Elsevier, Amsterdam.

Golub, Gene, and Van Loan, Charles F. (1983). "Matrix Computations," John Hopkins, Baltimore.

Golub, Gene, and Ortega, James M. (1993). "Scientific Computing, an Introduction with Parallel Computing," Academic Press.

Stewart, G. W. (1996). "Afternotes on Numerical Analysis," SIAM.

Stewart, G. W. (1998). "Afternotes goes to Graduate School: Lectures on Advanced Numerical Analysis," SIAM.

Young, David M., and Gregory, Robert Todd (1972). "A Survey of Numerical Mathematics," vols. I and II. Addison Wesley, Reading, MA.

IMA Journal of Numerical Mathematics, Institute of Mathematics and Its Applications, Oxford University Press.

Numerische Mathematik, Springer-Verlag.

SIAM Journal of Numerical Analysis, Society for Industrial and Applied Mathematics.



Operations Research

Salah E. Elmaghraby
Shu-Cherng Fang
North Carolina State University

I. Perspective
II. Mathematical Programming
III. Stochastic Processes
IV. The Fuzzy Paradigm
V. Selected Models of Common Processes
VI. Selected Fields of Application

GLOSSARY

Decision analysis Determination of rational decisions in the face of uncertainty.
Fuzzy set theory Mathematical theory of decision based on nonbinary logic of set inclusion.
Mathematical programming Optimization of real functions subject to constraints.
Operations research Use of scientific methodology in studying operational systems.
Queueing theory Mathematical theory for the analysis and optimization of units demanding service at a facility.
Simulation Representation of the essence, without the reality, of the system under study for the purposes of analysis and design.
Stochastic process Process that evolves over time in a nondeterministic fashion.

OPERATIONS RESEARCH (OR) deals with the analytical study of operational systems, where the latter are defined as systems that are subject to human decision. This is a vast collection, which includes, among others, production, transportation, warehousing, agriculture, communications, education, forestry, defense, trade, economic, and government systems. OR is mainly concerned with the study of the structure of such systems and with observing their behavior, first to gain basic understanding of their nature and of the factors that influence their performance, and second to modify their performance, with an eye toward improving such performance. Such improvement is achieved through the optimal allocation of resources, or the optimal utilization of existing resources. The approach of OR is in harmony with the classical view of the scientific approach, which embodies observation, modeling and establishing hypotheses to explain the phenomena under study, the validation of both models and hypotheses, and the use of these theories and models to describe the performance of the system under modified conditions.

Operations research is strongly based in the mathematical and physical and engineering sciences, with the computer and its technology playing an increasingly important role in the implementation of its findings. Its methodology

borrows from, and contributes to, the fields of mathematical programming, discrete mathematics, stochastic processes, control and optimal control, and economics and econometrics, among others. An OR study may concern itself with a small system, such as an individual assembly line or a truck route, or with a large system such as a multimodal evacuation system in case of emergency.

Typically, OR is a graduate program of instruction, with introductory treatment given at the undergraduate level in departments of mathematics, statistics, industrial engineering, and business. Its graduates are in high demand at both the master's and doctoral levels in industry, consulting, research centers, universities, and government.

I. PERSPECTIVE

Operations research (OR) is the use of scientific methodology in studying operational systems. A system is any collection of items that are related. Systems of concern to OR are purposeful systems, so that the objective(s) of study can be made purposeful also. Operational systems are systems whose design and/or operation require human decision making. OR provides the means for making the most effective decisions—some of which are concerned mainly with design (e.g., the optimal design of equipment and/or administrative organizations), while others are mainly operational in nature (e.g., the optimal operation of existing equipment and/or administrative organizations). The strength and versatility of OR stem from its diagnostic power through observation and modeling, as well as from its prescriptive power through analysis and synthesis. The techniques used are continually being improved and expanded, with the aid of both basic and applied research.

OR is interdisciplinary in nature, drawing on (and contributing to) the theories and techniques from many fields, including mathematics, statistics, computer science, engineering, economics, and the physical sciences. In distilling from such theories and practices those that apply to a given system, the OR practitioner initially acts like a scientist—one who observes how the system responds to various stimuli, so that it can be described by a model, mathematical or otherwise, that is capable of predicting the consequences of the various possible decisions. Although the insight produced from constructing such a model is frequently very valuable in itself, the OR practitioner usually goes on to thoroughly analyze the model—with the ultimate goal of determining those decisions that would make the given system perform as well as possible. In doing so, additional equipment and/or administrative organizations are frequently designed and/or synthesized.

OR practitioners have been confronted with a wide variety of real-world problems, varying from the optimal design of power plants in the face of uncertain demand over a planning horizon that extends for several decades, to the scheduling of flight arrivals and departures at an airport. They have found employment in both industry and government, as well as in the military (where OR originated during World War II). Most importantly, new applications are continually arising, with many originating from relatively recent societal problems, such as food and energy production and distribution, health services delivery, pollution control, and transportation planning.

Operations research has oftentimes been confounded with "systems science," or "systems engineering," and the terms have been used interchangeably. There is a difference. Traditionally, systems engineering has been identified with hardware systems, while OR has been identified with software. Systems science is concerned with more global problems and is more abstract in its approach than OR. A systems study may involve societal and/or biological components that have not been successfully quantified, while OR tends to be mostly concerned with more quantifiable problem settings, and rarely, if ever, has it concerned itself with biological systems.

A. Historical Perspective

OR was "born" during World War II, when the Allied Military Command sought the assistance of the scientific community to resolve certain operational problems (hence the British name of "operational research"), though it is generally recognized that the application of science in military affairs dates back to the days of Archimedes. Teams were formed of a variety of disciplines according to the problem in hand, mostly composed of physicists, mathematicians, statisticians, economists, and engineers. The reports of their investigations make penetrating and delightful reading, covering a wide spectrum of problems from the optimal depth to detonate antisubmarine devices to the arming of merchant ships to maximize the safety of convoys.

After World War II, progress of OR in the United States followed a different path from Europe. The devastation in Europe necessitated the rapid dismemberment of the military establishment and the conversion to civilian economy. This, in turn, shifted the contribution of OR to the industrial and commercial enterprises, where OR was rapidly accepted as a novel approach to the resolution of managerial problems. On the other hand, because of the ensuing cold war between the two superpowers after World War II, OR continued its close association with the military in the United States, with significant work sponsored by defense-related establishments such as the RAND Corp. of Santa Monica, California, the Office of Naval Research in Washington, D.C., and the Army Research Office in Durham, North Carolina. The shift to predominantly

civilian applications did not take hold in the United States except at a later date, in the early 1960s. Since then, it has flourished with the progress in computer technology, and has gained, together with its alter ego "management science," wide acceptance in the mainstream of corporate America, state and local government, and the service industry. In the nonmilitary domain, OR was received with a great deal of skepticism because of the long-standing tradition of "scientific management" in industry. It had to earn its colors by demonstrating its worthiness at all levels of application: strategic, tactical, and technical.

A discipline is distinguished by its field of application, its methodology, and the specialized education and training its practitioners must receive to qualify for working in the discipline. OR is no exception. At the time of this writing, there exist over 40 institutions of higher learning in the United States alone that offer advanced degrees (master and doctorate) in OR (see the next section). Furthermore, a number of professional societies have sprung up to serve the new scientific community, of which we mention the Institute for Operations Research and Management Science (INFORMS) and a number of OR societies all over the world (the United Kingdom, France, the Federal Republic of Germany, India, etc.) that are united in the International Federation of Operations Research Societies. Many other scientific and professional societies have "divisions" and/or "special interest groups" in OR (such as the Institute of Industrial Engineers and the Institute of Electrical and Electronics Engineers, among others).

The following sections give a synopsis of the methodologies, models, and fields of application of OR. For more information, consult the bibliography.

B. The Education and Training of the OR Practitioner/Researcher

Of necessity, the early OR professionals were "migrants" from other disciplines. Gradually, as the methodologies peculiar to the nascent field were developed, there also evolved a need for education and training in these methodologies, as well as the need to refine, hone, and expand them; the latter are evident activities of research.

One may identify several types of OR education: formal, semiformal, informal, and on-the-job. Formal education consists of regular university or college programs and courses for which credit is earned. There are a few undergraduate courses throughout the United States that may be designated as "OR courses"; these are typically offered in the departments of mathematics, statistics, and industrial engineering. The majority of formal education is at the graduate (master or doctoral) levels. The reason is simply that sufficient background has to be gained in the traditional mathematical and physical sciences before the student can absorb the new and more advanced concepts of OR. Normally, such background constitutes a large proportion of the requirements for a baccalaureate degree in a field of engineering or the mathematical sciences (mathematics, statistics, and computer science).

Schools of engineering seem to have embraced OR more than any other school, followed by graduate schools of business, and trailed by schools of mathematical sciences. On a percentage basis, formal OR educational programs are about 50% imbedded in regular engineering departments (with the lion's share going to industrial engineering), 30% in business schools, 12% in independent OR departments or interdisciplinary programs, 5% imbedded in regular mathematical sciences departments (mostly departments of statistics), and a miniscule fraction labeled "other."

Semiformal education can be gained through a plethora of short courses offered on a regular basis by universities and consulting firms alike. A typical course runs for 1–3 weeks, and ranges in level from the elementary to the advanced. Recently, however, interest in and attendance at introductory short courses has been waning, reflecting the existence of a pool of persons with formal OR education who are more inclined to attend a 3-day seminar aimed at mastering some new speciality that they perceive as important to their performance in their regular jobs.

Informal education is shouldered mostly by the professional societies and their local chapters (especially (INFORMS)), through the plethora of media of communication available to them (meetings, publications, special interest groups, circulars).

Perhaps the most effective learning is experiential or on-the-job learning. One learns more easily and more quickly if one is motivated by direct applicability of what is learned to one's immediate concern. Indeed, the other three forms of education are of little worth to the OR practitioner unless they are directly relevant to his/her assigned tasks. Learning is best as an apprentice to an experienced professional.

What is the subject matter of a "proper" OR program of instruction? Here is a "maximal list" at the doctoral level. What is actually studied will depend largely on the area of expertise sought by the individual. As to be expected, mathematics plays a dominant role in this list.

Mathematics: Discrete mathematics, calculus through real and functional analysis, complex analysis, linear algebra, modern algebra, topology and measure theory, probability theory.

Statistics: Inference, design of experiments, sequential analysis, time series analysis, decision theory.

Computer science: Data structures, complexity and theory of algorithms, graphs, artificial intelligence.

Operations research: The topics covered under "methodologies" in this article define specialization in OR, and all, or at least a majority of them, should be covered in

depth. Broadly speaking, one must gain proficiency in three broad areas: mathematical programming, stochastic systems, and optimal control theory.

C. The Conduct of an OR Study

Succinctly stated, the steps of an OR study do not differ significantly from those of the "scientific method" in general. Any differences present reflect the "purposeful" character of OR (in contrast with the pursuit of knowledge that characterizes "pure" scientific investigations). In particular, one may define eight steps as follows.

1. Define the objectives of the study. Other words commonly used are the deliverables, the goal of the study, and what the project is supposed to accomplish.
2. Develop the project plan. This should include the resources needed, their timing, and the activities to be undertaken.
3. Define the problem(s). Equivalently, diagnose "what is wrong" with the operational system. This is a vital step, since the wrong diagnosis will eventually lead to the wrong remedial action unless corrected later.
4. Determine the problem boundaries. Realize that systems do not exist in isolation, and that everything is related to everything else. But a study, no matter how global it is intended to be, cannot encompass all activities. At some point a decision must be made to define what shall be of concern to the OR study, and what shall be excluded and taken as "exogenous input."
5. Develop the model. It is a mathematical/computer/analog representation of the problem(s) as defined in Step 4. Such representation defines the controllable decision variables and the relations among them, the constraints imposed internally or externally that limit the analyst's freedom of action, and the criterion (or criteria) against which decisions can be measured. In model development we include also the specification of the data required and the method by which it is to be collected.
6. Develop the computational approach. In reality, one undertakes steps 5 and 6 concurrently, since they cannot be logically separated. The construction of a model of the system under investigation is dependent on the technology available to the analyst for securing answers, and conversely. In the development of the computational approach we also include any programming/debugging required to accomplish the computations specified.
7. Collect data and validate the model. Model validation has many components. They are:
   (a) Plausibility: Does the model fit special cases for which actual data are available?
   (b) Consistency: Are results logical when major parameters are varied (especially to extremes)?
   (c) Sensitivity: Are the relative magnitudes of changes in outputs appropriate for small changes in the inputs?
   (d) Workability: Is the model easy to solve routinely if that is required by the operational system?
8. Implement the proposed solution. This is more easily said than done. The real worth of the study will depend heavily on the degree of cooperation and participation of the various individuals or organizations involved in realizing the benefits anticipated by the OR study through its implementation.

II. MATHEMATICAL PROGRAMMING

The field of mathematical programming is a vast and rich mosaic of problems of optimizing a (real) function subject to constraints.

Unconstrained optimization has occupied mathematicians for centuries, especially since the development of calculus, and continues to do so. Constrained optimization, on the other hand, is a latecomer, with roots dating back to the work of Lagrange in the mid-nineteenth century. In the late 1940s, it came to the forefront with the resolution of the linear programming problem, which, to many scholars, marks the real birth of OR. Since then, the fever has not abated. The following sections give but a glimpse of this nascent field that is full of vitality and promise.

A. Linear Programming (LP)

The LP model may be defined mathematically as (the criterion function)

$$\text{optimize (maximize or minimize) } x_0 = \sum_j c_j x_j \qquad (1)$$

subject to (the constraint set)

$$\sum_j a_{ij} x_j (\leq, =, \text{ or } \geq) b_i \geq 0 \qquad i = 1, 2, \ldots, m \qquad (2)$$

$$x_j \geq 0 \qquad \text{for all} \qquad j = 1, 2, \ldots, n.$$

The basic assumptions of the LP model are (1) linearity (i.e., proportionality) of both criterion and constraints and (2) nonnegativity of variables (whence a feasible solution must lie in the positive orthant of the n-dimensional Euclidean space). The mathematical model has the following economic interpretation. There are n "activities" and $x_j \geq 0$ is the level at which activity j is undertaken. There are m "resources." Activity j, when carried out at unit level, consumes a_{ij} units of resource i. The objective is to optimize a linear function of the levels of the activities.

The LP defined by Eqs. (1) and (2) can be solved by the simplex method, proposed by Dantzig in 1947. For a long time this method reigned supreme because it was

the only known approach to the solution of LPs. However, the repertoire has been enriched by two other approaches: the "ellipsoid" method of Khatchiyan, and the "interior" method of Karmarkar. Active research is underway in several quarters around the world to polish all three approaches, and perhaps develop new ones.

The central theoretical result of LP is duality theory, usually credited to the mathematician von Neumann. It simply asserts that to each LP that is formulated, which is called the "primal LP," there is a "dual LP" that can be easily derived from it. Duality theory helped link LP with the theory of games, originally proposed by von Neumann and Morgenstern in 1944. The "marriage" between these two theories enriched both, and led to spectacular developments in the theory of economic planning.

Special structures of the coefficients of the LP model give rise to specialized, and therefore much simpler, algorithms. This, in turn, enables us to solve much larger problems (as measured by the number of constraints and/or the number of variables) with the same expenditure of (computing) effort. Prominent among these special cases are:

1. *The transportation model*, so designated because it represents the problem of shipping a homogeneous commodity of limited availability from "sources" to satisfy demand at "destinations." Here, all the coefficients of the A matrix of Eq. (2) assume the values $+1$ or 0, and any column vector of A has exactly two nonzero entries. The right-hand side coefficients are arbitrary nonnegative real numbers. Because of the *unimodular* characteristic of the matrix A, if the vector b is integer, then the optimal solution is also integer.

2. *The assignment model*, so called because it represents the problem of assigning n "jobs" to n "machines," in which the "worth" of an assignment varies. (In these scenarios, it is common that the number of jobs equals the number of machines, though such a restriction is not necessary.) The assignment model may be regarded as even a special case of the transportation model, in which the vector b is a vector of $+1$'s. It is solved by a "primal–dual" algorithm dubbed the "Hungarian method" by its originator.

3. *Flow networks*. There are two types of networks that lend themselves to treatment via LP: "regular" or "simple" flow networks, which conserve flow through the arcs, and "generalized" flow networks, in which entering flow in an arc is multiplied by a constant factor as it flows through the arc. Specialized algorithms have been devised for both types of networks, with, as to be expected, simple flow networks providing richer theory and simpler algorithms.

LP also provides the algorithmic vehicle for solving quadratic programming problems. A quadratic program possesses a quadratic criterion function of the form

$$\text{optimize} \quad x_0 = cx + x^t Dx,$$

where D is a symmetric matrix of rank n, assumed to be definite positive, subject to *linear* constraints of the form

$$Ax + IS = b \qquad x, S \geq 0,$$

where S is a slack vector of size m.

LP is, by definition, a deterministic model in continuous variables, the x_j's. However, there are applications in which probabilistic information is included directly in the formulation, yielding stochastic LP. Another approach is via chance-constrained LP, in which the constraints are stated in probabilistic terms.

Finally, LP is a model of optimization of a single criterion function. But it has been extended to optimization under multiple criteria. Then it is better known under the title *goal programming*.

B. Integer Programming (IP), Graphs, and Combinatorial Optimization

Loosely speaking, IP is the domain of mathematical optimization in which some or all of the problem variables are restricted to be integers. Combinatorial optimization is closely linked to IP, and is concerned with seeking the best subset of items (decisions, activities, etc.) satisfying particular criteria from a structured finite set of alternatives. An example of IP would be any LP whose variables are restricted to be integers, in which case one speaks of integer LP (ILP), which is clearly an abuse of language since the integer requirement vitiates the host of properties commonly associated with linearity! An example of combinatorial optimization would be to determine the minimum set of arcs in a connected graph $G = (N, A)$ which "covers" all the nodes.

As was hinted above, the presence of the integer requirement often transforms a solvable problem into an unsolvable one, or at least one that is several orders of magnitude more difficult than its continuous-variable analog. For instance, a LP in one constraint is trivially solvable in one pass by the so-called "greedy algorithm," while the same problem in integer variables is the notorious knapsack problem (KnP), which is NP-complete!

Interest in IP, and in particular in ILP, stems from the ease with which numerous real-life problems can be modeled with the use of IP. For instance, investment problems in which only a subset of the projects can be undertaken at any one time; scheduling problems in which precedence among the activities must be respected; problems in which a fixed cost is incurred only if the activity is undertaken, otherwise no cost is incurred (the so-called fixed-charge problems); and many other "matching," "covering," and "route selection" problems, to name but a few, have been successfully modeled using integer variables. Another

area of application in which ILP proves invaluable is in the piecewise linear approximation of a separable nonlinear function. If the nonlinear function is convex (concave) and the criterion is minimization (maximization), then there is no need to appeal to ILP; but if it is not, then an ILP formulation is mandatory.

ILPs may be solved by the use of the cutting plane method, originally proposed by Gomory in 1958, which treats the problem as a (regular) LP but continues to add constraints to eliminate portions of the feasible convex polytope without cutting any integer feasible point until the optimum integer point is reached. Another solution approach, which has generality beyond the linear case, is implicit enumeration through branch-and-bound (B&B). This approach, if permitted to run to completion, terminates with the optimum, and if it is aborted before that time it yields upper and lower bounds on the value of the optimum together with a feasible, albeit not necessarily optimal, point.

Some ILP problems have the fortunate property that their solution as regular LP problems (i.e., ignoring the integer requirements) automatically assigns integer values to the variables, such as in the classical "transportation" and "assignment" problems (already discussed). Consequently, the regular LP methods are entirely sufficient to solve these problems, and no special IP techniques are required. Other IP problems may possess special structures that invite special algorithms (such as Bender's decomposition procedure in the fixed-charge problem). Finally, it is interesting to note that the *linear* KnP, while being the simplest ILP, is also the archetype of all ILP problems, since it is possible to reduce $m \geq 2$ equations of any ILP to a single Diophantine equation through the appropriate choice of multipliers of the equations and linearly combining the equations with these multipliers.

It is extremely difficult, if not impossible, to disentangle the discussion of IP from that of "combinatorial optimization." Furthermore, the two are entwined with the *theory of graphs*. For instance, the "set covering problem" mentioned may be rightly classified as a "graph theoretic" problem!

A graph is a collection of nodes and arcs connecting them, which is a different kind of graph from the common one in analytic geometry and calculus. The theory of graphs is concerned with the structure of these connections among the nodes. When parameters are added to the graph, one usually refers to the questions asked then as "network" problems. For instance, the following is a graph-theoretic problem (famous under the name Hamiltonian circuit): does there exist a closed path (i.e., a path whose start and terminal nodes coincide) that passes by every node of the graph once and only once? On the other hand, the transportation problem of LP stated above may be represented as a ("bipartite") graph in which the "source" and "destination" locations are represented by nodes, and the route between every source–destination pair is represented by an arc (in this case, the arc is directed from the source node to the destination node). However, since the question asked relates to the cost of shipping, the model is usually referred to as a network model of the LP problem.

The theory of graphs has its genesis in the work of Euler in the eighteenth century (we refer to the famous "seven bridges of Konigsburg" problem and the theory that resulted from it), with sporadic contributions from different mathematicians over the past two and a half centuries. OR called upon, and contributed to, graph theory when it was realized that several discrete optimization problems that originated in the application of OR could be fruitfully modeled as graph theoretic problems. Examples are the matching, set covering, the "number of trees," the "four color," and the "traveling salesman" problems. The nascent field of computer science and technology added to the interest in graph theory. For instance, the theory of combinatorial complexity boasts numerous examples of NP-complete and NP-hard problems from graph-theoretic problems.

C. Variational Inequality and Complementarity Problems (VI & CP)

The existence and computation of economic and game theoretic equilibria have been of great interest to the academic and professional communities since the mid-1960s. The early developments include the fixed point approach and nonlinear optimization models. However, the former lacks computational efficiency and the latter lacks the generality for solving large scale equilibrium problems. Starting in the late 1970s and early 1980s, finite dimensional variational inequality and nonlinear complementarity problems have emerged as promising candidates.

The nonlinear complementarity problem first appeared in Richard Cottle's PhD dissertation in 1964. But the name "*complementarity problem*" was coined later in 1970 by Cottle, Habetler, and Lemke in dealing a linear case problem. The classical "variational inequality problem" was introduced by Philip Hartman and Guido Stampacchia in 1966 in the context of calculus of variations and optimal control theory for solving nonlinear elliptic differential equations. The finite dimensional variational inequality problems has taken its own tangent to become a new field since the late 1970s.

Let X be a nonempty subset of R^n and F be a mapping from R^n to itself. The *variational inequality problem*,

commonly denoted by $VI(X, F)$, is to find a vector $x^* \in X$ such that

$$F(x^*)^T(x - x^*) \geq 0, \quad \text{for all} \quad x \in X.$$

One typically assumes that X is a closed and convex set; in fact, X is often polyhedral in applications.

When F becomes a point-to-set mapping (called multifunctions) from R^n into a subset of R^n, the *generalized variational inequality problem*, denoted by $GVI(X, F)$, is to find vectors $x^* \in X$ and $y^* \in F(x^*)$ such that

$$(x - x^*)^T y^* \geq 0, \quad \text{for all} \quad x \in X.$$

When X is further generalized as a point-to-set mapping in $GVI(X, F)$, then the corresponding problem is called a *generalized quasivariational inequality problem*.

For X being a convex cone in R^n and F being a mapping from R^n to itself, the *generalized complementarity problem*, denoted by $GCP(X, F)$, is to find a vector $x^* \in X$ such that

$$F(x^*) \in Y \quad \text{and} \quad F(x^*)^T x^* = 0,$$

where $Y = \{y \in R^n \mid x^T y \geq 0, \forall x \in X\}$ is the dual cone of X. In case $X = Y = R^n_+$, the first orthant of R^n, the problem becomes the commonly seen (*nonlinear*) *complementarity problem*.

Karamardian was the first one to show that when X is a convex cone, then $x^* \in X$ solves the problem $VI(X, F)$ if and only if x^* solves the problem $GCP(X, F)$. In other words, every generalized complementarity problems is a variational inequality problem, but the converse is not true in general.

The existence and uniqueness of solutions to (generalized) variational inequality and (generalized) nonlinear complementarity problems have been extensively studied in the past three decades. The most basic result on the existence of a solution to the problem $VI(X, F)$ requires that X be compact and convex and F be continuous. Many corollaries have been derived from this basic result by replacing or relaxing the requirements on X and F. The most basic result on the existence and uniqueness of the solution to the variational inequality problem requires that F be strongly monotone over X.

Various computational algorithms including the pivoting scheme, fixed point methods, linear approximation methods, projection methods, simplicial decomposition methods, proximal point methods, cost approximation methods, and continuation methods have been developed in the past several decades. These solution methods basically are the traditional *sequential algorithms*. In order to solve real-life large scale variational inequality and complementarity problems, developing *parallel algorithms* takes high priority.

D. Nonlinear Programming (NLP)

The theory of nonlinear programming is the mathematical theory of optimizing (maximizing or minimizing) a nonlinear real function of a set of variables $x_1, \ldots, x_n$ subject to inequality and/or equality aggregate constraints in which the aggregating (real) functions are also nonlinear in the variables. The general mathematical form is to

$$\text{optimize} \quad f(X), X = (x_1, \ldots, x_n) \tag{3}$$

subject to

$$\begin{aligned} g_j(X) &\leq 0 \qquad j = 1, \ldots, m \\ h_j(X) &= 0 \qquad j = 1, \ldots, k, \end{aligned} \tag{4}$$

where $f: R^n \to R$, $g: R^n \to R^m$, and $h: R^n \to R^k$, that is, f, g, and h are functions that map each point of the n-dimensional real Euclidean space R^n into R, R^m, and R^k, respectively.

As can be suspected, nonlinear programming provides a general paradigm for many problems in the physical and social (in particular, economic) sciences. Unfortunately, unlike LP, there does not exist an algorithm that is universally applicable to all NLP models. What are available, though, are conditions for an optimum to exist, which, in many instances, characterize the optimum sufficiently well to lead to an algorithm for achieving it. Among these we mention the following.

1. The (necessary and sufficient) conditions of Kuhn and Tucker: under the assumptions that the functions f, g, and h are differentiable and that g and h satisfy a particular constraint qualification at the optimum.
2. The (necessary) conditions of Kuhn and Tucker: under the assumptions that f, g, and h are twice continuously differentiable around the optimum.
3. The (necessary and sufficient) first-order conditions of the augmented Lagrangean, due to Mangasarian: under the assumptions that f, g, and h are differentiable and that the constraint qualification is satisfied by g and h at the optimum.
4. The (necessary and sufficient) gradient projection conditions, due to Levitin and Polyak: under the assumptions that f is differentiable at the optimum, g is continuous and convex on R^n, and h is linear.

As in LP, duality theory also plays a central role in theoretical as well as computational investigations in NLP.

Practically speaking, one cannot guarantee global optimality of the solution achieved utilizing any algorithm except in very special circumstances. Research is directed toward speeding up the rate of convergence to a (local) optimum, and the devising of a sequence of starting points

of the algorithm to result in a high probability of capturing the global optimum.

E. Geometric Programming (GP)

Geometric programming is concerned with the optimization of an arbitrary real-valued function g over a subset $S \subset R^n$, which is restricted to be the intersect of the function domain C with an arbitrary cone $X \subset R^n$ (which is, in fact, a vector space for most examples). To fix ideas, consider the following definition of GP:

Determine

$$f = \inf_{x \in s} g(x) \qquad \text{where} \qquad S = X \cap C \tag{5}$$

and the optimal solution set

$$S^* = \{x \in S \mid g(x) = f\}.$$

Each optimization problem can generally be formulated as Problem (5) in more than one way by suitably choosing the function g and the cone X. The most striking example of the utility of GP was demonstrated in the study of the minimization of signomials by Zener and Duffin. A *signomial* (sometimes referred to as a generalized polynomial) is any function with the form

$$P(y) = \sum_i c_i \prod_j y_j^{a_{ij}}, \tag{6}$$

where the coefficients c_i and the exponents a_{ij} are arbitrary constants but the independent variables y_j are restricted to be positive. When the c_i coefficients of the signomials are positive, which is the case in the majority of problems stemming from economic or physical laws of nature, they are termed *posynomials*.

To cast the problem of minimizing $P(y)$ in the format of Eq. (5), simply make the change of variables

$$x_i = \sum_j a_{ij} \log y_j \qquad i = 1, 2, \ldots, n$$

and then infer that minimizing $P(y)$ is equivalent to solving Eq. (5) when

$$C = R^n \qquad g(x) = \sum_i c_i e^x$$

and

$$X = \text{column space of } [a_{ij}].$$

The advantages of studying this problem rather than its signomial predecessor are numerous. For example, unlike the signomial P of Eq. (6), the exponential function g is completely separable in the variables x_i. Consequently, if y^* minimizes a posynomial P, then the corresponding x^* must be a unique optimal solution to Eq. (5); in which event the set of all y that minimize P can be obtained from x^* simply by solving the displayed system of equations, a task that is relatively easy because the system is clearly linear in terms of $\log y_j$, $j = 1, 2, \ldots, m$.

GP can be generalized to explicitly incorporate *constraints*. The mathematical formalism must then be carefully structured to make sharp distinction between "cone conditions" of the form $x \in X$ and constraints of the form $g_i(x^i) \leq 0$, $i \in I$, where I is a positive integer index set.

The theory of GP can be gainfully partitioned into several topics. Optimality conditions describe important properties possessed by all optimal solutions, and in many cases collectively characterize all optimal solutions. Saddle-point characterizations of optimality can be used to introduce the even more significant concepts of duality, which, in turn, provides important existence and uniqueness assertions for optimal solutions, as well as provide useful algorithmic concepts. Duality is also a key ingredient in "parametric programming" and "sensitivity analysis."

F. Dynamic Programming (DP)

The name dynamic programming is not indicative of the scope or content of the subject, which led many scholars to prefer the expanded title: "DP: the programming of sequential decision processes." Loosely speaking, this asserts that DP is a mathematical theory of optimization. It also identifies DP with decision systems that evolve in a *sequential and dynamic* fashion.

DP possesses formalism. But unlike other areas of mathematical programming, many optimization problems that are normally stated in the form of other mathematical programs (such as ILP, NLP) can be cast in the formalism of DP. Consequently, DP is considered to be more of an approach, a way of thinking about a problem, rather than a fixed mathematical statement in which a problem is cast.

A DP model is structured from five elements that reflect the sequential paradigm underlying the approach. These are stages, states, state transformation function, stage reward function, and total system reward function. Following is a brief and somewhat loose description of each of these elements. A stage is an epoch of decision making. A state is either an "input" (i.e., start) or "output" (i.e., terminal) state, and is a parameter or a set of parameters (i.e., a vector) that summarizes the past history of the system and enables one to make a decision (hopefully optimal) at the current stage and all subsequent stages. A (state) transformation function is a mapping from the cartesian product of the input state space and stage decision space to the output state space at each stage. A stage reward is a real function that maps the cartesian product of the input state space and the stage decision space into the real line. Finally, the system reward function is a real function that maps the individual stage rewards into the real line. Let

the system be modeled with N stages, and let n index the stages. Let

s_n denote the input state to stage n, $s_n \in S_n$
x_n denote the decision made, $x_n \in X_n$
$\tilde{s}_n$ denote the output state and be defined by the state transformation function $\tilde{s}_n = t_n(s_n, d_n)$, $\tilde{s}_n \in \tilde{s}_n$
r_n denote the "reward" and be defined by $r_n = g_n(s_n, d_n)$
R_N denote the system reward and be defined by $R_N = R_N(r_1, r_2, \ldots, r_N)$.

Under the DP paradigm, the system is envisioned to evolve in stages. At any stage it starts in some state and, depending on the decision made, two events take place: it is transformed into another (output) state, and a reward is received.

The central concept of DP is embodied in its *principle of optimality*, first enunciated by Bellman, and is traditionally stated as:

An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first transition.

The principle is manifested in the *extremal equation* of DP, given as

$$f(n, s_n) = \text{opt}x_n[r_n(s_n, x_n) \circ f_n(s_n)] \quad \text{for} \quad n = 1, 2, \ldots, N \qquad (7)$$
$$f(0, s_1) = r_0(s_0) \qquad \text{a "terminal" reward}$$

where $s_n \in S_n$, $x_n \in X_n$.

There are several distinguishing themes of DP that are not shared by other mathematical programming models. First, the problem of interest is embedded in a much richer set of optimization problems. (The term "flooding" is sometimes used to indicate this fact.) The price paid for such enrichment is increased computational difficulty, sometimes to the point of impracticality (the so-called "curse of dimensionality"). Second, the problem of concern is transformed from its original setting, which usually involves the simultaneous optimization over N variables, to N optimizations, each in only one variable. Third, if iterations are halted at any intermediate point, one usually has the optimal solution of a partial problem.

The concepts of DP are easily extendable in two important directions. The first is to stochastic systems, that is, systems that involve probabilistic elements either in their evolution (in time, space, etc.) or in their payoff. This extension gave rise to the field of Markov programming, and has spurred significant advances in the theory of optimization of queues. The second is to systems with unbounded number of stages. As $N \to \infty$, the stage designation in the extremal equation of Eq. (7) disappears and the resultant equation is commonly referred to as the "functional equation of DP" because the same function $f(\cdot, \cdot)$ appears on both sides of the equality sign. Finally, there is a strong link between the formalism of DP and the theory of optimal control.

G. Entropy Optimization (EO)

The term "entropy" was coined by Rudolf Clausius around 1865 as a measure of the amount of energy in a thermodynamic system. Claude Shannon gave a new meaning to this term in 1948 as "a measure of uncertainty" in the context of communication theory. With E. T. Jaynes "maximum entropy principle" and its variations, the the concept of using Shannon's entropy to select a probability distribution that complies with known observations about an unknown distribution and yet remains most noncommittal (i.e., the one that carries a maximum amount of entropy) has penetrated a wide range of disciplines in science, engineering, and management.

Let $\mathbf{Y}$ denote a random variable with n possible outcomes $\{y_1, \ldots, y_n\}$ and $x = \{x_1, \ldots, x_n\}$ represents the corresponding probability mass function. Let $g_1, \ldots, g_m$ be m functions defined on $\mathbf{Y}$ with known expected values $a_1, \ldots, a_m$, respectively. The probability distribution determined according to the 'principle of maximum entropy' can be stated as a linearly constrained concave optimization problem:

$$\text{Maximize} \quad \sum_{j=1}^{n} x_j \log x_j$$
$$\text{s.t.} \quad \sum_{j=1}^{n} x_j g_i(y_j) = a_i, \quad i = 1, \ldots, m,$$
$$\sum_{j=1}^{n} x_j = 1, \quad x_j \geq 0, \quad j = 1, \ldots, n.$$

When the expectation is replaced by a moment function of higher order or a general nonlinear function of $\mathbf{x}$, the corresponding entropy optimization problem becomes versatile but difficult. Moreover, when a prior distribution $p = (p_1, \ldots, p_n)$ is given, the objective of achieving "maximum entropy" can be replaced achieving "minimum cross-entropy," which is defined by

$$\text{Minimize} \quad \sum_{j=1}^{n} x_j \log(x_j / p_j).$$

Various entropy optimization models have proven their usefulness with successful applications in image reconstruction, statistical inference, queueing theory, spectral

analysis, statistical mechanics, transportation planning, urban and regional planning, input–output analysis, portfolio investment, information analysis, and linear and nonlinear programming. Major solution methods for entropy optimization include the Bregman's balancing method, multiplicative algebraic reconstruction technique (MART), Newton's method, generalized iterative scaling method (GISM), interior-point method, and the unconstrained convex dual programming approach.

H. "Soft Optimization": Meta-Heuristics and Compusearch Approaches

Formal optimization theory as described in sections II, A through E have their limitations, to varying degrees, as the size of the problem (as measured by the number of variables and number of constraints) increases. Then the frontal attack on the problem becomes infeasible, and one must resort to alternative approaches to achieve meaningful answers. The gamut of approaches proposed for this purpose have been invariably labeled as "soft optimization," "meta-heuristics," and "compusearch" approaches. As the names indicate, these are approaches that do not *guarantee* optimality but approximate it to varying degrees. Also as the names indicate, the approaches rely heavily on *heuristics*, and on *organized search using the computer*.

A possible taxonomy of these approaches is as follows.

1. *Neighborhood search methods* work on complete feasible points (such as a complete schedule or a complete allocation of resources) and try to improve them at each step. They are so-called because they normally stop when they find an optimum, which is most probably a local one, and because at each step they study the local "neighborhood" of a given complete solution. The definition of that "neighborhood" is precisely what distinguishes one approach from another. They normally are constituted from three procedures: one which provides an initial (feasible) point, one which provides other feasible points in the "neighborhood," and a function which provides the "measure of performance" of a given point. They are usually classified depending on the use of these three procedures inside some general schemes. There are *stochastic* and *deterministic* schemes, depending on whether there is use of random decisions within the solution schema; and between *no-null* and *null* threshold schemes, depending on whether a neighbor can be selected that is worse, but only if the loss is smaller than a given computed threshold. The most popular approaches falling in this class are *simulated annealing* (SA), *tabu-search* (TS), and *genetic algorithms* (GA). The first two techniques, SA and TS, are very similar to one another. The last technique, GA, is somewhat more abstract; it can be argued that SA and TS are special cases of GA.

2. *Splitting-up methods* follow the dictum of "divide and conquer." It splits the original problem into two or more subproblems that are smaller in size and hence easier to solve optimally. The ingenuity of the analyst lies in how to split the original problem into the subproblems, and how to recombine the optima of the smaller problems in an answer to the original, global, problem. The splitting may be done *hierarchically* (e.g., strategic vs. tactical vs. *operational*), or *structually* (e.g., resource usage constraints vs. temporal constraints), or over the *set of solutions* such as in branch-and-bound (BaB). *Filtered beam* search methods constitute a further restriction on the BaB approach because it further restricts the search among the "descendents" of a particular node in the BaB search tree to a selected class. Others are ignored.

3. *Model changing methods*, better known as *relaxation methods*, "relax" (i.e., delete or replace) an original requirement of the problem. For example, ignore integer requirements and solve assuming continuous variables, then round off the solution in some rational fashion. Or approximate nonlinear relations by linear one. Or restrict the class of solutions sought, such as in flowshops scheduling problems where the restriction is to "permutation schedules"; etc.

4. *Artificial intelligence methods* include *constraint-guided search, expert systems*, and *knowledge-based procedures*. Limitation of space does not permit a more extensive discussion of these approaches, but they may be briefly described as search methods that are guided by using the available information intelligently. This information may be in the form of the constraints under which the process is functioning; or in the form of how established "experts" have solved the same, or similar, problems before, or on available classification of problems arising in the same class (such classification is sometimes referred to as "*artificial memory*").

III. STOCHASTIC PROCESSES

"Stochastic" is the Greek word for "random," which is much less forbidding. However, stochastic is commonly used to connote a process that evolves randomly over time, which leads to the mathematical definition of a stochastic process (SP) as "an indexed family of random variables," in which the indexing occurs over time.

The presentation that follows is couched in the vernacular of classical concepts of uncertainty, based on the calculus of probability and bivalent logic. Section IV discusses the fuzzy paradigm, which presents a different view of uncertainty that is based on "belonging" as a matter of degree.

Interest in SPs stems from their prevalence in real-world phenomena, such as waiting lines (or queues), Brownian motion, and the analysis of time series.

There are several issues for study in SPs: the construction of models that represent observed phenomena, the theoretical analysis of models of SPs, the solution of problems by either of the two approaches available (analytical and simulation), and the statistical analysis of data from SPs for either estimation of parameters or evaluation of performance. The field has gained prominence in recent years because of the need to optimize the design and utilization of networks of computer systems.

Theoretical research and applications in the field of SPs have resulted in a rough taxonomy along the following lines: random walk, Markov chains, recurrent events, renewal processes, martingales, and cumulative processes. A brief discussion of these topics follows.

Random walk is the name given to the following process. Let $X_1, X_2, \ldots$ be a sequence of real-valued, identically distributed, mutually independent random variables, (r.v.'s) and let $S_0 = 0$, $S_n = \sum_{i=1}^{n} X_i$, for $n = 1, 2, \ldots$. Then $\{S_n\}$ is a random walk. Many questions can be asked relative to this process, some of theoretical interest, and others of practical significance depending on the physical interpretation of the r.v.'s X_t. For instance, suppose X_t is of bounded variation; does there exist a limit to the arithmetic mean S_n/n? Or, if X_t represents profit or loss over a short period (say 1 month), what is the probability that $S_T < A$, where T is a finite planning horizon (say 1 year) and A is a least desirable cash-in-hand amount?

A recurrent event, as the name indicates, is an event that recurs over time. Imagine an event A to happen at times $S_0, S_1, S_2, \ldots$, where $S_n = X_0 + \cdots + X_n$ for $n = 0, 1, 2, \ldots$. The most important underlying assumption is that the r.v.'s $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.), taken to be multiples of some fixed period $w > 0$. (For simplicity, and with no loss of generality, it is usual to take $w = 1$.) Set $f_j = \Pr\{X_n = j\}$, $j = 1, 2, \ldots$. The case $X_0 = 0$ (i.e., when the process starts with event A occurring at the origin of the time scale) is what is usually meant by a recurrent event process. The more general situation, in which X_0 has some arbitrary distribution, is called a delayed recurrent event process. Let u_n be the probability that A occurs at time n, and let N_n be the number of occurrences of A in the time interval $[0, n]$. Thus $u_n = \Pr\{N_n > N_{n-1}\}$. Noting that $u_0 = 1$, it is easy to show that, for example,

$$E[N_n] = 1 + u_1 + u_2 + \cdots + u_n = U_n,$$

and that, if $m_1 = E[X_j] < \infty$, then

$$E[N_n] \approx n/m_1 \qquad \text{as} \quad n \to \infty.$$

If the X's happen (with probability 1) to be multiples of some integer greater than unity, then the recurrent event process is said to be *periodic*; otherwise, it is called *aperiodic*.

Let $f = \sum_1^\infty f_n$. If $f < 1$, then the event A is said to be *transient*; otherwise it is *persistent* (i.e., when $f = 1$). Actually, it is possible to show that event A is transient iff $\sum_0^\infty u_n < \infty$, and that, in such a case, $f = 1 - (1/\sum_0^\infty u_n)$.

Suppose the system under study can exist at any given moment in one, and only one, of a sequence of possible states, identified as, say, $s_0, s_1, s_2, \ldots$, finite or infinite. Suppose that in one unit of time the system "moves" from its present state to another one (which may be the same current state, in which case it is said that the system has undergone a "virtual movement"). Suppose further that, given that it is in state s_i, the probability that it moves to state s_j is p_{ij}, and that the selection of the next state of the system is independent of all previous history of the process except for the fact that it is presently in s_i. Such a system is called *Markov chain*. If the probability of move is dependent on the (absolute) time at which the transition occurs, written $p_{ij}(t)$, then the system is the more general Markov process.

If the transition from state i to state j does not occur in unit time, as was assumed in Markov chains, but the system sojourns (waits) in state i for a period of time whose duration is an r.v., denoted by T_{ij}, then the process is known as a *semi-Markov* process.

Markov and semi-Markov chains and processes are perhaps the most extensively studied models of stochastic processes, because of their valid representation of many real-world situations.

Renewal processes are the continuous-time analog of the recurrent event processes on the "discrete" time scale. Let $\{X_n\}$ be an infinite sequence of i.i.d. nonnegative r.v.'s, with distribution function $F(x) = \Pr\{X_n \le x\}$, and moments $m_1 = E[X_j]$, $m_2 = E[X_j^2]$, etc. We assume the renewal event to occur at times $0, S_1, S_2, \ldots, S_n, \ldots$, where S_n is defined as before. Denote the number of renewals in the time interval $(0, x)$ by $N(x)$. Let $H(x)$ denote the expected number of renewals by time x, the renewal function. Then it is known that it satisfies the integral equation

$$H(x) = F(x) + \int_0^x H(x - z)F(dz),$$

which is a commonly encountered equation in many OR models.

The elementary renewal theorem states that $H(x) \approx x/m_1$; the more profound Blackwell's theorem states that, for every fixed $c > 0$, $H(x + c) - H(x) \to c/m_1$, as $x \to \infty$, with the understanding that $(m_1)^{-1}$ is interpreted

as zero if $m_1 = \infty$. The key renewal theorem states that for a variety of absolutely integrable functions $k(x)$, as $x \to \infty$,

$$\int_0^x k(x-z)H(dz) \to (1/m_1)\int_0^\infty k(z)\,dz.$$

Martingales are models of SPs with the property that the expected value of the nth r.v., conditional upon knowledge of all "earlier" values of the r.v.'s, shall be equal to its $(n-1)$st value; or, mathematically,

$$E[X_n : X_1, X_2, \ldots, X_{n-1}] = X_{n-1}, \tag{8}$$

for $n = 1, 2, \ldots$, where $X_1, X_2, \ldots$ is an infinite sequence of r.v.'s such that $E[|X_j|] < \infty$ for all $j = 1, 2, \ldots$. The condition of Eq. (8) is quite general, since it does not require independence.

Although, to date, martingales have not played an important role in OR, it is anticipated that their use will increase because of the power their theory brings to bear on problems. For instance, a shortcut to the solution of some quite messy problems may be found via the optional stopping result, which asserts that, under some rather general conditions, a martingale (that is relatively easy) is derived from another martingale (that is difficult) by specifying a rule by which the latter may be "terminated."

A stationary process refers to the r.v.'s $\{X(t)\}$, $-\infty < t < \infty$, in continuous time, or to r.v.'s $\{X_t\}$, $-\infty < t < \infty$ and integer, in discrete time. The terms *time series* are commonly reserved for the discrete time stationary process. The sequence $\{X_t\}$ is stationary in the strict sense if the following conditions are satisfied: for every finite integer $n > 0$, every set of n integers $(k_1, k_2, \ldots, k_n)$, and every set of n reals $(x_1, x_2, \ldots, x_n)$, the $\Pr\{X_{k_1+r} \le x_1, X_{k_2+r} \le x_2, \ldots, X_{k_n+r} \le x_n\}$, is independent of the integer r, $-\infty < r < \infty$. Although many OR problems satisfy these conditions, it is acknowledged that an even wider class of problems satisfies the (weaker) secondorder conditions: (1) $E[X_t]$ and $E[X_t]^2$ exist for all t and are independent of n, and (2) if $m = E[X_t]$, the product moment $E[(X_{t_1} - m)(X_{t_2} - m)]$ is a function of $|t_1 - t_2|$ only.

Consider the real-valued second-order stationary process $\{X_t\}$. There exists a distribution function $F(\cdot)$ on the interval $[0, \pi)$ such that

$$\rho(n) = \int_0^\pi (\cos n\lambda)F(d\lambda), \tag{9}$$

for $n = 0, \pm 1, \pm 2, \ldots$ where $\rho(\cdot)$ is the autocorrelation function. The values of the autocorrelation function are thus seen to be the Fourier coefficients of a spectral distribution $F(\cdot)$. The converse of this statement is also true: if $F(\cdot)$ is any distribution function on $[0, \pi)$, then there exists a real-valued stationary (second order) process whose autocorrelation function is given by Eq. (9). In many problems, $F(u) = \int_0^u f(w)\,dw$, and the density function $f(w)$ is called the *spectral density*. The spectral distribution plays a significant role in the analysis of stationary processes.

A. Waiting-Line Theory

Also referred to as "queueing theory," this is the mathematical theory for the analysis and optimization of units demanding service at a facility. The problem is defined in terms of three constituents: (1) the input process: the source of arrivals, the type of arrivals, and the interarrival times; (2) the service mechanism: the number of servers, the number of customers getting served at any time, and the duration and mode of service; and (3) the queue discipline: all the factors regarding the rules of conduct of the queue, such as the priority in which jobs are processed by the facility.

The analysis of queueing systems is aimed at understanding the system behavior relative to certain major characteristics, such as the time variation of the queue length $Q(t)$, the time variation of the customer waiting time $W(t)$, and the time variation of the occupancy of the service mechanism. Various modifications of these characteristics are also of concern, such as throughput time (total time spent by the customer in the system), busy cycle (sum of two adjacent busy and idle periods), and server utilization (fraction of time the server is busy).

There are two problems associated with the study of queueing systems that demand special mention. They are the statistical estimation problem, and the operational problems inherent in the design, control, and measurement of effectiveness of such systems. Evidently, the results obtained from analysis shall depend, to a large extent, on the validity of the estimated parameters and on the operational structure of the system.

In formulating a queueing model for a system, one applies direct observation to the determination of its structure and discipline; but in the determination of the form and properties of the input and service processes, one must rely on statistical techniques. These comprise four major steps: (1) collection of data, (2) tests for stationarity in time, (3) tests for independence of arrival times and/or service times, and (4) tests for specific distributions, which involve the estimation and tests of hypotheses regarding specific parameters and forms of distribution. The most popular models structured for such studies are the Poisson process of events, the Erlangian model, and the hyperexponential model.

1. Stationarity

Real-life systems exhibit some sort of stable behavior in the long run. In terms of system characteristics, this

translates into the properties of the processes $\{Q(t)\}$ and $\{W(t)\}$ as $t \to \infty$ and $\{Q_n\}$ and $\{W_n\}$ as $n \to \infty$, where Q_n and W_n refer to the discrete-time analogs. In such a case, denote, for convenience, $Q(\infty)$ by Q and $W(\infty)$ by W, while retaining Q_∞ and W_∞ for the discrete time measures. As always, one seeks system performance indices, preferably given by simple measures. One of the most used performance measures is the server utilization, given by the traffic intensity

$$\rho = \frac{\text{effective arrival rate}}{\text{effective service rate}}.$$

2. The Single-Server Queue with Poisson Arrival and Exponential Service

This is the so-designated "M/M/1 queue," where the first symbol designates the input behavior, the second the output behavior, and the third the number of servers, and the "M" indicates Markovian. This process has been used more often than any other system in queueing theory. Let λ and μ be the arrival and service rates, respectively. Then the traffic intensity $\rho = \lambda/\mu$. Let $\pi_j = \Pr\{Q = j\}$, $j = 0, 1, 2, \ldots$. We have

$$\pi_j = (1-\rho)\rho^j \qquad j = 0, 1, 2, \ldots$$

$$E[Q] = \frac{\rho}{1-\rho} \qquad \text{and} \qquad \mathrm{Var}[Q] = \frac{\rho}{(1-\rho)^2}.$$

When the queue discipline is FCFS (first-come first-served), and noting that the waiting time of a customer is equivalent to the service load of the server, one gets

$$F(x) = \Pr\{W \le x\} = 1 - \rho e^{-\mu(1-\rho)x}$$

$$E[W] = \frac{\rho}{\mu(1-\rho)}$$

and

$$\mathrm{Var}[W] = \frac{\rho(2-\rho)}{\mu^2(1-\rho)^2}.$$

The following two relations are then immediately derived:

$$\frac{1}{\mu}E[Q] = E[W]$$

and

$$E[Q] = \lambda E\left[W + \frac{1}{\mu}\right].$$

Of these, the first is true only when the service times are exponential, and the second relation is true under very general conditions.

The speed with which the distribution of $Q(t)$ approaches its limiting value $\{\pi_j\}_{j=0}^{\infty}$ is given by its relaxation time $t \approx 2\lambda/(\mu-\lambda)^2$. Further, the asymptotic behavior of $E[Q(t)]$, $\mathrm{Var}[Q(t)]$, and $\Pr\{[Q(t)/\sqrt{2\lambda t}] \le x\}$, as well as similar results for $W(t)$, have been evaluated.

Apart from the completely random input, exponential service, and single-server queue, others have come under scrutiny. Unfortunately, the results of these models are not as rich as the M/M/1. We name a few that have received more extensive treatment.

The queue M/G/1, in which the input is Markovian, the service distribution is "general," and there is only one server.

The queue GI/M/1, in which the input possesses general distribution, the service is Markovian, and there is only one server. Results similar to those of the queue M/G/1 are available.

The queue GI/G/1, in which both arrival and service systems possess general distributions, but still one server. Let $F(x) = \Pr\{W_\infty < x\}$ and $K(x) = \Pr\{X_n \le x\}$; then

$$F(x) = -\int_{0^-}^{\infty} F(y)\,dK(x-y) \qquad (x \ge 0).$$

In special cases this equation may be solved to derive exact distributions. For the applicationsoriented individual, the following three results are significant.

1. Little's formula: $E[Q] = \lambda E[W + V]$, where V is an r.v. representing the service time.
2. Heavy-traffic approximation: when the traffic intensity is close to 1, and under the usual assumptions of independence, the waiting time distribution is approximately negative exponential with mean

$$\frac{1}{2}\left[\frac{\text{interarrival time variance} + \text{service time variance}}{\text{mean interarrival time} - \text{mean service time}}\right]$$

if the denominator is small compared with the square root of the numerator.

3. Bounds for $E[W]$:

$$l \le E[W] \le \frac{\text{interarrival time variance} + \text{service time variance}}{2(\text{mean interarrival time})(1-\rho)}$$

where l is the unique solution of the equation

$$x = \int_{-x}^{\infty} [1 - K(u)]\,du \qquad x \ge 0, \qquad \rho < 1.$$

3. Multi-Server Queues

These have been extensively analyzed under the assumption of Markovian arrival and service mechanisms. Let λ and μ be the arrival and service rates and $\rho = \lambda/s\mu$, where s is the number of (identical) servers. Then

$$\pi_j = \begin{cases} \dfrac{(s\rho)^j}{j!}\pi_0 & 0 \le j \le s \\ \dfrac{(s\rho)^j}{s^{j-s}s!}\pi_0 & j \ge s, \end{cases}$$

where

$$\pi_0 = \left[\sum_0^{s-1} \frac{(s\rho)^j}{j!} + \frac{(s\rho)}{s!}(1-\rho)^{-1} \right]^{-1}.$$

Also,

$$E[Q] = s\rho + \frac{\rho\pi_s}{(1-\rho)^2}$$

and

$$E[W] = \frac{\pi_s}{s\mu(1-\rho)^2}.$$

4. Finite Queues

The real world is replete with systems in which $Q(t)$ has a finite limit. Such systems can be purely loss systems or loss and delay systems. In a loss system, customers arriving when all the servers are busy are denied service and are lost to the system. In a loss and delay system, such customers wait provided the number in the system does not exceed a fixed number, which is the upper bound on the size of the queue.

5. Other Queueing Systems

Many queueing systems encountered in practice are more complex than the systems discussed above, which, in the most part, have simple structure and queue disciplines. The complexity is reflected in the three characteristics enumerated in the first paragraph. For instance, arrival may be in bulk, or the service is offered in quanta of fixed length, or the queue discipline incorporates priorities that change dynamically based on the arriving customer characteristics, etc.

B. Simulation

It would be presumptuous of OR to claim simulation as its own, since simulation has been advanced by other researchers, such as physicists, and used by them extensively long before the advent of OR on the scene. However, it is a fact that OR has adopted simulation, and in particular Monte Carlo methods, as among its most prominent tools, and thus helped popularize the approach well beyond the confines of theoretical studies in particle physics.

To simulate, according to a dictionary definition, is "to feign, to attain the essence of without the reality." In substance, therefore, every representation is a form of simulation and must involve some reduction or change from reality. As it is commonly used in the OR community, simulation is understood to be "digital simulation," and stands for "the use of numerical computation and, in particular, statistical sampling of a mathematical model to arrive at results such as estimates of probabilities of certain responses" [Gaver (1978) in Bibliography].

The degree of difference between the simulation and the real system depends, of necessity, on the objective of simulation. That objective is usually either the understanding of the behavior of the system (which may lead to the discovery of some "laws" governing the system), or the evaluation of various strategies relative to a criterion or a set of criteria in the hope of choosing the best decision (from among the given set of alternatives).

All simulation models are input–output models. In other words, they are descriptive, since they represent the structure of the systems they purport to emulate. They are also functional models, since they yield the output of any system given the input to its interacting subsystems. Another way to characterize a simulation model is that it is an "if . . . then . . . " device; that is, if a certain input is specified, which includes any strategy, then the output can be determined.

It immediately follows that a simulation is a model that is "run" rather than "solved" in order to obtain results. Consequently, by its very nature, simulation is incapable of generating its own "answer." The latter must be provided from the "outside." Simulation is therefore a tool of analysis of the behavior of the system under any specified conditions. Furthermore, simulation is not a theory—rather, it is a problem-solving methodology. The following is a brief list of the most prominent among the reasons for adopting the methodology of simulation.

1. Simulation may be used in the absence of any compressed and complete mathematical formulation of the problem or when great difficulty is encountered in solving such a formulation if one exists.
2. Simulation may be the only possible way of experimentation because of the difficulty of conducting the experiments and observing the phenomena in their actual environment.
3. It is a very powerful educational and training device because it enables the experimentor to "play with" the simulation.
4. It affords complete control over time, since a phenomenon may be "speeded up" or "slowed down" at will in the simulation.
5. It can be used as a check on analytical solutions.

Some simulations may involve humans in the conduct of the experiment, and these are referred to as participative simulations. In such cases, the model must be constructed to cope with external inputs to the simulation from the participating individuals. When competition is present as an integral part of the simulation, it is usually referred to

as "gaming." Such "play" of the simulation is often used in business and warfare contexts.

The conduct of a simulation study, assuming that the problem has been well defined, generally involves five steps.

1. Model formulation.
2. The definition and collection of the necessary data.
3. The design of experiment and the set up of hypotheses tested.
4. The running of the simulation, and the statistical analysis of its results.
5. Model validation.

The final phase of a simulation study, or, for that matter, any study, is to communicate the results of the study to the interested party who sponsored the study in the first place.

Simulation, and in particular Monte Carlo simulation, has spawned several research areas, of which we name the following two.

1. The generation of random numbers—that is, numbers that can be considered as uniformly distributed between 0 and 1. For the generation of random variables that follow a given distribution, the following methods have been used: the transformation method (for the exponential, Gaussian, and log-normal distributions); the composition method (for the gamma family, Poisson, and binomial distributions); and the rejection method (also for the gamma function as well as the beta distribution).
2. Variance reduction. These are sampling methods that result in estimates that possess smaller variance than would otherwise be obtained by "straightforward" sampling. The approaches are: antithetic variates, stratification, importance sampling, and control variates and regression.

Simulation may be viewed as a methodology that rests on three premises: the *model*, the *probabilistic framework* for the inputs and *statistical* design and analysis of the outputs, and the *language* used in constructing the simulation. There are several commercially available languages, all of which are designed to relieve the analyst of the chore of writing his or her own computer code, and to let the analyst concentrate on the other two aspects of the simulation, which are typically the analyst's forte. Among the more popular of discrete event simulation languages available are GPSS/H, Awesim, and Arena. A more recent development in the design and construction of simulations is *object-oriented simulation* (OOS). It views the process being simulated as a set of interacting "objects," hence the name, and the modes of interaction are themselves defined by other "objects." The object collections, or "classes," encapsulate the characteristics and functionality of the entities pertaining to the domain of interest. Simulations thus constructed have distinct advantages over others; in particular, the use of "*encapsulation*" (the attachment of pertinent characteristics to the objects), "*inheritance*" (the use of previously developed classes of objects), *polymorphism* (the same property applies to different objects), *run-time binding* (the tying of variables to the values generated at run time), *extensibility* (the ability of the user to modify the internal logic of the simulation), and *parametrized typing* (the characterization of the object class *via* parameters of the process). Some of these characteristics are shared by other non-OOS languages, but only OOS-based languages possess all of them. The ease and power of OOS can be easily gleaned from its optic. The concept of OOS simulation started with Simula, followed by Smalltalk, which eventually led to Eiffel and C++ as the general programming languages for object-oriented programming. A lucid exposition of the philosophy and structure of OOS simulations may be found in Joines and Roberts (1998).

The idea of simulation is so patently simple that it is not easy to resist the temptation to adopt the simulation approach in almost every problem one encounters. Care should be taken when such adoption is decided upon because of the following reasons. (1) To construct a simulation is a time-consuming and expensive undertaking that requires a high caliber of talent that is often not available. (2) There are several problems that are intrinsic to the process of simulation and to the manner in which a simulation is constructed, which, if not resolved in the correct manner, may confound the experiments and yield wrong results. (3) Simulation is imprecise and there is no measure of the degree of its imprecision. Unless the experimental design was laid out well at the outset, one can easily get swamped in the data spewed out by the computer that signifies very little. (4) Because the output of a digital simulation is in numbers, a certain degree of definitiveness is accorded to the results, and more is read in them than is really justified.

C. Decision Analysis

The term "decision analysis" (DA) is typically used to denote the activities of "rational" decision in the face of uncertainty. The definition of "rationality" is embodied in a set of axioms that, if accepted by the decision maker, should also require acceptance of the procedures and consequent actions. By its very definition, DA is a normative theory of decision: it does not claim to explain how individuals decide, but rather how they should decide.

DA is based on the assumptions that (a) a decision results in a set of outcomes, about which the decision maker possesses judgment (subjective or otherwise) relative to their likelihood of occurrence, and (b) the decision maker exhibits preference among the outcomes on the basis of its utility.

Decision problems that are legitimately considered in the purview of DA usually exhibit some or all of the following characteristics: multiple objectives and criteria, which are necessarily conflicting; the presence of intangibles and/or nonmeasurables; uncertainty in the data (which is typically sample data rather than population data), as well as in the outcome of a decision and its payoff (or utility); group decision makers who do not necessarily share the same utility function; and nonstatic behavior over time.

It is common to discuss DA relative to four stages.

1. Structuring the decision problem, which includes the specification of the alternative courses of action, the delineation of the objectives and the criteria, and the specification of the degree to which these objectives might be achieved by the various alternatives.
2. Describing likelihoods of the possible consequences of each alternative decision in quantitative terms.
3. Assigning preferences to the possible consequences on the basis of individual or group utility functions.
4. Deciding rationally among the alternative courses of action on the basis of the information provided. This step also contains sensitivity analysis.

Reference has been made above to the axioms of DA. These have been and continue to be the subject of extensive studies that are aimed at reducing them to the bare minimum independent set. A consequence of these axioms is the acceptance of the principles of maximizing the expected utility as the ultimate criterion in the process of decision. Some of these axioms have come to bear distinctive names, such as the axiom of existence of relative preference, or the axiom of transitivity, or the axiom of substitutability.

Theoretical issues in DA span three areas. The first is the problem of assessment of utility of an individual decision maker or a group of individual decision makers. The second is the quantification of judgemental uncertainties, that is, the assignment of probability measures to the various possible outcomes of a decision. And the third is the methodological approaches to determining the best decision. Imbedded in this last issue are the problems of the value of "perfect information" and the decision to acquire (or not to acquire) additional information; and the issue of sensitivity of the value of the outcome to the various model parameters.

D. Theory of Games

The theory of games is the brainchild of the great mathematician von Neumann, and it was first presented to the scientific community in the (now classical) book by von Neumann and Morgenstern in 1944. It is a normative theory of behavior in situations of conflict, and does not attempt to describe actual human behavior but specifies "rational" behavior, where the definition of "rationality" is based on a set of axioms that, if accepted, dictate the outcome. The theory of games is a mathematical model par excellence. Apart from its intellectual achievement, the theory has found application in situations of conflict in oligopolistic economics and in defense. In economics, it represents a departure from the "Robinson Crusoe" economics of a single entrepreneur in a free market, which is the model that had dominated economic thought for over two centuries.

A conflict situation, henceforth referred to as a "game situation," is either between two parties (a "two-person" game) or among several parties (an "n-person" game). In either case, the total "wealth" of all players is either fixed or variable. In the former case, the gain of one (or some) player(s) must be balanced by the loss of the other(s), and the game is referred to as "zero-sum." In the latter case, the total wealth of all players may be increased or decreased, and the game is referred to as "non-zero-sum." The fundamental difference between the two-person and the n-person game is the possibility of the presence of "collusion" among some of the players in the n-person game, which is clearly absent from the two-person game (else the game situation would be vitiated).

1. Representation of a Game

A game is represented either in extensive form or in matrix form. The extensive form of an n-person game gives, in logical order, the several possible moves in a game. Each move is assigned either to a player (personal move) or to chance (random move). At a personal move, the options available to the player and the information given to the player are made explicit. At a random move, a probability distribution is specified. Finally, at each terminal position of the game, a payoff or outcome can be expressed by a vector $(p_i, \ldots, p_n)$, where p_i represents the utility to player i of the given outcome.

Generally, the extensive form is represented by a game tree with a distinguished node as the initial node. Thereafter, each node of the tree represents a position of the game, and each arc a move. Information is indicated by the use of information sets: essentially, two positions belong to the same information set if the same player must move at each position and the player cannot distinguish between them.

A player "moves" according to a particular strategy, which tells which alternative to choose at each information set. If a player has k information sets, and r_i alternatives at set i, then, in total, the player has $r_1 r_2 \cdots r_k$ strategies for the game. The normal form of an n-person game is the function that assigns the vector of expected payoffs $p_i(\cdot)$ to each n-tuple of strategies. An n-tuple $(\sigma_1, \sigma_2, \ldots, \sigma_n)$ of strategies is said to be an equilibrium n-tuple if, for every player i, and for every strategy s_i of player i,

$$p_i(\sigma_1, \sigma_2, \ldots, \sigma_i, \ldots, \sigma_n) \geq p_i(\sigma_1, \sigma_2, \ldots, s_i, \ldots, \sigma_n).$$

Consequently, none of the players can "gain" by a unilateral change of strategies.

2. Zero-Sum Two-Person Games

The two players are commonly referred to as I and II. The gain to one player, say I, is necessarily the loss to II, and conversely. Then the payoff to both players can be represented by a single number a_{ij}, which, by convention, is the "gain" of I when I adopts decision i and II adopts decision j. (These two decisions are referred to as pure strategies, as opposed to randomized strategies, which assign a probability measure to each possible decision.) By a "strategy set" is meant the set of decisions available to a player. If the strategy sets S_{I} and S_{II} are finite, then the normal form can be represented by a matrix $A = (a_{ij})$, where the rows represent (pure) strategies for I, the columns are (pure) strategies for II, and a_{ij} is the payoff (from II to I) if I chooses the ith row, and II the jth column. An equilibrium pair, or a saddle point, is defined as a pair of (pure) strategies $\sigma^* \in S_{\mathrm{I}}$ and $\tau^* \in S_{\mathrm{II}}$ for which

$$P(\sigma^*, \tau) \geq P(\sigma^*, \tau^*) \geq P(\sigma, \tau^*),$$

where P is the payoff to player I. The fundamental theorem of two-person zero-sum games (the minmax theorem) asserts the existence of pair of randomization vectors x and y ($x_i \geq 0$ and $\Sigma x_i = 1$; similarly for $\{y_j\}$) so that

$$\min_{y \in Y} \max_{x \in X} x^t A y = \max_{x \in X} \min_{y \in Y} x^t A y.$$

The common value of these two is known as the value of the game. The variables $\{x_i\}$ and $\{y_j\}$ are determined as the solution of a linear program and its dual.

3. Some Extensions

1. *Constrained games*. These arise when there is some extraneous constraint on the mixed strategies, so that not all mixed strategies are permissible. The optimal constrained strategies are obtained by solving a pair of mutually dual linear programs.

2. *Infinite games*. These arise when the sets of pure strategies for a game are infinite. In general, the loss of compactness destroys the validity of the minmax theorem. The most extensively studied form of this class is the so-called "game on the square," in which the pure strategy sets of either player are isomorphic to the unit interval [0, 1]. Unfortunately, no analytic methods exist for solving such games except through approximation by finite games.

3. *Multistage games*. A multistage game occurs when, at fixed intervals, both players are given perfect information. Each of these intervals marks a stage of the game. In this case, the problem can be simplified by the employment of dynamic programming.

4. *Differential games*. These are, essentially, multistage games with continuous time. The transition from state to state is then described by means of differential equations. The usual assumption is that two players together control the motion of a particle in n-dimensional space. Thus, if the particle is in position x at time t, its motion is determined by a differential equation that embodies the control vectors ϕ and ψ of the two players. The game ends when the point x reaches some preassigned terminal surface. Two types of payoff are usually adopted: an integral payoff $\int_0^T K(x; \phi; \psi)\, dt$ for some function K from the beginning of the game at time 0 until its termination at time T; and a terminal payoff equal to some function defined on the terminal surface of the game. More generally, the total payoff may be expressed as a sum of these two.

4. Two-Person Non-Zero-Sum Games

The essential difference is that now the gain of one player need not be the loss to the other. Consequently, the two players need not act against each other and may act cooperatively to gain an advantage to both. Under such cooperative play, preplay communication, binding contracts, and correlated strategies are permitted.

Noncooperative games are characterized by two payoff functions $A(\sigma, \tau)$ and $B(\sigma, \tau)$ defined for all pairs of strategies (σ, τ), and corresponding to the payoffs for players I and II, respectively. If the strategy sets are finite, then the functions A and B can be represented as matrices $A = [a_{ij}]$ and $B = [b_{ij}]$, and the game is said to be a bimatrix game. In general, equilibrium pairs of pure strategies need not exist. However, it has been shown that every bimatrix game has at least one equilibrium pair of mixed strategies. Unfortunately, equilibrium pairs for non-zero-sum games are not a totally satisfactory solution concept, because, first, there may be more than one equilibrium pair but with distinct payoffs to the two players, and, second, the equilibrium pair need not yield the best payoff to both players!

In cooperative games it is assumed that both players can do better by acting together than separately against each other. This cooperation must be paid for, thus giving rise to bargaining. In the simplest case, the game is defined by a subset S of the Euclidean space, representing the (joint) payoffs that the two players can obtain by cooperation, and two numbers, u_0 and v_0, representing the amount that each player can obtain without the other's cooperation. It is agreed that S should be closed, convex, and bounded above. Then, the set of Nash axioms leads to the conclusion that the players "should" agree to obtain that point (u^*, v^*) in S that maximizes the product

$$(u - u_0)(v - v_0),$$

subject, of course, to $u^* \geq u_0$, $v^* \geq v_0$. This point is known as the Nash point, and can be shown to be unique.

In the more general case, each player can affect the other even when no agreement is reached. Then the game will have a more complicated structure, consisting of two payoff functions $A(\sigma, \tau)$ and $B(\sigma, \tau)$, which will be obtained if no agreement can be reached, and a set S similar to that discussed above. The idea then is that the strategies σ and τ are used as threats, to be implemented only if no agreement is reached. Then, with each player knowing the other's threat, bargaining can be conducted. It stands to reason then that the outcome of the bargaining, and of the game itself, will depend—to some extent at least—on the threats made. Then the objective would be to maximize the product

$$[u - A(\sigma, \tau)][v - B(\sigma, \tau)],$$

subject to the constraints $(u, v) \in S$, $u \geq A(\sigma, \tau)$, and $v \geq B(\sigma, \tau)$. From this point of view, the point (u^*, v^*) is the logical outcome of the threat pair (σ, τ). The equilibrium pairs (σ, τ) are now defined in terms, not of the (threat) payoffs, but rather of the logical outcome (u^*, v^*).

5. *n*-Person Games

As in the two-person non-zero-sum games, a distinction is made between cooperative and noncooperative games. For noncooperative games, however, there is little difference between the n-person and the two-person theories.

For cooperative n-person games, the interest is on collusion among the players, which is exhibited in the formation of coalitions. Thus such games are usually studied in their characteristic function form, which tells us how much utility the members of a coalition can guarantee for themselves (if the coalition forms). It is assumed that this global utility can be divided among the members of the coalition according to predetermined side payment conditions. Thus, if $N = \{1, 2, \ldots, n\}$ is the set of players and S is a subset of N, then the characteristic function is a function v that assigns a real number $v(S)$ to the subset S satisfying

$$\begin{aligned} v(\Phi) &= 0 \qquad & v(S \cup T) &\geq v(S) \\ &+ v(T) \qquad & \text{if } S \cap T &= \Phi \end{aligned}$$

and the superadditivity condition, which is sometimes weakened to

$$v(S \cup \{i\}) \geq v(S) + v(\{i\}) \qquad \text{if} \quad i \notin S.$$

The elements of N are the players, and the non-empty subsets of N are the coalitions.

Some additional basic concepts of n-person games are S-equivalence normalization, normalization, dominance, the core of a game, a stable set, the Shapley value, the bargaining set, and no-side-payments games (which assume that only certain divisions of the winnings of a coalition are possible or permissible).

E. Gaming

Gaming is the use of games not as normative models of behavior but to study and/or deduce human behavior under conditions of conflict. The exercise is used in teaching and training, experimentation, entertainment, and therapy. It is also used in "operational gaming" to determine the likely outcome of policies that may be chosen by the decision maker.

IV. THE FUZZY PARADIGM

For many centuries, scientists insisted that physical science should strive for certainty in all its manifestations (like the precise law of Newtonian mechanics), and uncertainty was regarded as unscientific and should be avoided by all means. By the end of the 19th century, the study of physical processes at the molecular level made scientists realize that uncertainty is not only an unavoidable plague, but it is essential to the understanding of nature. This understanding led to the creation of statistical mechanics. The role played by probability theory, which captures uncertainty of a certain type, in statistical mechanics is just like calculus, which involves no uncertainty, in Newtonian mechanics. Starting with the 20th century, scientists began to notice that most problems of interest and substance are usually nondeterministic, but not as a result of randomness that could yield probabilistic distributions. The seemingly unique connection between uncertainty and probability theory is no longer taken for granted, and it was felt that for the advancement of science and technology a non-probabilistic paradigm is needed in dealing with uncertainty.

Even though some modern concepts of uncertainty, such as Max Black's *vagueness*, were envisioned in the early part of the 20th century, it is generally agreed that Lotfi A. Zadeh's publication of "Fuzzy Sets" in 1965 is truly monumental. The new theory mathematically defines sets with boundaries that are not precise by assigning a 'degree of belongingness' to each of the elements. It challenged not only probability theory as the sole agent for uncertainty, but also Aristotle's two-valued logic that probability theory is based on. In the new paradigm, science is no longer "black or white"—"To be *and* Not to be" could coexist to certain degree. This capability of expressing gradual transition from membership to non-membership provides scientists not only a meaningful measurement of uncertainty, but also a meaningful representation of vague concepts expressed in natural language.

As pointed out by Zadeh, a fundamental paradigm shift underlying the organization of Fuzzy Set Theory is through *fuzzification* (using gradual membership). A field X and a theory Y can be fuzzified by replacing the concept of regular "crisp" sets in X and Y by that of fuzzy sets. In application to fields such as arithmetic, topology, graph theory, logic, mathematical programming, neural networks, stability theory, pattern recognition, and decision making, fuzzification leads to fuzzy arithmetic, fuzzy topology, fuzzy graph theory, fuzzy logic, fuzzy mathematical programming, fuzzy neural networks, fuzzy stability theory, fuzzy pattern recognition, and fuzzy decision making. What we are witnessing today and what we expect to witness tomorrow is a growing number of fuzzified fields and theories—a grand paradigm shift.

A. Fuzzy Sets

Let X be a collection of objects of interest. A *fuzzy set* $\tilde{A}$ in X is a set of ordered pairs

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) \mid x \in X\},$$

where $\mu_{\tilde{A}}(x)$ is a mapping from X to a membership space M, called the "membership function," that specifies the grade of membership (or degree of compatibility or degree of truth) of x in $\tilde{A}$. M is usually taken to be the closed real interval $[0, 1]$. In case M is taken to be $\{0, 1\}$, then $\tilde{A}$ reduces to a regular crisp set A in X.

The membership function is obviously the crucial component that defines a fuzzy set. The basic set-theoretic operations on fuzzy sets can also be defined through the membership functions. Given that $\tilde{A}$ and $\tilde{B}$ are two fuzzy sets in X, then

1. The intersection of these two fuzzy sets is a fuzzy set $\tilde{C} = \tilde{A} \cap \tilde{B}$, with a membership function
$$\mu_{\tilde{C}}(x) = \min\{\mu_{\tilde{A}}(x),\ \mu_{\tilde{B}}(x)\}, \qquad \forall x \in X;$$
2. The union of these two fuzzy sets is a fuzzy set $\tilde{D} = \tilde{A} \cup \tilde{B}$, with a membership function
$$\mu_{\tilde{D}}(x) = \max\{\mu_{\tilde{A}}(x),\ \mu_{\tilde{B}}(x)\}, \qquad \forall x \in X;$$
3. The complement of fuzzy set $\tilde{A}$ is a fuzzy set $\tilde{A}^C$, with a membership function
$$\mu_{\tilde{A}^C}(x) = 1 - \mu_{\tilde{A}}(x), \qquad \forall x \in X.$$

In a more general setting, the "min-operator" for intersection (the "logical and") and the "max-operator" for union (the "logical or") can be replaced by the so-called "t-norms" and "s-norms," respectively.

B. Fuzzy Measures

A fuzzy measure is an extension of the classical measure, essentially by relaxing the "additivity property." Let X be a set of interest (universe) and F be a Borel field of X. A set function g defined on F is called a *fuzzy measure*, if the following conditions are met.

1. $g(\phi) = 0$, $g(X) = 1$.
2. If $A, B \in F$ and $A \subseteq B$, then $g(A) \leq g(B)$.
3. If $A_n \in F$ and $A_1 \subseteq A_2 \subseteq \cdots$ then $\lim_{n\to\infty} g(A_n) = g(\lim_{n\to\infty} A_n)$.

Commonly seen fuzzy measures include the "belief measures," "plausibility measures," "necessity measures," and "possibility measures." In particular, the ordinary probability measure falls in the intersection of the plausibility and belief measures.

C. Fuzzy Arithmetic

One class of the most frequently used fuzzy sets is the so-called fuzzy numbers. A *fuzzy number* $\tilde{n}$ is a fuzzy set whose membership function assumes positive values in an interval around the real number n and peaks at n. For manipulation purposes, a formal definition of $\tilde{n}$ specifically requires that $\mu_{\tilde{n}}(n) = 1$ and the left branch of the membership function is piecewise continuous and nondecreasing (starts with value 0 and ends with value 1), while the right branch is piecewise continuous and nonincreasing (starts with value 1 and ends with 0).

Fuzzy numbers can perform arithmetic operations like ordinary numbers, but special attention should be given to make sure that a positive fuzzy number does not assume any non-positive value at any degree, similarly, a negative fuzzy number does not assume any non-negative value at any degree. Given that $\tilde{m}$ and $\tilde{n}$ are two fuzzy numbers and

"$\circ$" is a binary arithmetic operation (like addition, multiplication, subtraction, and division), then the membership function of a new fuzzy number $\tilde{m} \circ \tilde{n}$ is defined by the extension principle as

$$\mu_{\tilde{m}\circ\tilde{n}}(z) = \sup_{z=x\circ y} \min\{\mu_{\tilde{m}}(x), \mu_{\tilde{n}}(y)\}, \quad \forall z \in R.$$

The concept of fuzzy numbers plays a key role in formulating *quantitative fuzzy variables*, those variables whose states are fuzzy numbers. When, in addition, the fuzzy numbers represent linguistic concepts, such as *small*, *very tall*, *sort of young*, as interpreted in a particular context, the resulting constructs are usually called *linguistic variables*.

D. Fuzzy Relations

Fuzzy relations are widely used in the fuzzy paradigm. A fuzzy relation is an extension of an ordinary relation. It allows the expressions involving ambiguity such as "x and y are almost the same" or "z is much bigger than w."

Let X and Y be two sets of interest. A *fuzzy relation* $\tilde{R}$ between X and Y (or from X to Y) is a fuzzy set of the form

$$\tilde{R} = \{((x, y), \mu_{\tilde{R}}(x, y)) \mid x \in X, y \in Y\}.$$

In case $X = Y$, $\tilde{R}$ is known as a *fuzzy relation on* X. When X and Y are finite sets, say $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$, then a fuzzy relation can be represented as a *fuzzy matrix*:

$$\tilde{R} = [\mu_{\tilde{R}}(x_i, y_j)], \quad i = 1, \ldots, m; \quad j = 1, \ldots, n.$$

If $\tilde{R}$ is a fuzzy relation between X and Y and $\tilde{S}$ is a fuzzy relation between Y and Z, then the *composition* of these two fuzzy relations is a fuzzy relation $\tilde{R} \circ \tilde{S}$ between X and Z, whose membership function can be defined via the extension principle as:

$$\mu_{\tilde{R}\circ\tilde{S}}(x, z) = \max_{y\in Y} \min\{\mu_{\tilde{R}}(x, y), \mu_{\tilde{S}}(y, z)\}, \quad \forall x \in X, z \in Z.$$

The "max" and "min" operators can be replaced with general "s-norms" and "t-norms" for different applications. A very important class of fuzzy applications is to characterize a fuzzy system with fuzzy input and output by the so-called "fuzzy relational equation."

Let $\tilde{A}$ be a fuzzy set of X (viewed as a fuzzy relation on X) and $\tilde{R}$ be a fuzzy relation from X to Y, then the composition of $\tilde{A}$ and $\tilde{R}$ defines a fuzzy set $\tilde{B}$ in Y by the following *fuzzy relational equation*

$$\tilde{A} \circ \tilde{R} = \tilde{B},$$

with

$$\mu_{\tilde{B}}(y) = \max_{x\in X} \min\{\mu_{\tilde{A}}(x), \mu_{\tilde{R}}(x, y)\}, \quad \forall y \in Y.$$

In the above fuzzy relational equation, $\tilde{R}$ represents the dynamics of a fuzzy system with fuzzy input $\tilde{A}$ and fuzzy output $\tilde{B}$. A "fuzzy diagnosis system" first uses some known pairs of fuzzy input and output to build the so-called "cause-symptom" fuzzy relation $\tilde{R}$, and then use individual "fuzzy symptom" ($\tilde{B}$) to identify (infer) possible "causes" ($\tilde{A}$). How to effectively solve fuzzy relational equations remains a challenging job, even though many algorithms have been proposed by researchers in the field.

E. Fuzzy Logic

Logic is the study of the methods and principles of reasoning in all its possible forms. As bases for reasoning, different logics can be distinguished essentially by three "context independent" items, namely, the "truth values," "vocabulary" (operators), and "reasoning procedure" (tautologies, syllogisms).

In the classical Boolean logic, the truth value of a logic variable (say A, B, etc.) can be 0 (representing false) or 1 (representing truth), but not both at the same time. The vocabulary (operators) is defined by these truth values via the well-known "truth table." Actually, all operators can be represented as a combination of a sequence of "not" and "or" operators (or, equivalently, a sequence of "not" and "and" operators). For example, the "implication" operation of (A implies B) can be represented as (*not* (A) or B). The reasoning procedure is generally based on the tautologies, such as

- *modus ponens*: (A and (A implies B)) implies B,
- *modus tollens*: (*not* (B) and (A implies B)) implies *not* (A),
- *syllogism*: ((A implies B) and (B implies C)) implies (A implies C)

Note that every tautology remains a tautology when any of its variables is replaced with an arbitrary logic formula. This property is fundamental to the "rule of inference," referred to as a "rule of substitution" in the reasoning procedure.

Fuzzy logic is an extension of set theoretic multivalued logic in which the logic values are linguistic variables that may assume any value in the interval [0, 1]. Although there exist different philosophies to define fuzzy vocabulary (logic operators), given that $\tilde{A}$ in X and $\tilde{B}$ in Y, a common practice based on Zadeh's original idea defines the following.

$$\mu_{not(\tilde{A})}(x) = 1 - \mu_{\tilde{A}}(x), \quad \forall x \in X,$$

$$\mu_{\tilde{A}\ and\ \tilde{B}}(x, y) = \min\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(y)\}, \quad \forall x \in X, y \in Y,$$

$$\mu_{\tilde{A}\ or\ \tilde{B}}(x, y) = \max\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(y)\}, \quad \forall x \in X, y \in Y,$$

$$\mu_{\tilde{A}\Rightarrow\tilde{B}}(x, y) = \max\{1 - \mu_{\tilde{A}}(x), \mu_{\tilde{B}}(y)\}, \quad \forall x \in X, \quad y \in Y.$$

Another commonly practiced definition (less meaningful, but easy to compute) due to E.H. Mamdani is to replace the last equation with

$$\mu_{\tilde{A}\Rightarrow\tilde{B}}(x, y) = \min\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(y)\}, \quad \forall x \in X, \ y \in Y.$$

Fuzzy reasoning is referred to as "approximate reasoning" As mentioned before, in the classical Boolean logic, inference rules are based on the various tautologies. Fuzzy logic generalizes these inferences to facilitate approximate reasoning by using the so-called "composition rule of inference." In fact, approximate reasoning can be carried out by the composition rule of fuzzy relations.

A *generalized modus ponens* reads like

$$(\tilde{A}' \text{and} (\tilde{A} \text{ implies } \tilde{B})) \text{ implies } \tilde{B}'.$$

Note that when $\tilde{A}' = \tilde{A}$ and $\tilde{B}' = \tilde{B}$ in the crisp case, we have the classical *modus ponens*. The deviation of fuzzy set $\tilde{A}'$ from $\tilde{A}$ results in an approximation in the reasoning process. For example, the fact of ((this tomato is very red) and the rule of (if a tomato is red then it is ripe)) results in a conclusion of (this tomato is very ripe). Note that a fuzzy implication rule, $\tilde{A} \Rightarrow \tilde{B}$, is actually a fuzzy relation from X to Y. Moreover, the fuzzy set $\tilde{A}'$ is a fuzzy relation on X itself. Hence the generalized modus ponens can be represented by a fuzzy relational equation

$$\tilde{B}' = \tilde{A}' \circ (\tilde{A} \Rightarrow \tilde{B}),$$

with

$$\mu_{\tilde{B}'}(y) = \max_{x\in X} \min\{\mu_{\tilde{A}'}(x), \ \mu_{\tilde{A}\Rightarrow\tilde{B}}(x, y)\}, \quad \forall y \in Y.$$

In a similar fashion, the *generalized modus tollens* can be defined as

$$(\tilde{B}' \text{ and } (\tilde{A} \text{ implies } \tilde{B})) \text{ implies } \tilde{A}'.$$

Note that when $\tilde{B}' = \tilde{B}^C$ and $\tilde{A}' = \tilde{A}^C$ in the crisp case, we have the classical *modus tollens*. We also have

$$\mu_{\tilde{A}'}(x) = \max_{y\in Y} \min\{\mu_{\tilde{B}'}(y), \mu_{\tilde{A}\Rightarrow\tilde{B}}(x, y)\}, \quad \forall x \in X.$$

As to the *generalized syllogism*, one can view it as

$$((\tilde{A} \text{ implies } \tilde{B}) \text{ and } (\tilde{B} \text{ implies } \tilde{C})) \text{ implies } (\tilde{A} \text{ implies } \tilde{C})$$

with

$$\mu_{\tilde{A}\Rightarrow\tilde{C}}(x, z) = \max_{y\in Y} \min\{\mu_{\tilde{A}\Rightarrow\tilde{B}}(x, y), \mu_{\tilde{B}\Rightarrow\tilde{C}}(y, z)\}, \quad \forall x \in X, \ z \in Z.$$

More complicated fuzzy reasoning involving more than two premises can always be reduced to the unions (and intersections) of simple fuzzy reasoning with two premises. This property is particularly important in the area of 'fuzzy control' which applies a set of fuzzy 'if-then rules' to gain control for desired output of a fuzzy system.

V. SELECTED MODELS OF COMMON PROCESSES

OR has been successfully applied in a wide spectrum of operational systems. In this section is an abbreviated roster of these fields of application.

A. Aggregate Production Planning

One of the earliest fields of applications of OR was in the planning of aggregate production capacity and production quantity for the firm, as exemplified by the (now classical) study of Holt, Modigliani, Muth, and Simon in 1960 on a paint company. Briefly, assuming that quadratic cost functions may be considered as reasonable approximations to the actual costs of varying the workforce, the production schedule, and inventory about their respective "ideal" values, the authors were successful in modeling the total cost incurred by the firm over the finite horizon, and in deriving their famous "linear decision rules," which purport to regulate all three variables mentioned above. Several extensions of the basic model have been proposed in the literature since its appearance, but all lack the elegant simplicity of the original contribution.

If linear approximations are acceptable, even within a narrow interval, then linear programming can be used to advantage.

The "economic manufacturing quantity" (EMQ) is perhaps the most venerated of all simple mathematical models for optimizing production, and has been used by industrial engineers since 1916, long before the advent of OR. Unfortunately, the assumptions of the model were extremely restrictive: infinite planning horizon, continuous time, constant demand that is infinitely divisible, no stock-out condition permitted, and only two costs to contend with: the cost of setup and the cost of storing a unit of the product per unit time, both of which were assumed constant over time and invariant with the quantity produced. With the methodology of OR in hand, it has been possible to step out of the confinement of the EMQ model to the "economic lot scheduling problem" (ELSP) under discrete time and dynamically varying demand from period to period for a finite planning horizon, and, furthermore, to treat stochastic demand, with backlogging of demand permitted, with or without limitation on capacity, with linear or nonlinear costs that may vary with time, with or without the so-called "fixed charge" penalty that is incurred whenever setup is undertaken.

B. Sequencing and Scheduling

To sequence a finite set of tasks is to define the order in which they are to be performed. If the time at which each task is started is also defined, one ends with a schedule. Clearly, then, a schedule implies a sequence (and in fact the two terms are often used interchangeably in deterministic systems), but the converse need not be true in the case of uncertain task durations and fewer processors than tasks. In the latter case, it is understood that what is discussed is the sequence, although one continues to speak of the "schedule."

Attempts have been made to establish a strict taxonomy of scheduling problems, but with little success, mainly because of the extensive variety that can be met in practice, and the minute differences that exist among many of them. Typically, the problem is defined in terms of the following parameters.

1. The number of jobs to be scheduled. It is typically assumed that all tasks are available at the same time. However, it is easy to deviate from this assumption when job availabilities vary and the precise number of jobs to be processed is not known *a priori*.

2. The number of machines available for processing the jobs, the "route" of each job over the machines, and its processing times. The one-machine case, despite its deceiving simplicity, gives rise to extremely knotty scheduling problems, which are best exemplified by the renowned traveling salesman problem, in which a salesman is required to call on each of N cities once, and return to his base city having covered the shortest distance possible. In the case of $m > 1$ machines, the very arrangement of these machines relative to the job routes gives rise to an astronomical variety of problems. For instance, if all jobs possess the same route, with one machine per station, then the shop is known as a flowshop. Of course, flowshops exist in which stations have several machines "in parallel," which may or may not be identical. On the other hand, if there is no discernible pattern to the job routes, then one is faced with a jobshop. Setup times (and setup costs) may or may not be present, and they may or may not be sequence-dependent.

3. The pattern of processing the jobs. It is normally assumed that a machine can process one job at a time and, moreover, that each job can be processed at any time by only one machine. However, deviations from this assumption are quite common, where a machine (or operator) serves two or more jobs simultaneously (time-shared) or, more commonly, a job is served by two or more machines at the same time. Usually the jobs are permitted to queue ahead of a work station, but instances exist where such queueing is prohibited.

4. The relation, if any, among the jobs. Normally, the jobs are assumed to be independent. But they may also be "related" to each other either by precedence or some other relation. Assembly operations are common examples of jobs related by "tree-type" precedence.

5. The criterion function. Here, one again encounters a wide variety of measures of performance, depending on whether or not the jobs are "due-dated." These criteria typically refer to (a) the jobs (e.g., minimize the number of jobs tardy), (b) the values of the jobs (e.g., minimize the total cost of tardiness for due-dated jobs). (c) the machines (e.g., maximize the minimum machine occupancy), and (d) the value of the machine times (e.g., minimize the cost of machine idle time).

The methodologies used in resolving scheduling problems are as varied as the problems posed, and range over the full spectrum of discrete and combinatorial optimization, branch-and-bound methods, and heuristics. The latter methodology is gaining in popularity because the overwhelming majority of scheduling problems are in the class NP-hard, so that there is little hope of finding a procedure that resolves large-scale problems in reasonable and finite time.

C. Inventory Management

This has also been one of the earliest fruitful areas of application of OR, since it is intimately tied to production capacity planning, though not necessarily so. Development has progressed along two main streams of thought: the first is pragmatic and descriptive, and attempts to measure the various components of the system (statistical properties of demand for stocked items, the cost elements of carrying and replenishing inventories, the stratification of items according to some scheme, etc.); the second is formal and prescriptive, and attempts to specify what to stock, how much to stock, and when and where to stock it in the case of multiechelon inventory and distribution systems.

Pure inventory systems (i.e., ones that are divorced from production) are rare, and are typically found in applications such as water resources management and retail stores, Typically, however, in agriculture as well as in industry, inventory is linked to production, although the link may be intentionally weakened to divorce one from the other. This is done at the expense of increasing inventory.

An inventory problem is defined in terms of the following parameters.

1. The number of items in stock, their demand characteristics [deterministic or stochastic, constant (stationary) or variable (dynamic) over time], their physical characteristics (volume, dimensions, weight, physical/chemical

properties, and their rate of deterioration, if any), their relations to each other (if any), and their value and volume–price variation over the planning horizon.

2. The structure of the inventorying system, which may be single-echelon or multiechelon.

3. The criterion function. Basically, there are two schools of thought: one relates to "cost," or value, and the other to "service level." The two are intertwined, since "customer satisfaction" or the lack of it may be recast in monetary terms, though often this is well nigh impossible to accomplish because of the presence of nonmeasurable entities in the concept of "service" to customers.

Inventory management under stochastic demand draws heavily on, and has contributed significantly to, the theory of stochastic processes. In addition, it has spurred the rapid development of the theory of dynamic programming.

D. Project Planning and Control

For decades, the only quasianalytical tool available to project managers for the planning, control, and evaluation of performance of projects was the reputable Gantt chart. Then the concepts of CPM (for critical path method) and PERT (for program evaluation and review technique) appeared in 1959 in the seminal papers of Kelley and Walker and of Malcolm, Roseboom, Clark, and Fazar, respectively, and transformed the picture radically, ushering in the era of activity networks as the models for the interacting activities of the project. Both models (CPM and PERT) represent an activity, either by an arrow (in the activity-on-arc mode of representation) or by a node (in the activity-on-node mode of representation), and represent precedence by the direction of the arrow in the first mode, and by an arrow linking the nodes in the second mode of representation. The difference between the two models lies in the attempt of PERT to take cognizance at the outset of the fact that activity durations are random variables. The two models have enriched the vocabulary of OR with terms such as the critical path, activity floats (four such floats are defined for each activity), event slack, the probability distribution function of project completion time, path criticality index, and activity criticality index.

CPM and PERT have spurred intensive and extensive research in theoretical as well as applied problems that have plagued managers for a long time, especially relative to the following:

1. The estimation of the various statistical parameters in the PERT model, such as the probability distribution function of the various "key events" of the project, the criticality indices of paths and of activities, the ranking of activities according to their criticality, etc. Such estimation is accomplished through any of four approaches: exact analytical models, approximation schemes, bounding methods, and Monte Carlo sampling.

2. The optimal project "compression" in the CPM model under specified time–cost tradeoff functions, such as linear, concave, convex, and quadratic.

3. The optimal scheduling of activities and/or allocation of limited resources to the activities (or, alternatively, the optimal acquisition of resources and their allocation to activities) to complete a project in a specified time. Due to the extreme complexity of this problem (which has been demonstrated to be in the class of NP-hard problems), the approach to its resolution has been typically through the use of heuristics.

The models of CPM and PERT take the logic of the precedence relations among the activities of the project as given and that all the activities shall be deterministically realized. It is easy to visualize situations in which this need not be true. Such realization gave rise to the concepts of GERT (for graphical evaluation and review technique) and GAN (for generalized activity networks). These two models added much needed flexibility in the analytical (especially, simulation) study of complex systems.

Project planning and control algorithms have been programmed on almost all mainframe computers, and there is ample availability of such procedures on mini- and microcomputers. Such software does not presently involve any optimization or probability calculations, but is limited to performing the calculations on time and resource utilizations for any specified input in the CPM model. It is anticipated that the future will see further developments along these lines.

E. Facilities Location and Layout

The term "facility" can be defined broadly to encompass a factory (or a system of factories), a school (or a system of schools), a hospital (or a system of health delivery units), a warehouse (or a system of warehouses), etc. Problems of location and layout may refer to either a totally new set of facilities, or to a new facility (or facilities) that fit into an existing system, or to the redesign of an old facility.

The criteria adopted in the design of facilities typically relate to either cost or benefit, and are usually directed to either optimizing a weighted sum of functions of the costs incurred (or the utilities derived), or to the "minmax" (or "max-min") vernacular (e.g., maximize the minimum utility derived by the users).

A distinction is usually made between "continuous" and "discrete" location and facility design problems, mainly

because of the difference in methodologies used for the resolution of the problems posed. These methodologies are either mathematical programming or discrete and combinatorial optimization, respectively.

Real-life facility location and design problems are usually of formidable character that defies strict analytical treatment—hence the resort to approximations and heuristics. At the present time, there exists a large number of computerized facility layout programs, some of which are available on the microcomputer, and none of which claims optimality of the solution reached.

The nascent field of "microchip" design has spurred renewed interest in the optimization of layout in two and three dimensions, since there is great incentive to carefully place the maximum number of "gates" that perform different logic functions on the chip to achieve maximum packing of logic and extremely high speed of operation. Though the limits of physical proximity may have been recently approached, the optimization of the layout can still stand improvement.

F. Maintenance and Replacement

Maintenance and replacement usually stand for the variety of activities that are undertaken in controlling the condition of "equipment," including inspection, repair and overhaul, and replacement. OR has found a fertile field of application in determining the timing and degree of action, especially when such action involves the disruption of operations, such as when the equipment in question is a part of a larger system.

The basic purpose of inspection is to determine the condition of the equipment and, depending on the result of such determination, take action, which may range from "do nothing" to replacement. The equipment may be operated intermittently or continuously, and may or may not deteriorate at deterministic or stochastic rates.

Overhaul is taken to mean the restorative maintenance action that is taken before equipment has reached a defined failed state, while a repair is made after the failed state has occurred. The main decisions associated with overhaul and repair are the determination of (1) the interval between overhauls and (2) the degree to which equipment should be overhauled or repaired.

Replacement implies that the "new" condition is achieved on completion of the action. Deterministic replacement problems refer to equipment whose behavior (e.g., deterioration) may be described deterministically, while stochastic replacement problems refer to equipment whose behavior is uncertain and can best be described in probabilistic terms. The literature discusses "finite" versus "infinite" planning horizons, though the difference may be insignificant since a long, although finite, horizon with small time unit behaves very much like an infinitely long horizon.

Maintenance and replacement, whether or not they are analytically studied through OR methodologies, do not just "happen," but are made to happen through strategic and tactical decisions on the part of management to install such actions in a cohesive total system. The activity impacts many other facets of the enterprise that hitherto were thought to be unrelated, or at best weakly related, such as product quality, system productivity, and incentive payment to employees. Recognition of these facts has given added emphasis to studies in this area.

VI. SELECTED FIELDS OF APPLICATION

This article has chosen a few "models of common processes" for expanded treatment. Here, it is desired to mention a few of the "fields of human activities" to which OR has been applied. Obviously, the two discussions are intertwined. For instance, if inventory is discussed under the tutelage of a "model of a common process," it also clearly refers to the activity of inventory management as a "field of human activity." The same is true for production, project planning and control, etc.

Bearing this in mind, one can still legitimately contend that government is an overall activity and a field of human involvement, rather than a model of anything. OR has found application in all levels of government: local, state, and federal. In particular, OR has been, and continues to be, extensively involved in defense studies, where it originated, as was pointed out in the historical perspective given at the beginning of this treatise. Education is clearly a local and state responsibility, and OR has been applied to its manpower planning, facility planning, scheduling of classrooms, routing of school buses, resource allocation, and racial balance. It is a matter of taste whether health systems are included in discussions of government services or not; but OR has indeed been applied to the design and operation of health systems in all its aspects (emergency services, medical information systems, out- and in-patient services, patient screening, manpower planning). The infrastructure for transportation systems is the responsibility of government, either local or federal. Starting with traffic demand forecasting, OR has been applied to the design of the network and the facilities on the network that serve specific purposes (such as toll booths, maintenance and repair facilities), as well as to the modification of existing transportation networks through expansion or reorientation. Equilibrium, or steady-state, traffic flow conditions, bottleneck links, minimal cutsets that separate specific nodes (or subsets of nodes) of the network from each other, and minimal cover

nodes (i.e., the minimum number of nodes whose removal, together with the links incident on them, would disconnect all nodes of the network) have all been legitimate problems of transportation networks that have been studied by the methods of OR.

Furthermore, industry that is involved in transportation, such as the airline and trucking industries (and to a lesser extent the railroads) have extensively used OR in all aspects of their respective operations: capacity planning, resource allocation, short-term scheduling of operations, maintenance, repair, and replacement.

The utilities have seen extensive application of OR. In fact, telephony may be credited with the invention and subsequent development of waiting-line theory. It also has given impetus to, and has provided a fertile field of application for, the science and art of simulation, both analog and digital. Electric utilities have also provided a fertile field of application for OR in what is termed "operations planning," which includes "dispatching" [i.e., the allocation of electric load (demand for electricity) among electric power generators], hydrothermal coordination (when the source of power is multifaceted), security (i.e., operating the system so that creditable disturbances, such as the sudden loss of the most heavily loaded generating unit or transmission circuit, will not cause the outage of any other facility), maintenance scheduling, generation expansion planning, and transmission expansion planning. The last two applications involve heavy investment in plant and personnel that extends over a horizon of several decades and must rely on accurate forecasts of future demands for electricity.

SEE ALSO THE FOLLOWING ARTICLES

DYNAMIC PROGRAMMING • FUZZY SETS, FUZZY LOGIC, AND FUZZY SYSTEMS • GAME THEORY • GRAPH THEORY • LINEAR OPTIMIZATION • NONLINEAR PROGRAMMING • SIMULATION AND MODELING • SIMULATION, REALISTIC (ENGINEERING) • STOCHASTIC PROCESSES

BIBLIOGRAPHY

Bertsekas, D. (1987). "Dynamic Programming: Deterministic and Stochastic Models," Prentice-Hall, Englewood Cliffs, NJ.

Birtwistle, G. M., et al. (1973). "SIMULA Begin," Petrocelli. Charter, New York.

Coffman, E. G., Jr. (ed.) (1976). "Computer & Job/Shop Scheduling Theory," Wiley, New York.

Cottle, R. W., Pang, J.-S., and Stone, R. E. (1992). "The Linear Complementarity Problem," Academic Press, Boston.

Denardo, E. V. (1982). "Dynamic Programming: Models and Applications," Prentice-Hall, Englewood Cliffs, NJ.

Dreyfus, S. E. (1965). "Dynamic Programming and the Calculus of Variations," Academic Press, New York.

Elmaghraby, S. E. (1978). "Activity Networks: Project Planning and Control by Network Methods," Wiley, New York.

Fang, S.-C., and Peterson, E. L. (1982). "Generalized variational inequalities," *J. Optimiz. Theory Applic.* **38,** 363–383.

Fang, S.-C., Rajasekera, J. R., and Tsao, H.-S. J. (1997). "Entropy Optimization and Mathematical Programming," Kluwer Academic Publishers, Norwell, MA.

Fang, S.-C., and Puthenpura, S. (1993). "Linear Optimization and Extensions: Theory and Algorithms," Prentice Hall, Englewood Cliffs, NJ.

Garey, M. R., and Johnson, D. S. (1979). "Computers and Intractability: A Guide to the Theory of NP-Completeness," Freemanl, San Francisco, CA.

Gaver, D. P. (1978). Simulation theory. *In* "Handbook of Operations Research," Vol. 1 (eds. J. J. Moder and S. E. Elmaghraby). Van Nostrand-Reinhold, New York.

Goldberg, A., and Robson, D. (1989). Smalltalk-80: The Language, Addison-Wesley, Reading, MA.

Harker, P. T., and Pang, J.-S. (1990). "Finite dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications," *Math. Program.* **48,** 161–220.

Hax, A. C., and Candea, D. (1984). "Production and Inventory Management," Prentice-Hall, Englewood Cliffs, NJ.

Heyman, D. P., and Sobel, M. J. (1982, 1984). "Stochastic Models in Operations Research," Vols. 1 and 2. McGraw-Hill, New York.

Joines, J. A., and Roberts, S. D. (1998). "Object-oriented simulation," Chapter 11 in "Handbook of Simulation," (J. Banks, ed.), John Wiley & Sons, New York.

Kapur, J. N., and Kesavan, H. K. (1992). "Entropy Optimization Principles with Applications," Academic Press, Boston.

Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. *Combinatorica* **4,** 373–395.

Kaufmann, A., and Gupta, M. M. (1985). "Introduction to Fuzzy Arithmetic, Theory and Applications," Van Nostrand-Reinhold, New York.

Khachiyan, L. G. (1980). Polynomial algorithms in linear programming. *Zhu. Vichixlitel' noi Matematiki i Matematichesdkoi Fiziki* (in Russian) **20,** 51–58.

Klir, G. J., and Yuan, B. (1995). "Fuzzy Sets and Fuzzy Logic—Theory and Applications," Prentice Hall, Upper Saddle River, NJ.

Lawler, E. L., Lenstra, J. K., Rinnooy Kan, A. H. G., and Shmoys, D. B. (1986). "The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization," Wiley, New York.

Lippman, S. B. (1991). "C++ Primer," 2nd ed., Addison-Wesley, Reading, MA.

Mangasarian, O. L., Meyer, R. R., and Robinson, S. M. (eds.) (1975). "Nonlinear Programming 2," Academic Press, New York.

Markowitz, H. (1978). SIMSCRIPT. *In* "Encyclopedia of Computer Science and Technology" (eds. J. Belzer, A. T. Holtzman, and A. Kent). Marcel Dekker, New York.

Meyer, B. (1992). Eiffel: "The Language," Prentice-Hall, Upper Saddle River, NJ.

Miser, H. J., and Quade, E. S. (1985). "Handbook of Systems Analysis," North-Holland, New York.

Moder, J. J., and Elmaghraby, S. E. (eds.) (1978). "Handbook of Operations Research," Vols. I and II. Van Nostrand Reinhold, New York.

Nemhauser, G. L., and Wolsey, L. A. (1988). "Integer and Combinatorial Optimization," Wiley, New York.

Papadimitriou, C. H., and Steiglitz. K. (1982). "Combinatorial Optimization: Algorithms and Complexity," Prentice-Hall, Englewood Cliffs, NJ.

Patriksson, M. (1999). "Nonlinear Programming and Variational Inequality Problems—A Unified Approach," Kluwer Academic Publishers, Dordrecht, Netherland.

Portmann, M.-C. (1997). "Scheduling methodology: Optimization and compusearch approaches," Chapter 9 in "The Planning and Scheduling of Production Systems," (Artiba, A., and Elmaghraby, S. E., eds.), Chapman & Hall, London.

Pritsker, A. A. B. (1984). "Introduction to Simulation and SLAM II," Wiley, New York.

Sauer, C. H., and Chandy, K. M. (1981). "Computer Systems Performance Modeling," Prentice-Hall, Englewood Cliffs, NJ.

Trefethen, F. N. (1954). History of operations research. *In* "Operations Research for Management" (eds. J. F. McCloskey and Florence N. Trefethen). Johns Hopkins Press, Baltimore, MD.

Whittle, P. (1983). "Optimization Over Time: Dynamic Programming and Stochastic Control," Wiley, New York.

Zadeh, L.A. (1965) "Fuzzy sets," *Inform. Control*, **8,** 338–353.

Zimmermann, H.-J. (1993) "Fuzzy Set Theory and Its Application," 2nd ed. Kluwer Academic Publishers, Boston/Dordrecht/London.



Percolation

D. Stauffer
University of Cologne

A. Aharony
Tel Aviv University

I. Introduction
II. Simple Static Properties of Percolation
III. Fractals, Scaling, and Renormalization
IV. Kinetic Aspects of Percolation
V. Modifications and Generalizations
VI. Other Developments
VII. Conclusion

GLOSSARY

Animals Clusters in which every cluster containing the same number of sites has the same importance in averages over all clusters; correspond to percolation clusters with concentration $p \to 0$.
Bond Connection between two neighboring sites.
Cluster Group of neighboring occupied sites.
Concentration Probability p that a site is occupied.
Conductivity Ratio of electric current density to electric field.
Diffusion Motion on a lattice where at every step the new direction is selected randomly, that is, independent of the previously selected direction.
Fractal Object the mass of which increases with $(\text{length})^D$, where D differs from the Euclidean dimension d of the lattice into which the fractal is embedded.
Incipient infinite network Infinite network at the percolation threshold.
Infinite network In an infinite lattice, an infinitely large cluster; in finite systems, usually the largest cluster.
Site Single element of a lattice.
Threshold That concentration p of occupied sites at which for the first time an infinite network appears in an infinite system if p increases; denoted as p_c and often called the percolation threshold.

THE THEORY OF PERCOLATION aims to obtain quantitative estimates for the properties of disordered systems. In its simplest mathematical version, one considers a periodic lattice of sites, each of which is randomly occupied (with probability p) or empty (with probability $1 - p$). Such a lattice contains clusters of neighboring occupied sites. A simple example is shown in Fig. 1. As the concentration increases from zero, larger and larger clusters appear. The mean size of these clusters grows with p and diverges at a well-defined threshold concentration p_c. For $p > p_c$ there exists an "infinite" cluster, which connects the two sides of an arbitrarily large sample.

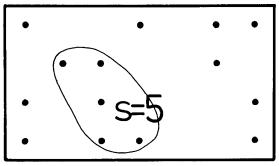


FIGURE 1 Example of site percolation and clusters.

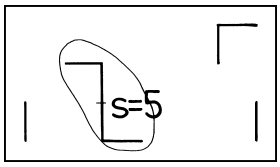


FIGURE 2 Example of bond percolation and clusters.

I. INTRODUCTION

Although the percolation problem is easily defined by one simple rule, it cannot be solved exactly. As we show in Section II, the model has very interesting properties, which exhibit universal features (independent, e.g., of details of the lattice). These features are closely related to the special geometric structure of the infinite cluster at p_c, which exhibits self-similarity, that is, invariance under the change of length scale. Graphs that have such properties, called fractals, can be described as having noninteger (fractal) dimensionality.

In this article we restrict ourselves to the simplest meaning of the complex concept of self-similarity: A structure is self-similar if subsections of the structure have a mass M proportional to some power of their linear dimension L, at least for large L:

$$M \propto L^D$$

In special cases, like the Sierpinski structures to be mentioned in Section III, the concept of self-similarity attains a more direct geometric meaning, which for percolation clusters is valid only in the average sense.

The absence of a basic length scale inherent in self-similar structures is shared by many systems that undergo continuous phase transitions. Power laws like $M(L) \propto L^D$ appear in these structures for similar and other quantities. Scaling theory describes the interrelation of the exponents of the power laws describing different quantities. The study of the changes in the properties of systems when the length scale is changed was the basis for 1982 Nobel Laureate K. Wilson's renormalization group theory explaining scaling and universality.

Although the percolation problem is defined in geometric and statistical terms, it has had many applications in the physical sciences. Historically, the first percolation theory came from chemistry in connection with polymer gelation during World War II. Usually, polymers are linear structures, but sometimes branches occur. Macromolecules with many branches may form an infinite network that is no longer a liquid solution. The formation of pudding is an example of this so-called sol-to-gel transition or gelation; the pudding is a jelly. Another example is the boiling of an egg, where the initially more liquid egg becomes more solid after it is heated for some time. In these cases, the molecules are always there, but the chemical bonds between them are either formed or not formed. For the primitive case of our lattice this means that we need here a modified form of percolation: All sites are occupied, but bonds between neighboring sites are formed randomly with probability p and remain absent with probability $1 - p$. A cluster is a group of neighboring sites connected by bonds formed between them.

We call the second type of problem bond percolation (Fig. 2), whereas the first case defined earlier is called site percolation (Fig. 1). P. Flory, and by a different method W. Stockmayer, solved the gelation problem with approximations that make it equivalent to bond percolation on the so-called Bethe lattice of Fig. 3, which contains no cyclic links and which ignores the constraint that two different parts of the macromolecule have to keep a certain minimum distance from one another and cannot occupy the same volume. This Flory–Stockmayer, or classical, theory plays a role in gelation similar to that of the van der Waals equation for the gas-to-liquid transition or the molecular field approximation for the ferromagnetic Curie point. In particular, all three examples predict a sharp phase transition; that is, a qualitative change takes place in the system at a certain value of a continuous variable. With increasing temperature T, the difference between a liquid and its vapor vanishes at the critical point, and the spontaneous magnetization of a ferromagnet vanishes at the Curie temperature. Similarly, with increasing time the heated polymer solution gels at the gel point.

What is the nature of this phase transition? Figures 1 and 2 make it plausible that the polymer solution becomes more solidlike once an infinite network of connected molecules is formed, that is, once an infinite cluster

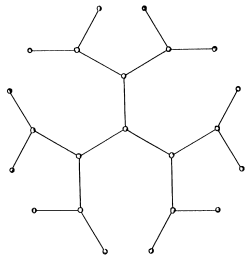


FIGURE 3 Central part of a Bethe lattice.

is formed in an infinitely large sample at the percolation threshold p_c. In the Flory–Stockmayer theory that threshold is

$$p_c = 1/(f-1) \tag{1}$$

where the functionality f is the maximum number of bonds a molecule can form with its neighbors. Equation (1) holds for both site and bond percolation on the Bethe lattice ($f = 3$ in the example of Fig. 3). The name percolation and its application to regular lattices, as in Figs. 1 and 2, were introduced in 1957.

In spite of its history, the gelation phenomenon is not that application on which modern percolation theory is usually tested. The main interest of physicists in percolation centered on the critical phenomena in the immediate neighborhood of the percolation threshold p_c, described by critical exponents to be defined later. Not many experiments are performed with real polymers to determine reliably the critical exponents of gelation. The most precise information exists about the behavior of the viscosity, which diverges as one approaches the sol-to-gel transition from the sol. However, here the connection between the viscosity and the geometric cluster properties of percolation theory is not entirely clear. Clearer are the predictions of percolation theory for the fraction of mass contained in the infinite network or for the average mass and diameter of clusters as measured by light scattering. There now seems to be a consensus that experimental critical exponents for masses and diameters agree with those of lattice percolation, while elasticity and viscosity are still under discussion.

More successful was the application of percolation to thin films of materials sputtered onto a surface. Small droplets form, which grow to larger sizes if more material is added. Here a one-to-one correspondence between laboratory experiment and computer simulation of percolation was established, for example, for the fractal nature of the largest cluster, as reflected by the dependence of its "mass" on the system size.

II. SIMPLE STATIC PROPERTIES OF PERCOLATION

Perhaps the most fundamental question of percolation theory is, what is the value p_c of the percolation threshold; that is, what is the probability p_c for which an infinite network is first formed in an infinite lattice? This problem can be investigated by exact mathematics, by expansion of suitable quantities into power series in p or $1-p$, and by computer simulation. The results for common two- to four-dimensional lattices are given in Table I. Here the results for triangular site, square bond, triangular bond, and honeycomb bond percolation are exact; the others are numerical estimates, usually accurate to the last decimal given here. We see that Eq. (1) is not fulfilled but describes the general trend: The more neighbors a site has, the smaller is the percolation threshold.

TABLE I Percolation Thresholds for Common Lattices

Lattice	Site	Bond
Honeycomb	0.696	0.653
Square	0.593	0.500
Triangular	0.500	0.347
Simple cubic	0.312	0.249
bcc[a]	0.245	0.178
fcc[b]	0.198	0.119
$d = 4$[c]	0.197	0.160

[a] bcc, Body centered cubic.
[b] fcc, Face centered cubic.
[c] $d = 4$, Four-dimensional hypercubic lattice.

It is rather easy to produce by computer a percolation picture, as in Fig. 4. One simply lets the computer go through a large two-dimensional array and, at every array element, decides with the help of a random number whether that site is full or empty. If the random number is distributed evenly between zero and unity, the site is taken as occupied for random numbers smaller than p, and empty otherwise.

It is not so easy to count clusters in a sample like Fig. 4 or to determine whether there is a cluster connecting the top and the bottom of the sample. That aim can be achieved efficiently with the Hoshen–Kopelman algorithm. There is a fortran program for counting clusters in bond percolation on simple cubic lattices. Only one line of a two-dimensional lattice, or one plane in three dimensions, has to be stored at any one time in the computer. Each cluster that is started new when we go through the lattice is given an index, and an index of index. If later two initially separated clusters merge into one cluster, this fact is recorded by changing the index of the involved index. The index itself remains unchanged; we do not have to go back in the lattice to relabel previously investigated sites. Fast computers can handle each site within a few microseconds; as of January 1987 the largest system simulated to our knowledge was a $160{,}000 \times 160{,}000$ square lattice.

One may also investigate which fraction P_∞ of sites belongs to the infinite network in an infinite system or to the largest cluster in a finite system (in the limit of system size $\to \infty$). Obviously for probability p below the threshold p_c this fraction, often called the percolation probability P_∞, is zero. One may instead look at the mean cluster size S, which is found if one selects randomly an occupied lattice site, counts how large the cluster is to which it

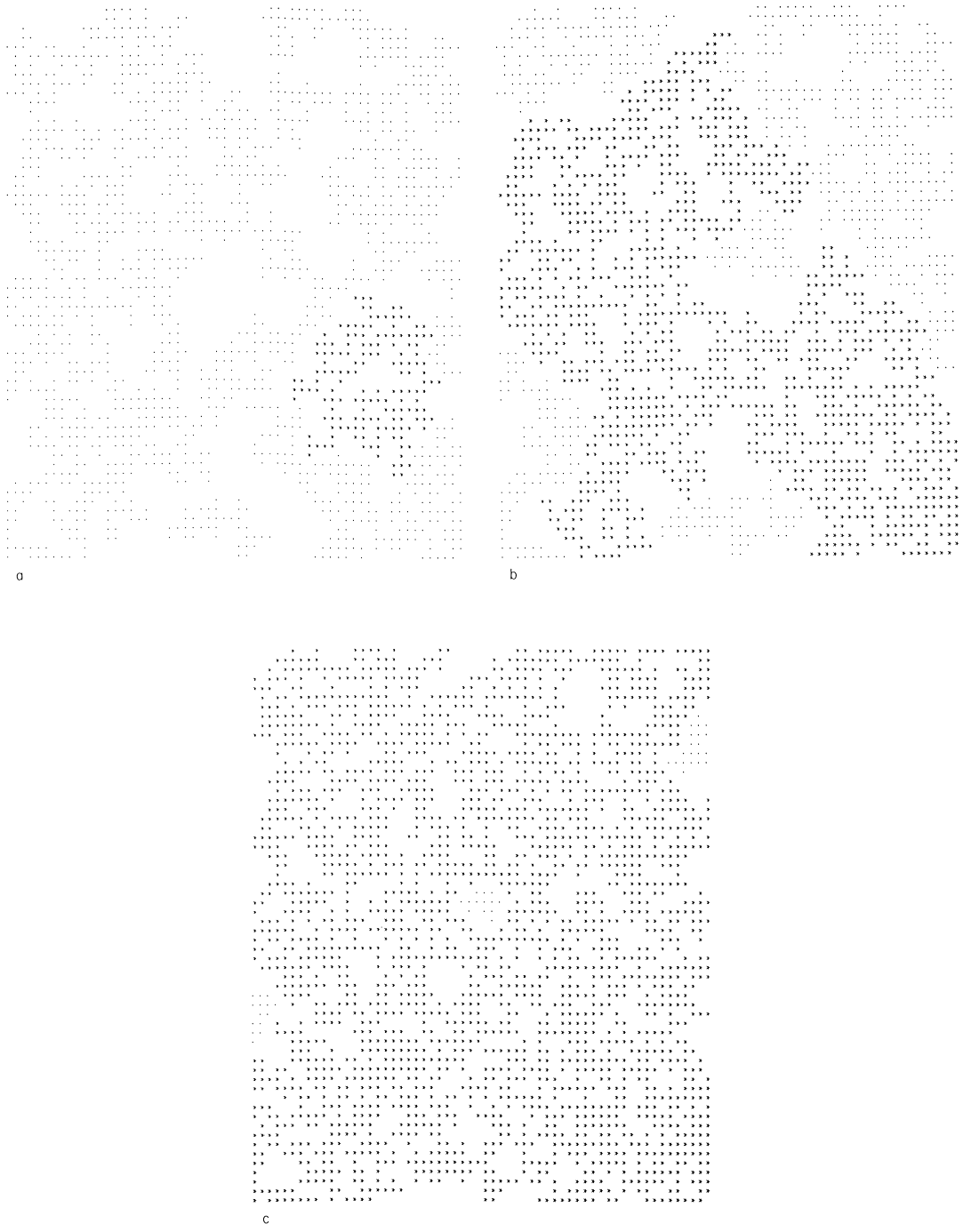


FIGURE 4 Clusters of sites connected by bonds formed in random site percolation on a square lattice (a) below ($p = .543$), (b) at ($p = .593$), and (c) above ($p = .643$) the percolation threshold. The largest cluster is denoted by stars, the smaller clusters containing >10 sites by dots.

belongs, and averages over all lattice sites selected in this way. It is plausible that this mean cluster size S diverges if the percolation threshold p_c is approached from below since, above this threshold, an infinite network is present. Finally, one may define a correlation length ξ as the root mean square average distance between two randomly selected occupied sites within the same cluster. In short, ξ is a typical cluster radius; and this length must also diverge

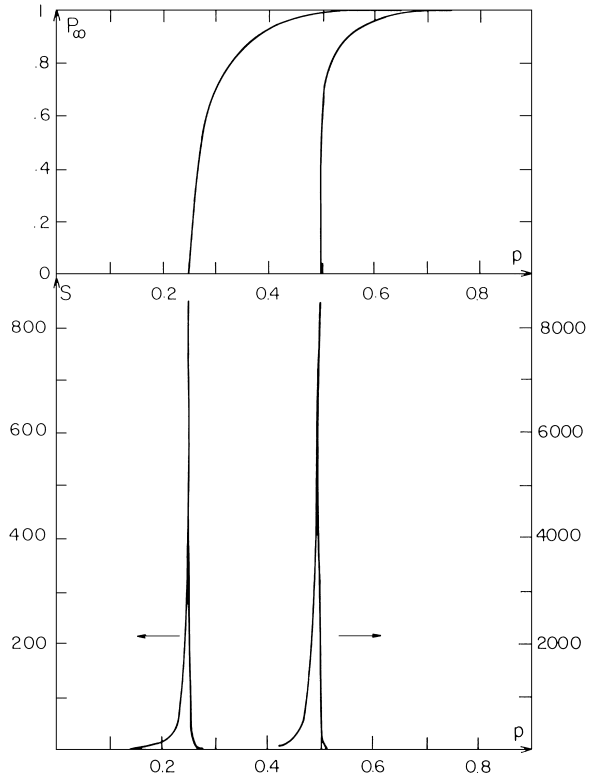


FIGURE 5 Variation of percolation probability P_∞ and mean cluster size S for bond percolation on the square ($p_c = .50$) and simple cubic ($p_c = .249$) lattice.

if $p \to p_c$ from below. For p above p_c one can define a finite mean cluster size S and a finite correlation length ξ by restricting the initially selected sites to be on some finite cluster, not on the infinite network.

How do the quantities P_∞, S, and ξ vanish or diverge at the threshold? Figure 5 shows the variation of P_∞ and S on the square and simple cubic lattices (bond percolation). For p very close to p_c the curves may be fit by power laws:

$$P_\infty \propto (p - p_c)^\beta, \qquad p > p_c \tag{2a}$$

$$S \propto (p_c - p)^{-\gamma}, \qquad p < p_c \tag{2b}$$

$$\xi \propto (p_c - p)^{-\nu}, \qquad p < p_c \tag{2c}$$

These proportionalities are supposed to be valid only asymptotically for $p \to p_c$, just as $\sin(x) = x$ is valid only for $x \to 0$. With the above restriction to finite clusters, Eqs. (2b) and (2c) also apply to the case $p > p_c$, with $p_c - p$ replaced by $p - p_c$ and with a different proportionality constant. The quantities β, γ, and ν and similar exponents governing other power laws are called critical exponents. A large part of modern percolation research has centered on these critical exponents. Similar exponents have been thoroughly investigated for critical phenomena near many other phase transitions. These critical phenomena refer to the asymptotic region $p \to p_c$, length $\to \infty$, time $\to \infty$, ..., on which we also concentrate for percolation.

TABLE II Values of Percolation Exponents in Two and Three Dimensions and for High Dimensions[a]

Exponent	$d = 2$	$d = 3$	$d \to \infty$
β	$\frac{5}{36}$	0.42	1
γ	$\frac{43}{18}$	1.80	1
ν	$\frac{4}{3}$	0.88	$\frac{1}{2}$
σ	$\frac{36}{91}$	0.45	$\frac{1}{2}$
τ	$\frac{187}{91}$	2.19	$\frac{5}{2}$
$D(p = p_c)$	$\frac{91}{48}$	2.52	4
$D(p < p_c)$	1.56	2	4
$D(p > p_c)$	2	3	4
μ	1.3	2.0	3

[a] Rational numbers presumably are exact; the others are numerical estimates [μ is defined in Eq. (15)].

Values of the exponents β, γ, and ν are listed in Table II. Unlike Table I, this table contains only one entry for all the two-dimensional (and one for all the three-dimensional) lattices, all of which have the same exponents. This universality is one of the main reasons that critical exponents for percolation theory have attracted so much attention. Our understanding of universality is based in part on self-similarity, as explained in Section III.

III. FRACTALS, SCALING, AND RENORMALIZATION

One of the powerful methods of confirming Eq. (2) uses extrapolation of data obtained in numerical simulations. These are done on finite lattices of size L^d ($L \times L$ in two dimensions, $L \times L \times L$ for three dimensions). Above p_c, if the correlation length ξ is small compared with L, the largest cluster fills the sample uniformly, with a density equal to P_∞. Thus, the number of sites in the sample that belong to this cluster is

$$M(L) = L^d P_\infty \propto L^d (p - p_c)^\beta \propto L^d \xi^{-\beta/\nu} \tag{3}$$

This situation is exhibited in Fig. 4c. As p decreases toward p_c, the correlation length ξ gradually increases. Since ξ represents a typical size of a finite cluster, it also represents a typical size of the "empty" areas in the infinite cluster. As can be seen from Fig. 4c, at $p = p_c + 0.05$ there appear empty areas of *all* sizes up to ~7 lattice units, which is a rough estimate of ξ. The appearance of all these sizes reflects the property of self-similarity at distances smaller than ξ.

If one continues to decrease p all the way to p_c (Fig. 4b), ξ diverges to infinity and the picture in Fig. 4b contains empty islands (i.e., lattice sites not belonging to the largest cluster) at *all* length scales. A direct counting of the number M of sites on the largest cluster within the sample now yields

$$M(L) \propto L^D \tag{4}$$

with D having the noninteger values listed in Table II. Since D replaces d in the power of L, as compared with Eq. (3), it is natural to consider D as a generalized "fractal" dimensionality.

A power law dependence like Eq. (4) characterizes fractal graphs, an example of which is shown in Fig. 6. In this example, called the Sierpinski carpet, the number of filled squares changes by a factor of 8 when the length scale changes by a factor of 3, so that $8 = 3^D$, or $D = \log 8/\log 3 = 1.893$. Indeed, fractals like the Sierpinski carpet have been widely used to investigate some of the properties of the infinite cluster at p_c.

Looking at Fig. 4c, one may divide it into unit squares of sizes $\xi \times \xi$. Within each square the cluster is self-similar, obeying Eq. (4), whereas the squares form a uniform distribution. Since the number of squares is $(L/\xi)^d$, and their "mass" (number of occupied sites on the largest cluster) scales as ξ^D, we find that the total mass scales as $\xi^D(L/\xi)^d \propto L^d \xi^{D-d}$. Comparison with Eq. (3) now yields

$$D = d - \beta/\nu \tag{5}$$

The power laws in Eq. (2) are direct consequences of the self-similarity.

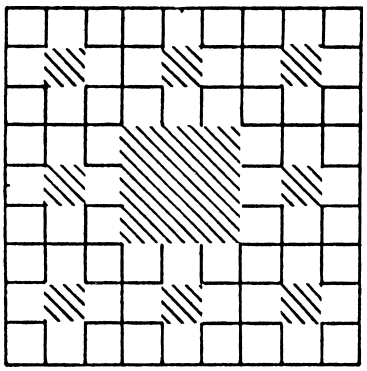


FIGURE 6 Example of a Sierpinski carpet as a model for large clusters at the percolation threshold. Empty squares are shadowed; all other squares are occupied and will be divided in the next iteration into nine squares, of which the center square will be empty. The length L of this structure is the number of elementary squares on each side; it increases by a factor of 3 at each iteration. The "mass" M of the cluster is the number of occupied elementary squares in it; it increases by a factor of 8 at each iteration. Thus, $M = L^D$, and the fractal dimension $D = \log 8/\log 3 = 1.893$ is close to that of the largest percolation cluster at the threshold, $D = \frac{91}{48} = 1.896$.

More generally, Eqs. (3) and (4) can be combined into a scaling form (with + for $p > p_c$ and − for $p < p_c$):

$$M(L, \xi) = L^D m_{\pm}(L/\xi) = L^D \tilde{m}\big((p - p_c)L^{1/\nu}\big) \tag{6}$$

We see that ξ is the only relevant length scale in the problem, so that the dependence of the ratio M/L^D on L should be only through the ratio L/ξ (i.e., L is measured in units of ξ). The scaling function $m_+(z)$ has the limiting behavior $m_+ \to$ const for $z \ll 1$ and $m_+(z) \propto z^{\beta/\nu}$ for $z \gg 1$. Equation (6) can be interpreted as describing a crossover from the fractal dimension D at $L \ll \xi$ to the Euclidean dimension d at $L \gg \xi$. For $L \gg \xi$ one may thus say that $D(p > p_c) = d$, at least for $d < 6$ (Table II).

Measurements of $M(L, \xi, p)$ yield the exponents D, β, and ν and indeed confirm the scaling form [Eq. (6)]. A direct indication of universality arises from the fact that Eq. (5) has been shown to apply not only to simple model computer experiments, but also to the electron microscope picture of real sputtered films (Fig. 7). Note that our theories refer to percolation on lattices, whereas these experiments are done on a continuum. However, computer simulations of continuum percolation seem to give the same exponents as lattice percolation.

Scaling forms like Eq. (6) can be written for other quantities, for example, for the mean cluster size,

$$S = L^{\gamma/\nu} s_{\pm}(L/\xi) = L^{\gamma/\nu} \tilde{s}\big((p - p_c)L^{1/\nu}\big) \tag{7}$$

They are called finite size scaling and are crucial in the analysis of data from small samples. In many computer simulations, exponent ratios like γ/ν can be estimated more accurately through Eq. (7) and its analogs than the exponents γ and ν separately through Eq. (2).

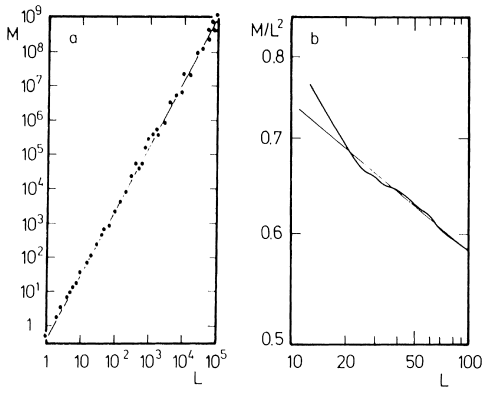


FIGURE 7 (a) Log–log plot of the number of sites, $M(L)$, in the largest cluster of an $L \times L$ triangular lattice at its site percolation threshold. The slope of the straight line is the fractal dimension $D = \frac{91}{48}$. (b) Log–log plot of the fraction of sites, $M(L)/L^d$, measured experimentally on sputtered metal films of size $L \times L$ at their two-dimensional percolation threshold. The slope of the straight line is the prediction $D - d = -\beta/\nu = -\frac{5}{48}$ of percolation theory. The experimental data are summarized by the wavy line.

As can be seen qualitatively from Fig. 4, finite clusters also exhibit self-similar features. Indeed, similar to Eq. (6), the average linear size R_s of a finite cluster with s sites is expected to behave as

$$R_s = s^{1/D} r_{\pm}(s/\xi^D) = s^{1/D} \tilde{r}\big((p - p_c)s^{\sigma}\big) \tag{8}$$

with

$$\sigma = 1/(D\nu) \tag{9}$$

Assumption (9) shows that the finite clusters have the same fractal dimension as the largest cluster as long as $R_s \ll \xi$, or $s \ll \xi^D$. The probability of finding much larger finite clusters of size $R_s \gg \xi$ is exponentially small. However, the function $r_{\pm}$ or $\tilde{r}$ yields a different behavior in that limit. For $p > p_c$ these rare large clusters are very compact, and $R_s \propto s^{1/d}$, that is, $D(p > p_c) = d$. In contrast, the rare large clusters found for $p < p_c$ are very loosely structured. These clusters turn out to have a lower fractal dimension $D(p < p_c)$, called that of lattice animals and listed in Table II. These lattice animals (percolation clusters for $p \to 0$) may be a model for branched polymers in a very dilute solution, where interactions between different macromolecules are negligible. In contrast, percolation clusters near p_c may be a model for gelation, the formation of an infinite network out of many intermediate-size macromolecules.

Scaling relations like Eq. (6) or (8) were introduced into percolation theory as phenomenological assumptions, in analogy to critical phenomena. Renormalization group arguments give further theoretical support for their validity. In a typical example of the renormalization group analysis one transforms a given dilute lattice into a coarse-grained version of itself, in which the length scale is enlarged by a factor $b (b = \sqrt{3}$ in Fig. 8). Basic cells, of b^d sites (three sites in the example), are replaced by single new sites, which are considered to be occupied if two opposite edges are connected by a cluster of previous sites, and empty otherwise. The result is a renormalized concentration p' of new sites; p' is a function of the concentration p of previous sites, $p' = R(p)$. In the simple example of Fig. 8,

$$p' = p^3 + 3p^2(1 - p)$$

since a triangle is defined as occupied if all three, or at least two, of its sites are occupied, thus connecting at least two of the three corners.

One may coarse-grain the new sites again, that is, replace b^d of the new sites by one supersite with occupation probability $R(p')$, and so on. The relation $p' = R(p)$ applies to each of these "renormalizations." After many iterations, the renormalized concentration approaches unity if one started with p near 1, and approaches zero if initially p was near 0. There exists one invariant fixed point $R(p^*) = p^* (0 < p^* < 1)$ of the transformation; in our example this fixed point is at $p^* = p_c = \frac{1}{2}$. If one starts very close to p^*, one can linearize $R(p) = p^* + (p - p^*)\Lambda + \cdots$ and set $\Lambda = b^y$:

$$p' - p^* = (p - p^*)b^y \tag{10}$$

Since the renormalization changed the length scale by a factor b, one has $\xi' = \xi/b$. Using $\xi \propto (p - p^*)^{-\nu}$ [Eq. (2c)], one identifies $y = 1/\nu$, at least for large b.

In our example, $b = \sqrt{3}$, $b^y = \frac{3}{2}$ from expanding $R(p)$ about $p^* = \frac{1}{2}$, and thus $y = 1/\nu = \log \frac{3}{2}/\log \sqrt{3} = 0.738$, close to the correct result $y = \frac{3}{4}$ (Table II). This renormalization method for finite b is only approximate, because all collective phenomena are simulated by a small cell with only three sites. The approximation is better for larger values of b; large cells can be studied by computer simulation, whereby one checks with what probability $R(p)$ opposite edges are connected.

If one had a different (e.g., square) lattice, details of the renormalization group transformation $p \to p' = R(p)$ would change, and $p_c = R(p_c)$ is in general not valid. However, the properties near the fixed point, like the values of the critical exponents, reflect the behavior at very large length scales. On these scales, local short-range differences become irrelevant, and universality results.

Equations like (10) imply power law behavior at $p = p_c$ for the quantities of interest and yield scaling forms like Eq. (6) or (8). Another quantity of interest is the average number n_s of clusters containing s connected sites each. (We normalize these cluster numbers by dividing them by the total number of lattice sites.) This quantity is related to the finite clusters only: Since there seems to be at most a small number of infinite networks present in percolation for $d < 6$, this number divided by the lattice size goes to zero for infinite lattices. As in the above examples [Eqs. (6) and (8)], the cluster numbers n_s obey the scaling form

$$n_s = s^{-\tau} N_{\pm}(s/\xi^D) = s^{-\tau} \tilde{N}\big((p - p_c)s^{\sigma}\big) \tag{11}$$

with the new critical exponent τ. One can now relate the mean cluster size S and the percolation probability P_∞ to the cluster numbers via

$$S \propto \sum_s s^2 n_s \tag{12a}$$

$$P_\infty = p - \sum_s s n_s \tag{12b}$$

at least close to the threshold. Here the sums run over all finite cluster sizes, with $s = 1$ corresponding to isolated sites. [In Eq. (12b), p has to be replaced by unity for bond percolation.] These equations lead to the scaling relations

$$\tau = 2 + \beta/(\beta + \gamma); \qquad \sigma = 1/(\beta + \gamma) \tag{13}$$

Generally (also for other phase transitions), one describes relations between critical exponents as scaling laws. They

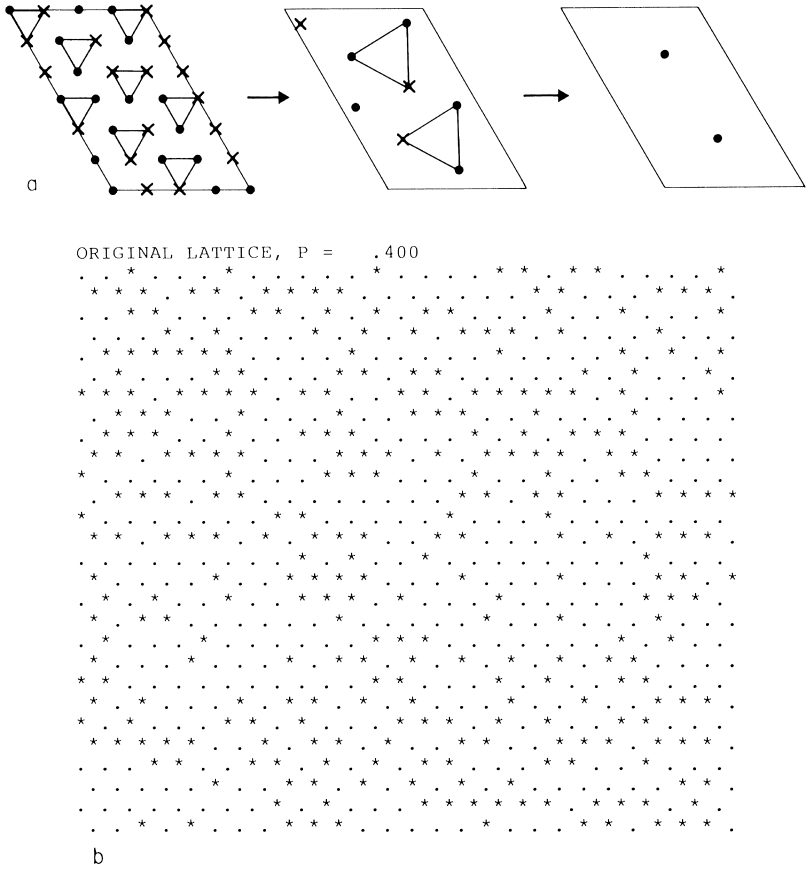


FIGURE 8 (a) Renormalization of small cells on the triangular lattice. Every lattice site (dot for occupied, cross for empty sites) belongs to one triangle; each triangle is renormalized (coarse-grained, averaged) into a new site. The new sites form again a triangular lattice, with a lattice constant $b = \sqrt{3}$ times larger than the old distance between nearest neighbors. A new site is occupied if in the small cell that it represents two opposite edges are connected. (b)–(p) Parts of a large computer-generated sample (stars for occupied and dots for empty sites). The large lattices are renormalized at an initial concentration of .4 (which is approaching zero as indicated through subsequent renormalizations), of .5 (which stays at this fixed point $\frac{1}{2}$ throughout all further iterations), and finally of .6 (which is approaching unity as indicated through subsequent renormalizations). Only for the threshold $p = .5$ is there self-similarity in the sense that the renormalized lattice has on the average the same properties as the unrenormalized lattice, apart from a change in lengths. This self-similarity is the basis of fractal properties and of scaling laws.

are derived from similarity assumptions like Eqs. (6), (7), (8), and (11). Equation (11) states that the cluster size distributions near sizes $s = s_1$ and $s = s_2$ are the same apart from a factor, provided that we scale the distance $p - p_c$ proportional to $s^{-\sigma}$ in order to keep the argument of the scaling function f in Eq. (11) the same.

The number n_s of small clusters can be evaluated exactly. For example, on a square lattice each pair of neighboring occupied sites has six empty neighbor sites (also called the perimeter) and can be oriented either horizontally or vertically. Thus, $n_s = 2p^2(1-p)^6$ for $s = 2$. For $s = 3$ we may have either a horizontal or a vertical straight line of three occupied sites [contribution $2p^3(1-p)^8$] or one of four corner configurations [contribution $4p^3(1-p)^7$]. Thus,

$$n_3 / p^3 = 2(1-p)^8 + 4(1-p)^7$$

Similar, but more complicated formulas have been derived for s up to 10 or 20. They are used to express quantities like the mean cluster size S as a power series in p or $1-p$, an important technique for analyzing critical behavior. We see that for $p \to 1$ only the most compact configurations with the smallest perimeter survive. Thus, in this limit (and actually for all very large clusters above p_c) the clusters have a rather compact structure, with the radius R_s increasing as little as possible with increasing mass s, that is, $R_s \propto s^{1/d}$, as mentioned in relation to Eq. (9). In the opposite limit, $p \to 0$, for fixed size s, all cluster configurations occur equally often and one ends up with the lattice animal limit.

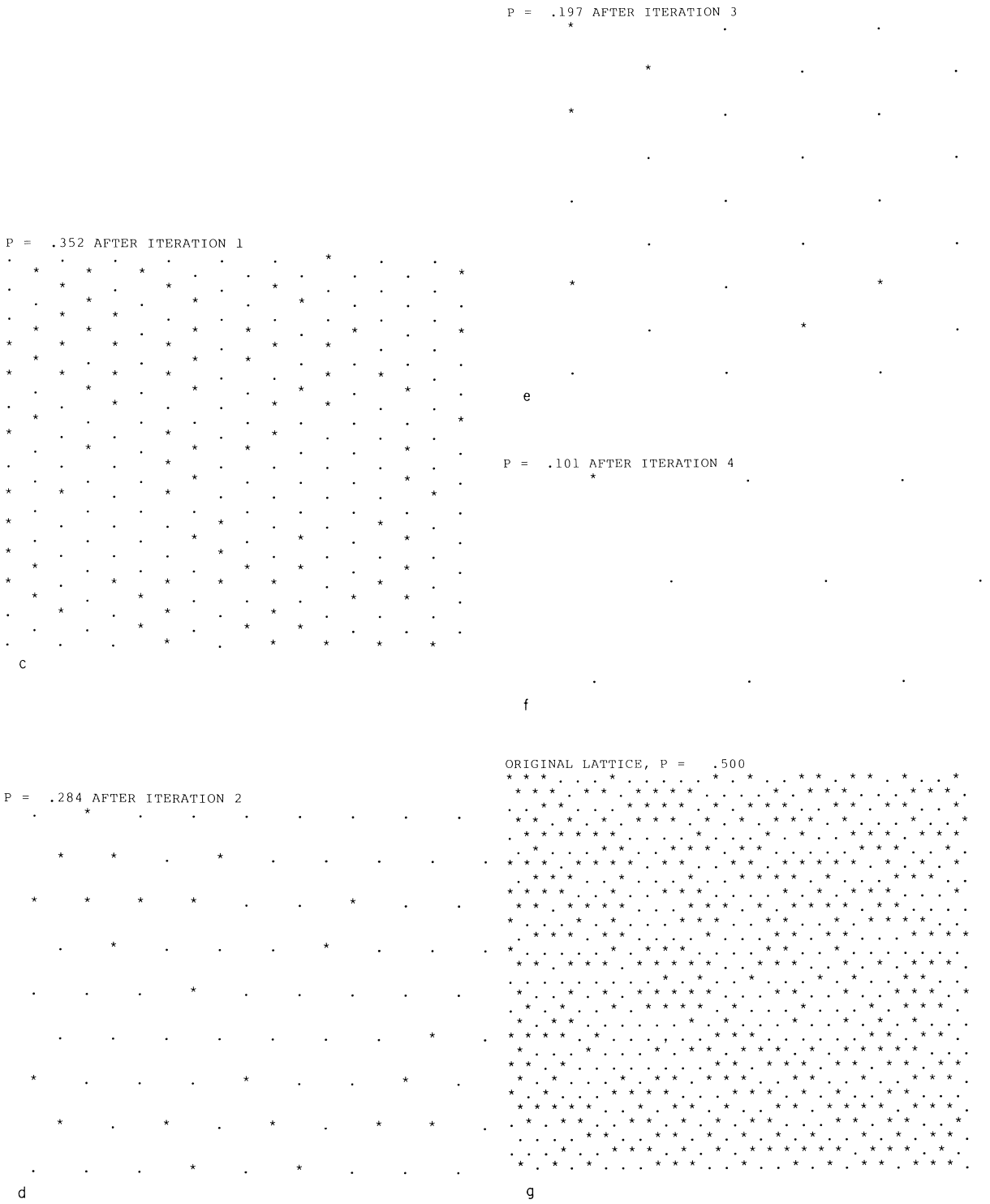


FIGURE 8 (*Continued*)

All these scaling forms [Eqs. (6), (8), and (11)] are valid only asymptotically for large clusters at or very close to the percolation threshold. Only here are the various critical exponents defined. Through the scaling laws [Eqs. (5), (9), and (13)] relating these exponents to one another, we can predict all these exponents once we know two of them. Again, this property is shared by many other phase transitions and their static critical phenomena.

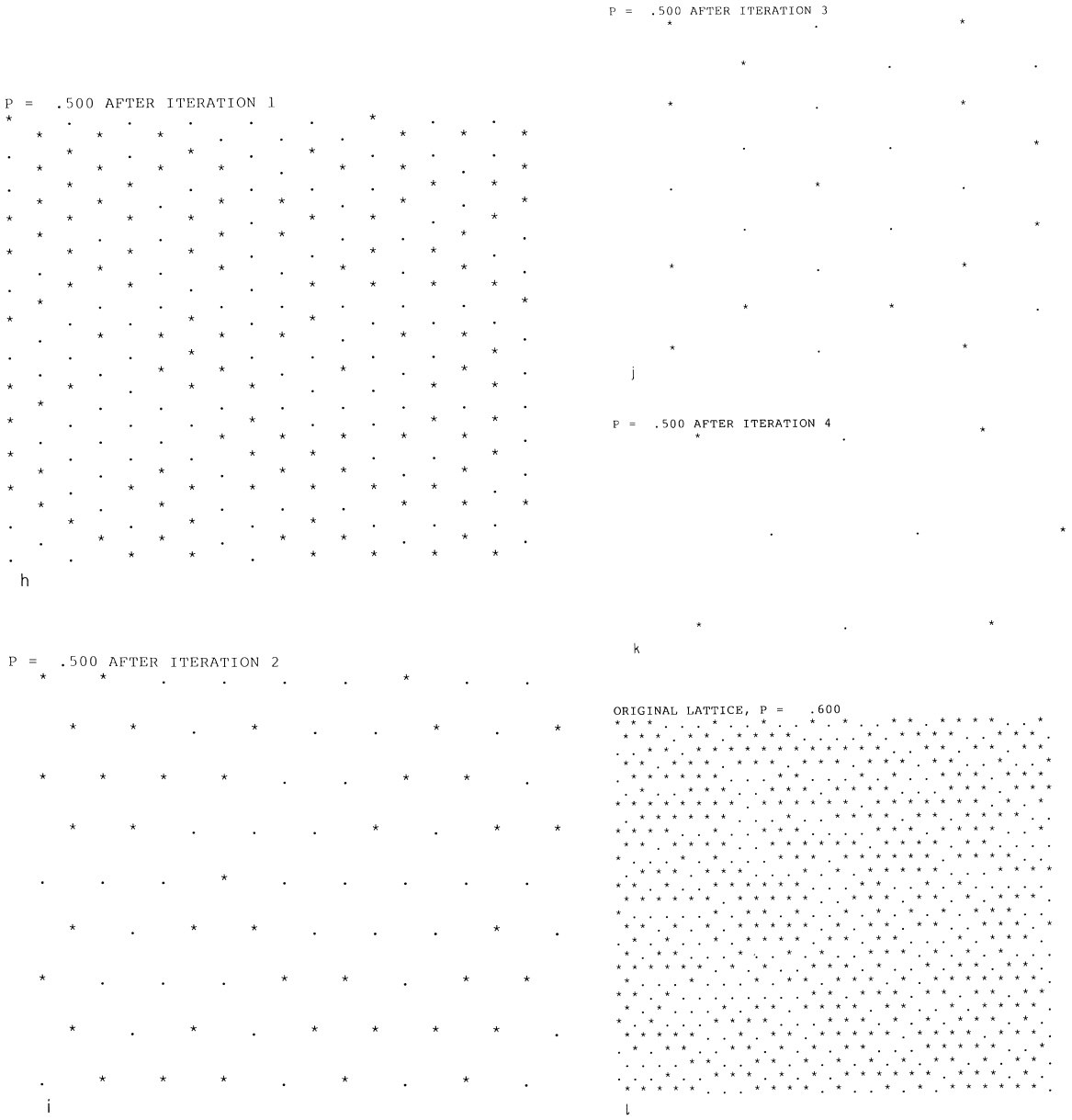


FIGURE 8 (*Continued*)

These scaling theories of percolation seem quite well confirmed by computer or laboratory experiments. For example, at the percolation threshold the cluster numbers n_s follow the simple power law $s^{-\tau}$ predicted by Eq. (11), as Fig. 9 indicates. In two dimensions there are good arguments, though no mathematically rigorous proof, that the critical exponents of percolation are rational numbers like $\nu = \frac{4}{3}$. Similar rational exponents also arise in many models for magnets and fluids where the two-dimensional exponents are known exactly. In three dimensions, no exact exponent is known, either for percolation or for magnets and fluids.

IV. KINETIC ASPECTS OF PERCOLATION

In this section we deal with percolation processes rather than percolation structures. Thus, we consider time-dependent phenomena connected with percolation.

A certain numerical method of building a percolation cluster by computer simulation can be interpreted as a time-dependent process. One starts with a given isolated site. Then one looks at the shell of neighbors surrounding that site and determines for each neighbor whether that site is occupied or empty; this process completes the first time step. In the second time step, one looks at the neighbors

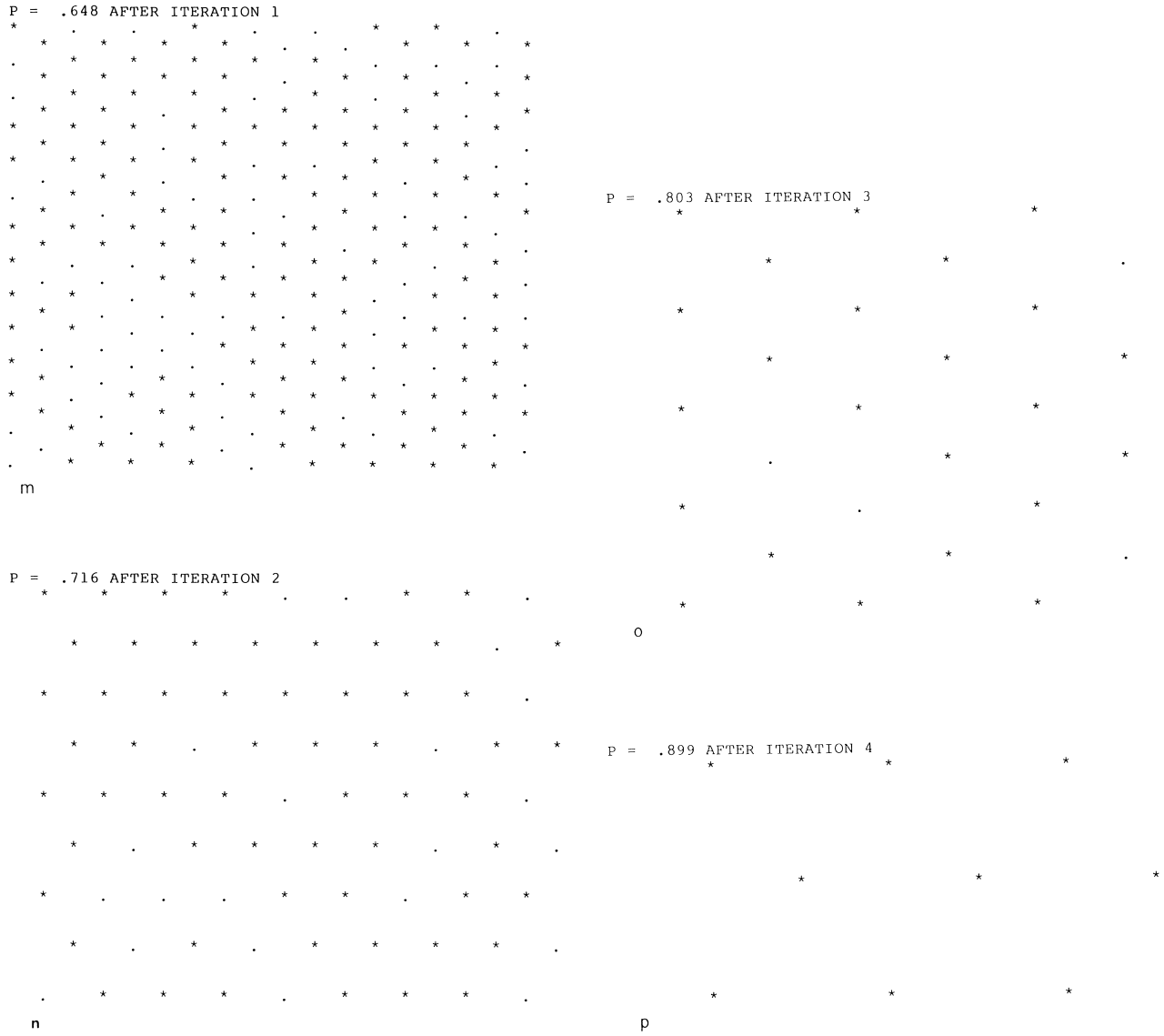


FIGURE 8 (*Continued*)

of the occupied sites that were not visited before, and so on. Once a site is identified as empty or occupied, it will always have this status during the growth process. In this way one cluster of s sites is created with a probability proportional to $n_s s$. In addition, one can count how many time steps were needed to build this cluster. For example, to build a large two-dimensional cluster of s sites, one needs a time

$$t_s \propto s^{0.6} \propto R_s^{1.1} \tag{14}$$

at the percolation threshold. In this time-dependent interpretation, percolation can be used as a model for the spread of epidemics or forest fires. At any given time, however, the cluster structure is that of random percolation, as described in the previous section. We restrict ourselves in this section to those kinetic percolation models that lead to the same static behavior as random percolation. The time t can also be interpreted as the "chemical" or "topological" distance in the sense of the shortest connection of a point with the origin, with all steps in between taken within the cluster.

A much more extensively studied transport property of percolation is the conductivity of random resistor networks. Imagine every occupied site (or bond) in the site (or bond) percolation problem to be a conductor, and every empty site or bond to be an insulator. Electric currents can flow only between neighboring conductors. The purpose of such a network is to model the electric properties of real

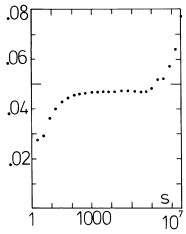


FIGURE 9 Semilogarithmic plot, versus s, of the number of clusters containing at least s sites, multiplied by $s^{\tau-1} = s^{96/91}$, at the site percolation threshold $p = .5927$ of a $160{,}000 \times 160{,}000$ square lattice with periodic boundary conditions. Each site took on average ~ 3 μsec on an IBM 3081 computer. According to the scaling assumption [Eq. (11)] $n_s \propto s^{-\tau}$, the quantity plotted here should approach a plateau for large s. The deviations for the largest cluster sizes come from the boundaries or inaccuracies in p_c (or both).

alloys of metals with insulators, the flow of fluids through porous rocks, and so on. This random resistor network has a conductivity Σ, which is zero for $p < p_c$ and nonzero for $p > p_c$, since only in the presence of an infinite network of neighboring conductors can electric charges flow from one end of the sample to the other. For p approaching the threshold p_c from above we define a new critical exponent μ through

$$\Sigma \propto (p - p_c)^{\mu} \tag{15}$$

Since at present the relation between μ and the static critical exponents like β and γ is unclear, we listed this exponent μ in Table II as if it were independent of the others. At the threshold $p = p_c$, in a finite system of linear dimension L the conductivity varies as $L^{-\mu/\nu}$, entirely analogous to Eq. (6) for the percolation probability: $P_\infty = M(L)/L^d \propto L^{-\beta/\nu}$.

One can also replace the conductors in the above random resistor network by superconductors (zero resistivity) and the insulators by resistors. Then the conductivity is infinite for $p > p_c$ due to the infinite superconducting cluster. The conductivity therefore diverges if p approaches p_c from below. The exponent for this divergence, often called s in the literature, equals μ in two dimensions and is ~ 0.75 in three.

The dynamic aspect of the conductivity becomes clearer if one looks at diffusion processes. Let us put some particle, or "walker," on an occupied site in the site percolation problem. At every time step, the walker selects randomly one of its nearest neighbors. If that neighbor is occupied, then the walker moves there; otherwise, it stays where it is. Then, in the next time step, the walker again selects a random neighbor of its current site and moves there if it is occupied. Repeated again and again, this process is like an ant in labyrinth. Many different "ant" investigations have been undertaken since then.

If all sites are occupied ($p = 1$), the above process is the usual diffusion process of a random walker on a regular lattice; the averaged squared distance traveled in time t is $R^2 = t$. This rms distance R (as the crow flies) must be distinguished from the cluster radius R_s or the renormalization function $R(p)$ of the preceding section and is also, of course, different from the chemical distance, the shortest path connecting the present site with the origin of the walk. The relation between the time t (number of steps) and the averaged distance R can be generally written as

$$t \propto R^{d_w} \tag{16}$$

with d_w representing the fractal dimensionality of the random walk. For usual diffusion (at $p = 1$) one has simply $d_w = 2$. In other words, we identify here the "mass" M of Eq. (4) with time t, a similarity that is particularly plausible if the walk is taken as a model for polymers.

If nearly all sites are empty ($p \to 0$), most walkers will start on an isolated occupied site where they cannot move at all: $R = 0$. For all $p < p_c$, only finite clusters exist, and R thus cannot increase to infinity for time $t \to \infty$. More precisely, the ant species defined above visits all sites of a finite cluster equally often, if given enough time. Thus, ants starting on clusters with s sites will walk a distance R of the order of the cluster radius R_s. An ant starts on such an s cluster with probability $n_s s$ since for the origin of a walk a large cluster is selected more often than a small cluster. Thus, R^2 approaches for long times below p_c a limit proportional to $\Sigma_s R_s^2 n_s s$. The scaling assumptions of the previous section predict this sum to diverge for $p \to p_c$:

$$R^2 \propto (p_c - p)^{-m}, \qquad m = 2\nu - \beta \tag{17}$$

This static limit was confirmed by accurate computer simulations after initial difficulties were overcome.

For p above p_c but below unity, the many holes in the infinite network slow down the diffusion of the walking ant: $R^2 < t$. Moreover, if the walk starts on a finite cluster, which it may still do even above p_c, the distance will again be limited by the cluster radius. Thus, the ant still diffuses, $R^2 \propto t$, but the diffusivity R^2/t is diminished compared with the pure case, $p = 1$. If p approaches p_c from above, this diffusivity goes to zero as $(p - p_c)^{\mu}$, since Einstein predicted diffusivities to vary as the conductivity [Eq. (15)]. (The current density is proportional to the velocity and the density of charge carriers. The velocity is the product of mobility and electric field. The ratio of current density to electric field is the conductivity. According to Einstein the mobility varies as the diffusivity.) Thus, we have for very long times near the threshold

$$R^2 \propto (p - p_c)^{\mu} t \tag{18a}$$

If we restrict the starting point to be on the infinite network only, and no longer on any occupied site, then the finite clusters no longer enter the average and the distance increases by a factor $1/P_\infty \propto (p - p_c)^{-\beta}$:

$$R^2 \propto (p - p_c)^{\mu-\beta} t \tag{18b}$$

Both types of averages, diffusion only on the infinite network or anywhere on the lattice, have been investigated by computer simulations and were found to agree with the above predictions.

Similar to the scaling assumptions [Eqs. (3), (6), and (8)] we may postulate

$$R^2 = t^{2k} f_\pm(t/t_\xi) = t^{2k} \tilde{f}\big((p - p_c)t^x\big) \tag{19a}$$

with a characteristic time

$$t_\xi \propto (p - p_c)^{-1/x} \tag{19b}$$

The scaling function $\tilde{f}(y)$ varies asymptotically in such a way that Eq. (18a) is recovered for large positive argument y and Eq. (17) for large negative y. This requirement leads to

$$2k = mx; \quad x = 1/(m + \mu); \quad m = 2\nu - \beta \tag{20a}$$

with the asymptotic behavior $f \propto y^\mu$ and $\propto(-y)^{-m}$ for large positive and negative arguments y, respectively, provided that the random walk starts on any occupied site. If instead it starts on the infinite cluster only, we have to match Eq. (18b) instead of Eq. (18a) in the diffusion regime. Denoting the corresponding exponents and scaling functions by primes, one finds

$$2k' = 2\nu x \tag{20b}$$

with the same $x' = x$ but a different scaling function $\tilde{f}'(y)$ varying as $y^{\mu-\beta}$ for $y \to \infty$.

Right at the percolation threshold $p = p_c$, the scaling functions $\tilde{f}$ and $\tilde{f}'$ approach some constants for zero argument, and one has

$$R \propto t^k \propto t^{(\nu-\beta/2)/(2\nu+\mu-\beta)} \tag{21a}$$

for walks anywhere at the percolation threshold, and

$$R \propto t^{k'} \propto t^{\nu/(2\nu+\mu-\beta)} \tag{21b}$$

for walks on the largest cluster at $p = p_c$. Thus, the squared distance R^2 does not increase linearly with t nor does it approach a constant limit for large times. Instead, it increases with some smaller power of the time, an effect also called anomalous diffusion.

From Eqs. (19b) and (20b) we see that the characteristic time varies as $t_\xi \propto \xi^{1/k'}$; it is the time to travel a distance ξ on the infinite cluster [Eq. (21b)]. The exponent $1/k'$, which is equal to the anomalous fractal dimension of the random walk on the largest cluster, is analogous to the dynamic critical exponent z in second-order phase transitions. Combined with Eq. (4), the volume spanned by the sites visited within time t is of the order $R^D \propto t^{Dk'}$. In comparison with the same quantity for random walks on the full ($p = 1$) lattice, $t^{d/2}$, the combination $d_s = 2Dk'$ has been called the fracton or spectral dimension. Numerically, d_s seems to be very close (although not exactly equal) to $\frac{4}{3}$ for percolation clusters in $d > 1$ dimensions. Were this Alexander–Orbach relation, $d_s = \frac{4}{3}$, exact, it would imply an explicit dependence of μ on the static exponents β and ν. However, because this seems to be only an approximation and no alternative relations are at present generally accepted, the critical exponent μ (and therefore k, k', or d_s) is regarded here as an independent new exponent. Possible relations between the dynamic and static exponents were the subject of much research.

Computer simulations have confirmed the above scaling theory within their numerical errors. For too short times, the exponents "measured" for k or k' are too high; a million time steps on lattices containing millions of sites might be necessary to get the three-dimensional exponent $k = 0.20$ or $k' = 0.27$ with good accuracy (Fig. 10). In two dimensions these systematic deviations for short times are smaller, and $k = 0.33$ or $k' = 0.35$ can be determined more accurately. The measurement of the diffusivity slightly above p_c [Eq. (18a)] is particularly difficult, since Eq. (19) tells us that the normal diffusion behavior can be observed only for times much longer than the characteristic time $\propto (p - p_c)^{\beta-\mu-2\nu}$. For times much shorter than this characteristic time, one observes the same anomalous diffusion as for $p = p_c$, even if p is slightly different from p_c.

As in biology, one can study more than one ant species. It does not matter much if the diffusing ant, having tried to move in a prohibited direction, immediately tries another one before time goes by. (In the earlier definition time goes by even after an unsuccessful attempt.) For this species the diffusion process is faster, but the critical exponents seem to remain the same (same dynamic universality class). Exponent oscillations periodic in the logarithm of time occur

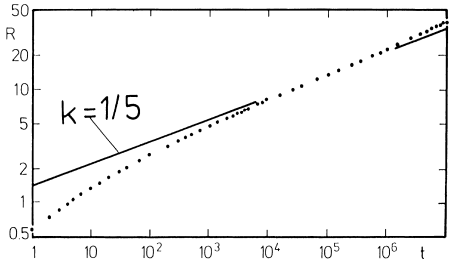


FIGURE 10 Distance R versus time t for "ant" diffusion at the percolation threshold of a 256^3 simple cubic lattice. Each step took $\sim$0.2 μsec per "ant" on a Cyber 205 vector computer. The scaling assumption [Eq. (19a)] predicts for large times $R \propto t^{0.20}$, as symbolized by the straight line in this log–log plot. The deviations for very large t are due to the finite lattice size.

if a bias is introduced, due to which one direction is chosen more often than the others. This bias corresponds to an electric field, if we identify the random walker with an electron in a disordered material; however, in that case interactions between electrons may be important and must be studied by computer simulations. A magnetic field incorporated into the "ant" model of electron diffusion does not seem to change the universality class. "Butterflies" can jump to more distant sites and not only to nearest neighbors. Finally, "termites" are diffusing random walkers on a superconductor-resistor network, and "parasites" are ants diffusing on lattice animals, as defined earlier.

V. MODIFICATIONS AND GENERALIZATIONS

The butterfly example suggests that in percolation one need not look only at clusters defined via nearest-neighbor connections. In the square site problem, when one allows sites to be connected to a cluster if they are nearest or next nearest neighbors, then the percolation threshold is lowered to $1 - p_c(\text{nn})$, where $p_c(\text{nn}) = 0.593$ is the nearest-neighbor threshold discussed so far. For other lattices no such simple relation is known. Of particular interest is the limit of long-range connections allowed between cluster sites, when $p_c \to 0$.

In all previous examples the sites (or bonds) were occupied randomly. There may also be correlations between the sites. For example, neighboring occupied sites might be bound together with a certain energy $-J$, which favors such a configuration with a probability $\propto \exp(-J/k_B T)$, where k_B is Boltzmann's constant and T the absolute temperature. Then we have the clustering problem of the lattice gas or Ising model. It seems that its critical exponents are the same as those for random percolation except when phase separation occurs due to the site–site interaction J.

For normal bond percolation, a bond once formed connects its two ends. One may instead define a bond as an arrow in which one end is connected with the other, like a one-way street, but not backward. If the preferred directions are the same throughout the lattice we get the different behavior of "directed percolation."

Among the kinetic generalizations of percolation we mention a polymerization model in which diffusion of "initiators" forms macromolecules: Chemical bonds are formed along the path of these catalysts, but not more than f bonds can emanate from one site.

More drastically different from usual percolation are diffusion-limited aggregates in which single particles diffuse from far away to a cluster, stick to it, and thus let the cluster grow. In cluster–cluster aggregation models, whole clusters diffuse toward one another and thus form larger clusters. Some aspects of this approach are well described by the Smoluchowski equation, formulated at the beginning of this century, in which the coagulation probability for two different cluster sizes s and s' is simply the product $n_s n_{s'}$, multiplied by a coefficient $K(s, s')$; by suitable assumptions for $K(s, s')$ one can reproduce at least one of the percolation exponents.

Fast viscous flow is represented by this diffusion-limited aggregation model. However, when water is injected very slowly into a porous medium filled with oil, the capillary forces dominate the viscous forces, and the dynamics is determined by the local pore radius r. Capillary forces are strongest at the narrowest pore necks. It is consistent with experimental observations to represent the displacements as a series of discrete jumps in which at each time step the water displaces oil from the smallest available pore.

In a lattice model, one assigns random numbers r in the range [0, 1], representing the pore sizes, to the sites. Growth sites are identified as the sites that belong to the "defending" fluid and are neighbors to the "invading" fluid. At every time step the invading fluid is advanced to the growth site that has the lowest random number r.

The invasion into new sites ends when the "invader" reaches the other end of the sample. For easily compressible fluids, the invaded region at that stage has the same fractal structure as the usual infinite percolation cluster. For incompressible fluids in two dimensions the amount of the trapped "residual oil" is larger, such that the fractal dimension of the invaded region is diminished to about 1.82.

VI. OTHER DEVELOPMENTS

During the second half of the 1980s it became clear that random *elastic* networks and random resistor networks may behave quite differently. The electrical conductivity of random networks of resistors and insulators vanishes at the percolation threshold with an exponent t (often also called μ). If we replace the resistors by elastic springs and omit the insulating bonds, the resulting mechanical network may become unstable against shear forces even if an infinite cluster of geometrically connected springs exists. This central-force spring model can be made more realistic by including bond-bending forces, so that, for example, on a honeycomb lattice the energy is lowest if neighboring bonds have angles of 120 or 240 degrees. Then the threshold for mechanical stability agrees with the percolation threshold in the dilute network, and Young's elastic modulus vanishes with an exponent close to the scaling prediction $t + 2\nu$, which is nearly 4 in two and three dimensions. Entropic effects are still neglected in this model and could dominate in real polymer networks at finite temperatures.

Multifractality has been observed in many parts of statistical physics, including percolating networks. When we look at moments of the cluster size distribution like $\sum_s s^q n_s$ and if we know the critical exponents of these moments for $q = 2$ and 3, then scaling theory predicts correctly the critical exponents for all larger q. The exponent simply grows linearly with q. This is not so if we look at the above random resistor networks and the moments of the voltage distribution, $\sum_b V_b^q$, where the sum goes over all bonds of the random network and V_b is the voltage drop across bond b. For these moments, $q \to 0$, $q = 2$, $q = 4$ corresponds to the size of the current-carrying backbone, the resistance, and the noise, respectively. In this case, knowledge of the critical exponents for some q values does not predict the exponents for all other positive q by a linear relation as in standard scaling theory. Instead, an infinite hierarchy of apparently independent exponents appears.

VII. CONCLUSION

We have seen that percolation, though very simple to define, has a rather rich behavior. This article has centered on its critical exponents, where two exponents (e.g., ν and D) determine the static behavior and a third (e.g., μ) seems necessary (at least presently) for the dynamic aspects. Thus, the fractal dimension D of the large clusters at the percolation does not completely determine the critical exponents, but it is certainly an important parameter for studying similarities and differences between percolation and, say, the aggregates of the preceding section.

Percolation is a good way to enter modern research without too much background in physics or mathematics. It can be used, for example, to teach scaling laws, as well as renormalization group and critical phenomena.

SEE ALSO THE FOLLOWING ARTICLES

APPROXIMATIONS AND EXPANSIONS • FRACTALS • KINETICS (CHEMISTRY)

BIBLIOGRAPHY

Aharony, A. (1986). *In* "Directions in Condensed Metaphysics" (G. Grinstein and G. Mazenko, eds.), p. 1. World Scientific, Singapore.

Aharony, A. (1986). *In* "Scaling Phenomena in Disordered Systems" (R. Pymn and A. T. Skjeltorp, eds.), p. 289. Plenum, New York.

Alexander, S., and Orbath, R. (1982). *J. Phys.* (*Paris*) **43,** L625.

Alexandrowitz, Z. (1980). *Phys. Lett.* **80A,** 284.

Balberg I., and Binenbaum, N. (1985). *Phys. Rev.* **B32,** 527. But see B. I. Halperin, S. Feng, and P. S. Sen, *Phys. Rev. Lett.* **54,** 2391 (1985).

Brodbent, S. R., and Hammersley, J. M. (1957). *Proc. Cambridge Philos. Soc.* **53,** 629.

Bunde, A., and Havlin, S. (1996). "Fractals and Disordered Systems," Springer, Berlin–Heidelberg.

Burkhardt, T. W., and van Leuwen, J. M. J., eds. (1982). "Real-Space Renormalization," Topics in Current Physics, Vol. 30. Springer-Verlag, New York.

Coniglio, A. (1986). *Physica* **140A,** 51.

Deutscher, G., Zallen, R., and Adler, J., eds. (1983). "Percolation Structures and Processes," Adam Hilger, Bristol. In particular, Chap. 10 by G. Deutscher, A. Kapitulnik, and M. Rapaport.

Domb, C., Green, M. S., and Lebowitz, J. L., eds. (1972–1984). "Phase Transitions and Critical Phenomena," Vols. 1–9. Academic Press, Orlando, Florida.

Family, F., and Landau, D. P., eds. (1984). "Kinetics of Aggregation and Gelation," North Holland, Amsterdam.

Gefen, Y., Aharony A., and Alexander, S. (1983). *Phys. Rev. Lett.* **53,** 77.

Gouker, M., and Family, F. (1983). *Phys. Rev.* **B28,** 1449.

Grassberger, P. (1983). *Math. Biosciences* **62,** 157.

Grassberger, P. (1985). *J. Phys.* **A18,** L 215.

Havlin, S., and Nossal, R. (1984). *J. Phys.* **A17,** L 427.

Herrmann, H. J. (1986). *Phys. Rep.* **136,** 153.

Hoshen, J., and Kopelman, R. (1976). *Phys. Rev.* **B13,** 3428.

Kapitulnik, A., Aharony, A., Deutscher, G., and Stauffer, D. (1983). *J. Phys.* **A16,** L 269.

Kesten, H. (1982). "Percolation Theory for Mathematicians," Birkhauser, Boston.

Mandelbrot, B. B. (1982). "The Fractal Geometry of Nature," Freeman, San Francisco.

Nienhuis, B. (1982). *J. Phys.* **A14,** 199.

Rapaport, D. C. (1985). *J. Phys.* **A18,** L 175.

Rapaport, D. C. (1986). *J. Phys.* **A19,** 291.

Sahimi, M. (1994). "Applications of Percolation Theory," Taylor and Francis, London.

Stauffer, D., and Aharony, A. (1994). "Introduction to Percolation Theory," Taylor and Francis, London.

Stauffer, D., Coniglio, A., and Adam, M. (1982). *Adv. Polym. Sci.* **44,** 103.



Perturbation Theory

Carl M. Bender
Washington University

I. Perturbation Theory
II. Elementary Illustrative Example
III. Summation Theory
IV. Perturbative Solution of Differential Equations
V. Unusual Ways to Insert the Perturbation Parameter
VI. Regular Versus Singular Perturbation Theory
VII. Boundary-Layer Theory
VIII. WKB Theory
IX. Perturbative Calculation of Eigenvalues
X. Multiple-Scale Perturbation Theory

GLOSSARY

Asymptotic matching A powerful perturbative method for obtaining accurate global approximations to the solution of singular perturbative problems for differential equations.

Boundary-layer theory A singular perturbation technique for solving a boundary-value problem for a differential-equation whose highest derivative is multiplied by ϵ.

Multiple-scale perturbation theory The most general kind of perturbation theory; this theory is based on the idea that problems exhibit various kinds of phenomena on different time scales (or distance scales). Multiple-scale perturbation theory was developed by astronomers who used it to calculate the effect of small perturbations on periodic systems.

Padé summation A particularly powerful summation method in which a formal power series, which may be divergent, is truncated and then converted to a rational fraction called a Padé approximant. The Padé approximant is designated $P_N^M(\epsilon)$. The Padé approximants converge to a limit as M and N approach ∞ even though the power series may be divergent.

Perturbation coefficients The coefficients, a_n, in the perturbation series.

Perturbation parameter A small parameter, ϵ, that, when introduced into the hard problem, converts the hard problem to an infinite sequence of relatively easy problems, each of which can be solved analytically.

Perturbation series A formal series in powers of the perturbation parameter, ϵ, that represents the answer to a hard problem.

Perturbation theory A large collection of analytical

techniques for obtaining highly accurate solutions to mathematical problems that do not have simple exact analytical solutions.

Regular perturbation problem A perturbation problem for which the perturbation series is a Taylor series; that is, the perturbation series is a series in integer powers of ϵ and this series has a nonzero radius of convergence.

Singular perturbation problem A perturbation problem for which the perturbation series is not a Taylor series and/or it does not a have nonzero radius of convergence.

Summation theory A collection of methods for accelerating the convergence of a slowly converging perturbation series and/or assigning a meaningful value to a divergent perturbation series.

Uniform approximation An asymptotic perturbative approximation valid on an extended interval.

Unperturbed problem The problem obtained by setting the perturbation parameter, ϵ, equal to zero.

WKB theory A singular perturbation technique, named after Wentzel, Kramers, and Brillouin, that is used it to construct perturbative solutions to a Schrödinger equation in which the highest derivative is multiplied by ϵ.

PERTURBATION THEORY is a large collection of analytical techniques for obtaining highly accurate solutions to mathematical problems that do not have simple exact analytical solutions. Perturbative approaches fall into two classes known as *regular* and *singular* methods. In this article the basic ideas of perturbation theory are explained and illustrated by the use of examples.

I. PERTURBATION THEORY

Perturbation theory is a large collection of powerful analytical methods for constructing highly accurate approximations to the solution of a hard mathematical problem. (By a *hard problem* we mean one for which there is no simple exact closed-form analytical solution.)

There are two possible ways to attack a hard problem; one can use numerical methods or perturbative methods. The advantage of perturbative methods is that they are analytical and therefore reveal very quickly and easily the mathematical structure and behavior of the solution. Perturbation theory works by reducing a hard problem to an infinite sequence of relatively easy problems that can be solved by elementary paper-and-pencil calculations. There are three fundamental steps in the perturbative solution of a hard problem:

i. Introduce a small parameter ϵ, called a *perturbation parameter*, into the hard problem. Inserting the parameter ϵ converts the original hard problem to an *infinite class* of hard problems, one for each value of ϵ. It is crucial, however, that for the special value $\epsilon = 0$ the problem be analytically solvable. The problem corresponding to $\epsilon = 0$ is called the *unperturbed* problem. There may be many ways to introduce the perturbation parameter ϵ and it may require art and insight to decide on the best way to do so.

ii. Assume that the solution to the infinite class of hard problems parameterized by ϵ has the form of a *perturbation series*, which is a formal power series in ϵ:

$$\text{solution}(\epsilon) = \sum_{n=0}^{\infty} a_n \epsilon^n.$$

(This strong assumption is not always valid. The solution may take the more general form of a power series multiplied by a singular function of ϵ such as $e^{1/\epsilon}$.) Then one calculates the *perturbation coefficients* a_n in the perturbation series. By assumption, the first coefficient a_0 in the series can be found analytically because it is the solution to the unperturbed problem. Typically, the higher coefficients can be found recursively, with each subsequent coefficient requiring more calculation than the previous coefficient. In general, one never expects to find all of the coefficients in the perturbation series; rather, one calculates enough coefficients to give an accurate representation of the solution. The more coefficients that one calculates, the more accurate the representation.

iii. The final step is to set $\epsilon = 1$ in the perturbation series to recover the solution to the original hard problem. While the calculation of the perturbation coefficients is a routine procedure, this step may require great subtlety because the perturbation series may be divergent at $\epsilon = 1$. Thus, to complete this final step, one may have to perform a *summation* of the perturbation series, which is a way to obtain the finite solution that is represented by the divergent perturbation series. Summation theory is crucial; without it, perturbation methods would be useful only for a small and mostly uninteresting class of problems. For a general discussion of perturbative methods see Bender and Orszag (2000), Cole (1968), Kevorkian (1981), and Nayfeh (1973, 1981).

II. ELEMENTARY ILLUSTRATIVE EXAMPLE

To illustrate the three steps of perturbation theory we consider the problem of finding the real root of the quintic polynomial equation

$$x^5 + x = 1.$$

We consider this problem to be a *hard problem* because there is no analytical quadrature formula for the roots of a polynomial whose degree is higher than 4. Newton's method gives the numerical value of the real root as

$$x = 0.75487767\cdots.$$

The first step in solving this problem perturbatively is to introduce the perturbation parameter ϵ in such a way that the unperturbed problem is exactly solvable when $\epsilon = 0$. A simple way to insert the parameter ϵ is to have it multiply the term x in the polynomial:

$$x^5 + \epsilon x = 1. \tag{1}$$

The roots of this class of polynomials are now functions of the parameter ϵ. Note that when $\epsilon = 0$, the polynomials reduce to $x^5 = 1$, whose real root can be found exactly: $x(0) = 1$.

The second step is to seek a solution for the real root in the form of a power series in ϵ:

$$x(\epsilon) = \sum_{n=0}^{\infty} a_n \epsilon^n. \tag{2}$$

Note that the first term in this series is $a_0 = 1$ because the solution to the unperturbed problem is $x(0) = 1$. If we substitute this power series into the polynomial (1), we obtain a series in powers of ϵ, whose coefficients are required to vanish. Each of these coefficients is an increasingly complicated algebraic equation, the first four of which read

$$\begin{aligned} 5a_1 + 1 &= 0, \\ 5a_2 + 10a_1^2 + a_1 &= 0, \\ 5a_3 + 20a_1a_2 + a_2 + 10a_1^3 &= 0, \\ 5a_4 + 20a_1a_3 + a_3 + 10a_2^2 + 30a_1^2a_2 + 5a_1^4 &= 0. \end{aligned}$$

While the original quintic polynomial equation cannot be solved analytically, each member of this sequence of algebraic equations is extremely easy to solve in closed form. Thus, perturbation theory has reduced a hard problem to an infinite sequence of easy problems. The first few perturbation coefficients are

$$\begin{aligned} a_1 &= -\frac{1}{5}, \quad a_2 = -\frac{1}{25}, \quad a_3 = -\frac{1}{125}, \\ a_4 &= 0, \quad a_5 = \frac{21}{15625}, \quad a_6 = \frac{78}{78125}, \end{aligned} \tag{3}$$

and so on.

The third step in the perturbative solution is to substitute the perturbation coefficients into the perturbation series (2) and to sum the series at the value $\epsilon = 1$. Using the coefficients in (3), we obtain the sixth-order result

$$x(\epsilon) = 1 - \frac{1}{5}\epsilon - \frac{1}{25}\epsilon^2 - \frac{1}{125}\epsilon^3 + \frac{21}{15625}\epsilon^5 + \frac{78}{78125}\epsilon^6 + \cdots. \tag{4}$$

Can we set $\epsilon = 1$ and sum the series to get an accurate value for the solution $x(1)$ to the original problem? For this elementary problem there is an easy way to answer this question because there is a closed-form expression for the nth perturbation coefficient:

$$a_n = \frac{\Gamma[(4n-1)/5]}{5\,\Gamma[(4-n)/5]n!}. \tag{5}$$

(Note that it is very rare to be able to find a simple, closed-form expression for the nth perturbation coefficient.) From (5), we know that the radius of convergence of the perturbation series is $1.64938\cdots$. Thus, we expect the perturbation series to yield a good approximation for the real root, and setting $\epsilon = 1$ in (4) gives the sixth-order perturbative result

$$x(0) = 0.75434\cdots,$$

which agrees with the value of the root found by Newton's method to a relative error of 0.07%. This completes the perturbative treatment of the problem. Clearly, we can determine the root to any desired accuracy if we calculate enough of the perturbation coefficients.

A. Same Problem, Different Approach

Now let us see what happens if we introduce the perturbation parameter ϵ in a different way in the first step of the perturbative solution. Let us insert the parameter ϵ so that it multiplies the x^5 term rather than the x term in the polynomial:

$$\epsilon x^5 + x = 1. \tag{6}$$

Again, the roots of this class of polynomials are functions of the parameter ϵ and when $\epsilon = 0$, the root is exactly $x(0) = 1$.

The second step in the perturbative solution is to find the coefficients b_n in the perturbation expansion

$$x(\epsilon) = \sum_{n=0}^{\infty} b_n \epsilon^n. \tag{7}$$

Substituting (7) into the polynomial equation (6) again gives a sequence of easy-to-solve algebraic equations. The sixth-order perturbation series has the form

$$x(\epsilon) = 1 - \epsilon + 5\epsilon^2 - 35\epsilon^3 + 285\epsilon^4 - 2530\epsilon^5 + 23751\epsilon^6 - \cdots. \tag{8}$$

Note that the perturbation series coefficients in (8) grow rapidly with increasing n in contrast with the perturbation

series coefficients in (4), which decay with n. It is possible to determine the growth of the series coefficients because, once again, it is possible to find a closed-form expression for the perturbation coefficients b_n:

$$b_n = (-1)^n \frac{(5n)!}{(4n+1)!\,n!}. \tag{9}$$

From this result we determine that the radius of convergence of the perturbation series (8) is $\frac{256}{3125} = 0.08192$. Hence, if we attempt to evaluate the perturbation series at $\epsilon = 1$ to recover the value of the root by direct summation of the series, we will obtain a divergent answer because we are summing the power series outside of its circle of convergence. In fact, evaluating the series (8) at $\epsilon = 1$ gives the sixth-order prediction $x(1) = 21476$, which is a totally useless result for the real root at $0.75434\cdots$.

It may thus appear that the divergence of the perturbation series prevents us completing the third step in the perturbation procedure. Fortunately, we can overcome this difficulty by using a summation procedure to sum indirectly the perturbation series outside of its circle of convergence.

III. SUMMATION THEORY

Expressing the solution to a perturbation problem as a formal series in powers of ϵ is useful in the second step of perturbation theory because, as we have seen, collecting the coefficients of powers of ϵ yields an infinite sequence of easy-to-solve problems for the perturbation coefficients. However, once the second step of perturbation theory is completed we often find that the perturbation series is not an appropriate way to represent the solution to the perturbation problem because the value $\epsilon = 1$ is larger than the radius of convergence of the series. We emphasize that it is wrong to infer from the divergence of the perturbation series does not exist at $\epsilon = 1$. Indeed, the solution may exist even though the Taylor series representation of the solution about the point $\epsilon = 0$ diverges.

In order to perform the third step of perturbation theory, we must first find a better way to represent the solution for large values of ϵ. Of the many possible ways to do this, a commonly used technique is to rewrite the Taylor series $\sum_{n=0}^{\infty} a_n \epsilon^n$ in the form of a continued fraction of the form

$$\cfrac{c_0}{1 - \cfrac{c_1\epsilon}{1 - \cfrac{c_2\epsilon}{1 - \cfrac{c_3\epsilon}{1 - \cdots}}}}.$$

Note that the Taylor coefficients a_n uniquely determine the continued-fraction coefficients c_n via a sequence of invertible equations:

$$\begin{aligned} c_0 &= a_0, \\ c_1 &= \frac{a_1}{a_0}, \\ c_2 &= \frac{a_0 a_2 - a_1^2}{a_0 a_1}, \\ c_3 &= \frac{a_0 a_1 a_3 - a_0 a_2^2}{a_0 a_1 a_2 - a_1^3}, \end{aligned}$$

and so on. (These equations are *invertible* because given the coefficients a_n, we can calculate the coefficients c_n, and given c_n we can calculate a_n.) Note that if we use these continued-fraction coefficients and calculate the Taylor expansion of the continued fraction in powers of ϵ, then we recover the Taylor series $\sum_{n=0}^{\infty} a_n \epsilon^n$.

The advantage of using a continued-fraction representation in place of a Taylor series is that the continued fraction often converges in a much larger region of the complex-ϵ plane than the Taylor series. Rather than converging in a circle, a continued fraction often converges in the entire complex plane or the entire complex plane except for a cut (a ray emanating from the origin).

If we truncate the continued-fraction representation after the coefficient c_n and rationalize the result, we obtain a ratio of two polynomials of degree M in the numerator and degree $N = M$ or $N = M + 1$ in the denominator (depending on whether n is even or odd), where $M + N = n$. This rational fraction is called the (M, N) Padé approximant of the Taylor series and it is designated $P_N^M(\epsilon)$.

To summarize, by converting a Taylor series to a Padé sequence, we replace the sequence of partial sums of the Taylor series

$$\sum_{n=0}^{K} a_n \epsilon^n \qquad (K = 0, 1, 2, 3, \cdots)$$

by a sequence of Padé approximants called the *main sequence*:

$$P_0^0(\epsilon),\ P_0^1(\epsilon),\ P_1^1(\epsilon),\ P_1^2(\epsilon), \cdots.$$

This padé sequence often converges for values of ϵ outside the circle of convergence of the Taylor series. Padé approximants are used routinely when one encounters slowly convergent or divergent power series. However, from a theoretical point of view much work needs to be done; no general and comprehensive theory concerning the convergence of Padé approximants exists. Padé summation is only one of many summation techniques available; another useful technique is called the Wynn epsilon algorithm method.

To illustrate Padé summation we use it to sum the perturbation series in (8). Recall that a direct summation of the first seven terms in this Taylor series evaluated at $\epsilon = 1$

gives the useless result $x(1) = 21476$. However, if we first convert the Taylor series to a (3,3)-Padé approximant and then evaluate this approximant at $\epsilon = 1$, we obtain the result $x(1) = 0.76369$, which agrees with the exact numerical answer to a relative error of 1.2%.

There are many good references on divergent series, acceleration of convergence, Padé theory, and continued fractions; see, for example, Borel (1975) and references therein.

IV. PERTURBATIVE SOLUTION OF DIFFERENTIAL EQUATIONS

Perturbation theory can be used to solve nontrivial differential-equation problems. Consider, for example, the Schrödinger equation initial-value problem

$$y''(x) = Q(x)y(x), \quad y(0) = 1, \quad y'(0) = 0, \tag{10}$$

where $Q(x)$ is an arbitrary continuous function of x. This is a *hard problem* because there is no quadrature solution for a Schrödinger equation. However, it is extremely easy to solve this problem using perturbative methods. We introduce the parameter ϵ so that it multiplies the function $Q(x)$:

$$y''(x) = \epsilon Q(x)y(x), \quad y(0) = 1, \quad y'(0) = 0 \tag{11}$$

and seek a solution in the form of a series in powers of ϵ:

$$y(x) = \sum_{n=0}^{\infty} y_n(x)\epsilon^n, \tag{12}$$

where we incorporate the initial conditions by requiring that

$$\begin{aligned} y_0(0) &= 1, \\ y_n(0) &= 0 \quad (n \geq 1), \\ y_n'(0) &= 0 \quad (n \geq 0). \end{aligned}$$

Note that we have introduced ϵ in such a way that it is possible to solve the unperturbed problem in closed form. The solution is simply

$$y_0(x) = 1.$$

Solving the higher-order problems is equally easy. The function $y_n(x)$ is obtained by integrating the product $Q(x)y_{n-1}(x)$ twice:

$$y_n(x) = \int_0^x ds \int_0^s dt\, Q(t)y_{n-1}(t).$$

Recovering the function $y(x)$ from the perturbation series (12) is straightforward because, as we will now show, this series is rapidly convergent if $Q(x)$ is continuous. Let M be the maximum value of $|Q(x)|$ on the interval $0 \leq x \leq a$. Then, $|y_n(x)|$ is bounded by $a^n M^n/(2n)!$. Hence, only a small number of terms in the series (12) are needed to calculate the value of $y(x)$ with extremely high precision.

For further references on Perturbation methods for differential equations see Kevorkian and Cole (1996) and O'Malley (1991).

V. UNUSUAL WAYS TO INSERT THE PERTURBATION PARAMETER

An interesting and surprising way to introduce the perturbation parameter ϵ is to have it measure the nonlinearity of a problem. To illustrate this idea, let us return to the polynomial equation considered in Section II. Rather than inserting ϵ so that it multiplies the x term or the x^5 term in the polynomial equation, let us insert ϵ into the *exponential*:

$$x^{1+\epsilon} + x = 1.$$

Note that when $\epsilon = 0$, this equation can be solved exactly and the solution for the real root is $x(0) = 1/2$.

We assume that the real root $x(\epsilon)$ of this equation has a perturbation expansion in powers of ϵ of the form in (2). We then find that after substituting this series into the polynomial equation and collecting powers of ϵ, the coefficients in this expansion are

$$\begin{aligned} a_0 &= \frac{1}{2}, \\ a_1 &= \frac{1}{4}\log 2, \\ a_2 &= -\frac{1}{8}\log 2, \\ a_3 &= -\frac{1}{48}\log^3 2 + \frac{1}{32}\log^2 2 + \frac{1}{16}\log 2, \\ a_4 &= \frac{1}{32}\log^3 2 - \frac{3}{64}\log^2 2 - \frac{1}{32}\log 2, \end{aligned}$$

and so on. It can be shown that the radius of convergence of this expansion is 1. Thus, we cannot recover the positive root of the original problem by setting $\epsilon = 4$ in the perturbation expansion and summing the perturbation series directly. However, if we first convert the perturbation expansion to a Padé sequence and evaluate the Padé approximants in the sequence at $\epsilon = 4$, we obtain extremely accurate numerical results. For example, the numerical value of the (3, 3) Padé approximant is 0.75448 (relative error of 0.05%) and the numerical value of the (6, 6) Padé approximant is 0.75487654 (relative error of 0.00015%).

Let us use this technique to solve a particularly difficult boundary-value problem for a nonlinear differential

equation known as the Thomas–Fermi equation. The boundary-value problem reads

$$y''(x) = \frac{y^{3/2}(x)}{\sqrt{x}}, \quad y(0) = 1, \quad y(\infty) = 0. \tag{13}$$

This problem is particularly important in physics because it describes the distribution of electrical charge in a nucleus. The objective is to find the value of $y'(0)$. A numerical analysis of this problem gives the result

$$y'(0) = -1.5880710\cdots. \tag{14}$$

To solve this problem perturbatively, we introduce the parameter ϵ into the exponent so that it measures the nonlinearity of the problem:

$$y''(x) = y(x)\left(\frac{y(x)}{\sqrt{x}}\right)^{\epsilon}, \quad y(0) = 1, \quad y(\infty) = 0. \tag{15}$$

Note that when $\epsilon = 0$, the exact solution to the problem is $y(x) = e^{-x}$, so that $y'(0) = -1$. We then assume that $y(x)$ has a formal perturbation expansion in powers of ϵ:

$$y(x) \;=\; e^{-x} + \sum_{n=1}^{\infty} y_n(x)\epsilon^n.$$

Substituting this expansion into the boundary-value problem (15), collecting powers of ϵ, and routinely calculating the coefficient functions $y_1(x)$, $y_2(x)$, $y_3(x)$, and so on gives an accurate result for $y'(0)$. Specifically, substituting $\epsilon = 1/2$ into the first-order perturbation expansion gives $y'(0) = -1.19259$ (25% relative error), substituting $\epsilon = 1/2$ into the second-order perturbation expansion gives $y'(0) = -1.38428$ (13% relative error), and substituting $\epsilon = 1/2$ into the third-order perturbation expansion gives $y'(0) = -1.47892$ (6.9% relative error). Furthermore, the (1,2) Padé of the third-order perturbation expansion reduces the error substantially and gives $y'(0) = -1.61629$ (1.8% relative error).

This technique of introducing the perturbation parameter ϵ into the exponent of a nonlinear problem (so that ϵ is a measure of the nonlinearity of the problem) is useful for solving a wide variety of nonlinear problems (the Blasius equation, the Korteweg-de Vries equation, the Duffing equation, and so on) (Bender *et al.*, 1991).

VI. REGULAR VERSUS SINGULAR PERTURBATION THEORY

All of the examples of perturbative problems we have considered so far are called *regular* perturbation problems. The identifying characteristic of a regular perturbation problem is that the perturbation series is a power series in integer powers of ϵ and that the power series has a nonzero radius of convergence. Thus, the perturbation series is a Taylor series. If the perturbation series is not a Taylor series and/or it does not a have nonzero radius of convergence, the perturbation problem is called a *singular* perturbation problem.

Roughly speaking, in a singular perturbation problem the limit $\epsilon \to 0$ is not smooth. Rather, one observes that as the perturbation parameter ϵ vanishes, there is an abrupt change in the character of the problem. Typically, a singular perturbation problem arises if the perturbation parameter ϵ is inserted into a differential equation so that it multiplies the highest derivative in the equation. Thus, as ϵ tends to 0, the order of the differential equation decreases and the solution to the lower-order differential equation is unable to satisfy all of the boundary conditions or initial conditions. Thus, the solution abruptly ceases to exist as $\epsilon \to 0$.

In a singular perturbation problem we often cannot use the term *unperturbed problem* because the solution to the problem with ϵ set to 0 may not exist. Instead, we refer to the first term in the perturbation expansion as the *leading-order* term. The leading-order term often ceases to exist in the limit $\epsilon \to 0$.

To illustrate the kinds of abrupt changes that may occur in the limit $\epsilon \to 0$ we consider two different boundary-value problems:

$$\epsilon y''(x) + y'(x) = 0, \quad y(0) = 0, \quad y(1) = 1, \tag{16}$$

whose exact solution is

$$y(x) = \frac{1 - e^{-x/\epsilon}}{1 - e^{-1/\epsilon}}, \tag{17}$$

and

$$\epsilon y''(x) + y(x) = 0, \quad y(0) = 0, \quad y(1) = 1, \tag{18}$$

whose exact solution is

$$y(x) = \frac{\sin(x/\sqrt{\epsilon})}{\sin(x/\sqrt{\epsilon})}. \tag{19}$$

Observe that as $\epsilon \to 0$ it is not possible to satisfy the boundary conditions in these problems. Thus, in both cases the exact solutions become discontinuous in this limit. However, in the first example, the exact solution only becomes discontinuous at a point (at $x = 0$). The localized region in which the solution becomes discontinuous is called a *boundary layer.* In the second example the solution becomes discontinuous on the entire region from $x = 0$ to $x = 1$ in this limit. Singular perturbation problems in which the solution becomes discontinuous in a narrow region are solved using a perturbative technique called *boundary-layer theory*. *WKB theory* is used to solve singular perturbation problems in which the solution becomes discontinuous throughout an extended region.

Boundary-layer theory and WKB theory require the use of a technique called *asymptotic matching*. Examples of applications of these perturbation theories are considered in Sections VII and VIII (Lagerstrom, 1988).

VII. BOUNDARY-LAYER THEORY

We explain and illustrate the ideas of boundary-layer theory in the context of a simple but general example. Consider the boundary-value problem

$$\begin{aligned} y''(x) + a(x)y'(x) + b(x)y(x) &= 0, \\ y(0) = A, \qquad y(1) &= B. \end{aligned} \tag{20}$$

There is no quadrature solution to a second-order linear differential equation, so we will solve this problem as a perturbation problem. We begin by introducing the parameter ϵ into this problem. If we introduce ϵ so that it multiplies the $y'(x)$ or the $y(x)$ terms, then we have a regular perturbation problem. In this section we are concerned with the singular perturbation problem that arises if we introduce ϵ so that it multiplies the highest-derivative term in the equation:

$$\begin{aligned} \epsilon y''(x) + a(x)y'(x) + b(x)y(x) &= 0, \\ y(0) = A, \qquad y(1) &= B. \end{aligned} \tag{21}$$

We will now make a crucial technical assumption; namely, that for all x $(0 \le x \le 1)$ the function $a(x)$ is strictly positive $[a(x) > 0]$. If this assumption holds, then as $\epsilon \to 0$, the exact solution to the boundary-value problem (21) remains smooth except in a narrow boundary-layer region located at $x = 0$. We do not prove this assertion here. However, we give a simple physical argument: If we think of the variable x as representing time, then $y''(x)$ is the acceleration of a particle of mass ϵ, the $y'(x)$ term represents a friction (velocity-dependent) force, and the $y(x)$ term represents a position-dependent force. If the function $a(x)$ is positive, then the friction force causes *damping* (as opposed to runaway behavior). For very small mass ϵ, the particle accelerates rapidly and its position varies rapidly with time until the damping and position-dependent forces balance. From this point on the position of the particle is a slowly changing function of time.

We begin the analysis by considering the *outer region*; namely, all x away from $x = 0$ where the solution is slowly varying as $\epsilon \to 0$. In this region we assume that the solution has a series expansion in powers of ϵ:

$$y_{\text{out}}(x) = \sum_{n=0}^{\infty} y_n(x)\epsilon^n, \tag{22}$$

where $y_0(1) = B$, $y_n(1) = 0$ $(n \ge 1)$. Substituting this series into the differential equation (21) gives a sequence of equations for the functions $y_n(x)$:

$$\begin{aligned} a(x)y_0'(x) + b(x)y_0(x) &= 0, \\ a(x)y_1'(x) + b(x)y_1(x) &= -y_0''(x), \\ a(x)y_2'(x) + b(x)y_2(x) &= -y_1''(x), \\ a(x)y_3'(x) + b(x)y_3(x) &= -y_2''(x), \end{aligned}$$

and so on. Each of these equations can be solved in quadrature form because they are *first*-order rather than second-order linear equations.

The solution to the first of these equations is

$$y_0(x) = B \exp\left[\int_x^1 dt \frac{b(t)}{a(t)}\right], \tag{23}$$

where we have incorporated the boundary condition at $x = 1$. We can just as easily solve for the functions $y_1(x)$, $y_2(x)$, and so on.

Next, we consider the *inner region*; namely, all x near the point $x = 0$. In this region the solution becomes a rapidly varying function of x as ϵ tends to 0. Let us begin by determining the width δ of this region. We introduce the *inner variable* X by means of a scaling transformation:

$$x = \delta X. \tag{24}$$

In terms of the inner variable, the differential equation (21) reads

$$\frac{\epsilon}{\delta^2} Y_{\text{in}}''(X) + \frac{1}{\delta} a(\delta X) Y_{\text{in}}'(X) + b(\delta X) Y_{\text{in}}(X) = 0, \tag{25}$$

where $Y_{\text{in}}(X) = y(x)$.

We must now identify the distinguished limits of this equation. (A *distinguished limit* is an asymptotic dominant balance between the two largest terms in the equation.) For small ϵ, there are two possible distinguished limits. First, we can have a balance between the second and third terms with the first term being negligible. This occurs if $\delta = 1$. This distinguished limit is just the outer limit. Second, we can have a balance between the first and second terms with the third term being negligible. This occurs if $\delta = \epsilon$. It is this distinguished limit that characterizes the inner region. We conclude that the thickness of the inner region is ϵ and that the differential equation in the inner region is

$$Y_{\text{in}}''(X) + a(\epsilon X) Y_{\text{in}}'(X) + \epsilon b(\epsilon X) Y_{\text{in}}(X) = 0. \tag{26}$$

We seek a perturbative solution to this equation of the form

$$Y_{\text{in}}(X) = \sum_{n=0}^{\infty} Y_n(X)\epsilon^n. \tag{27}$$

The boundary condition $y(0) = A$ then takes the form $Y_0(0) = A, Y_1(0) = 0, Y_2(0) = 0, Y_3(0) = 0$, and so on. Substituting (27) into (26) and collecting powers of ϵ gives the following sequence of differential equations for the functions $Y_0(X), Y_1(X), Y_2(X), \ldots$:

$$Y_0''(X) + a(0)Y_0'(X) = 0,$$
$$Y_1''(X) + a(0)Y_1'(X) = -a'(0)Y_0'(X) - b(0)Y_0(X),$$
$$Y_2''(X) + a(0)Y_2'(X) = -a'(0)Y_1'(X) - \frac{1}{2}a''(0)Y_0'(X) - b(0)Y_1(X) - b'(0)Y_0(X), \quad (28)$$

and so on, where we assume that $a(x)$ and $b(x)$ possess Taylor expansions about $x = 0$ of the form

$$a(x) = a(0) + a'(0)x + \frac{1}{2}a''(0)x^2 + \cdots,$$
$$b(x) = b(0) + b'(0)x + \frac{1}{2}b''(0)x^2 + \cdots.$$

Note that each of the equations in (28) is solvable because the homogeneous parts are constant-coefficient differential equations. The first of these equations has the same structure as that in (16) and its general solution is

$$Y_0(X) = C_0 + (A - C_0)e^{-a(0)X}, \quad (29)$$

where C_0 is an arbitrary constant. [The solutions for $Y_1(X)$, $Y_2(X)$, ... each contain new arbitrary constants C_1, C_2, and so on.]

The problem is now to determine the constants in $Y_n(X)$. This is accomplished by the use of asymptotic matching. We argue as follows: The exists a region in which the inner expansion (27) and the outer expansion (22) are *both* valid. The points x that are common to both regions are said to belong to the *overlap region*. In this overlap region we perform an asymptotic match by (i) finding an *asymptotic approximation of the asymptotic expansion* $Y_{\text{in}}(X)$ that is valid in the overlap region; (ii) finding an *asymptotic approximation of the asymptotic expansion* $y_{\text{out}}(x)$ that is valid in the overlap region; (iii) demanding that these two asymptotic approximations be identical in the overlap region.

We perform this procedure to leading order in ϵ as follows: The outer solution in (23) is valid throughout the entire outer region, but in the matching region, where x is small compared with 1, we have

$$y_0(x) \sim B \exp\left[\int_0^1 dt\, \frac{b(t)}{a(t)}\right].$$

The inner solution in (29) is valid throughout the entire inner region, but in the matching region, where x is large compared with ϵ, we have

$$Y_0(X) \sim C.$$

Requiring that these two approximations to the solution to the differential equation agree in the overlap region gives the result

$$C = B \exp\left[\int_0^1 dt \frac{b(t)}{a(t)}\right].$$

The process of asymptotic matching reveals the extent of the matching region. We have just seen that to leading order in ϵ the asymptotic matching occurs in the overlap region defined by

$$\epsilon \ll x \ll 1.$$

As the asymptotic matching is performed to higher order in ϵ the size of the overlap region shrinks. Typically, the overlap region to nth order in ϵ is given by

$$\epsilon \ll x \ll \epsilon^{n/(n+1)}.$$

The final step in the boundary-layer procedure is to combine the inner and outer approximations to the solution of the boundary-value problem to obtain a single expression that is a *uniform* approximation valid on the entire interval $0 \le x \le 1$. We do so by adding the inner solution to the outer solution and subtracting off the doubly-counted solution in the overlap region:

$$y_{\text{uniform}}(x) \equiv Y_{\text{in}}(X) + y_{\text{out}}(x) - y_{\text{match}}(x).$$

Thus, to leading order, we obtain

$$y_{\text{uniform},0}(x) = B \exp\left[\int_x^1 dt \frac{b(t)}{a(t)}\right] \times \left\{A - B \exp\left[\int_0^1 dt \frac{b(t)}{a(t)}\right]\right\} e^{-a(0)x/\epsilon}.$$

The function $y_{\text{uniform},0}(x)$ is a *global* approximation to $y(x)$ over the entire region in the sense that the difference between the exact solution $y(x)$ and $y_{\text{uniform},0}(x)$ is of order ϵ for all x $(0 \le x \le 1)$. The higher-order uniform asymptotic approximations to $y(x)$ require more algebra, but are routinely obtained by using the same procedure.

Boundary-layer theory can be used to solve (21) even if the restriction that $a(x)$ be strictly positive is lifted. If $a(x)$ is strictly negative, the ensuing analysis is identical except that the boundary layer occurs at the right boundary $x = 1$ instead of at $x = 0$. If $a(x)$ is not one-signed, then wherever it passes through 0 it is possible to have an *internal boundary layer* (a narrow region of rapid variation between $x = 0$ and $x = 1$).

Boundary-layer techniques also apply to differential-equations boundary-value problems in which the differential equation is not second order and even if the differential equation is nonlinear. In these more complicated problems the boundary layers that arise may be thicker or thinner than ϵ; indeed, we often find that δ is a fractional power of ϵ. Also, there may be multiple boundary layers and even nested boundary layers (boundary layers inside of boundary layers).

Boundary-layer methods are often used to solve problems in fluid mechanics for which we may assume that the viscosity of the fluid is small. For a low-viscosity fluid

flowing along a boundary surface there are two physically distinct regions. Near the boundary surface there is a narrow boundary layer in which the velocity profile of the fluid changes rapidly as a function of the distance from the boundary. Right at the boundary surface the velocity of the fluid is zero because we assume that there is a *no-slip* condition. Away from the boundary (in the outer region) the velocity profile is a slowly varying function of the distance from the boundary because the effects of the boundary do not propagate deeply into a fluid of low viscosity.

VIII. WKB THEORY

WKB theory is a perturbation technique named after Wentzel, Kramers, and Brillouin, who used it to construct perturbative solutions to the Schrödinger equation

$$\epsilon^2 y''(x) = Q(x)y(x), \tag{30}$$

where the perturbation parameter ϵ is proportional to Planck's constant $\hbar$. Problems like this cannot be solved using boundary-layer theory because for small ϵ the solution varies rapidly over regions that are not narrow. To solve (30) we seek a solutions of the form

$$y_{\mathrm{WKB}}(x) = \exp\left[\frac{1}{\epsilon}\sum_{n=0}^{\infty} S_n(x)\epsilon^n\right]. \tag{31}$$

Substituting (31) into (30) and collecting powers of ϵ gives a sequence of easy-to-solve equations for the functions $S_n(x)$:

$$\begin{aligned} [S_0'(x)]^2 &= Q(x), \\ 2S_0'(x)S_1'(x) + S_0''(x) &= 0, \\ 2S_0'(x)S_n'(x) + S_{n-1}''(x) + \sum_{j=1}^{n-1} S_j'(x)S_{n-j}'(x) &= 0 \\ (n \geq 2). \end{aligned} \tag{32}$$

The solution to the first of these equations, which is called the *eikonal* equation, is

$$S_0(x) = \pm\int^x dt\sqrt{Q(t)},$$

and substituting it into (31) gives what is known as the *geometrical-optics* approximation to $y(x)$:

$$y_{\mathrm{geometricaloptics}}(x) = \exp\left[\pm\frac{1}{\epsilon}\int^x dt\sqrt{Q(t)}\right].$$

The solution to the second of these equations, which is called the *transport* equation, is

$$S_1(x) = -\frac{1}{4}\ln Q(x).$$

Substituting both $S_0(x)$ and $S_1(x)$ into (31) gives what is known as the *physical-optics* approximation to $y(x)$:

$$y_{\text{physical optics}}(x) = Q^{-1/4}(x)\left\{C_1\exp\left[\frac{1}{\epsilon}\int_a^x dt\sqrt{Q(t)}\right] + C_2\exp\left[-\frac{1}{\epsilon}\int_a^x dt\sqrt{Q(t)}\right]\right\},$$

where the lower limit of integration is any fixed point a.

For those values of x for which $Q(x) \neq 0$ the physical-optics approximation is the leading-order asymptotic approximation to the solution $y(x)$ of the differential Eq. (30). That is, for small ϵ it differs from $y(x)$ by an amount of order ϵ. Values of x for which $Q(x) = 0$ are called *turning points* and near these turning points the physical-optics approximation is not valid. In the vicinity of a turning point at $x = x_0$, one may replace the differential Eq. (30) by

$$\epsilon^2 y''(x) = \alpha(x - x_0)y(x), \tag{33}$$

where we have replaced the function $Q(x)$ by the first term in its Taylor series:

$$Q(x) \sim \alpha(x - x_0) \qquad (|x - x_0| \ll 1).$$

The differential Eq. (33) is a scaled version of the Airy equation $y''(x) = xy(x)$ whose solutions are $\mathrm{Ai}(x)$ and $\mathrm{Bi}(x)$. Thus, to solve a differential equation of the form (30) when there is a turning point, it is necessary to match asymptotically the WKB physical optics solutions on either side of the turning point to the solution of the Airy equation in the vicinity of the turning point.

The uniform asymptotic approximation to the solution of the one-turning-point problem was first discovered by Langer. For the case $x_0 = 0$ the leading-order one-turning-point solution that vanishes as $x \to +\infty$ has the form

$$y_{\mathrm{uniform}}(x) = C\left[\frac{3}{2\epsilon}S_0(x)\right]^{1/6} Q^{-1/4}(x)\mathrm{Ai}\left[\frac{3}{2\epsilon}S_0(x)\right]^{2/3}, \tag{34}$$

where $S_0(x) = \int_0^x dt\,\sqrt{Q(t)}$ and C is an arbitrary constant.

The one-turning-point solution in (34) is exponentially decaying to the right of the turning point ($x > 0$) and oscillatory to the left of the turning point ($x < 0$). The region in which the solution decays exponentially is called the *classically forbidden* region; the region in which the solution is oscillatory is called the *classically allowed* region. These different kinds of regions underscore the differences between classical and quantum mechanics: In classical mechanics a particle may not penetrate into a region in which its kinetic energy is negative; a quantum particle may do so with exponentially small probability.

For the case of two turning points, there are two possibilities: (1) The classically allowed (oscillatory) region may lie between the two turning points, or (2) the classically forbidden region may lie between the two turning points. Let the two turning points be located at x_1 and x_2. When the classically allowed region lies between these two turning points and if the constraint that $y(x) \to 0$ as $x \to \pm\infty$ is imposed, this translates into the WKB *quantization condition*

$$\frac{1}{\epsilon}\int_{x_1}^{x_2} dt\sqrt{-Q(t)} = \left(n+\frac{1}{2}\right)\pi \qquad (n=0,1,2,\ldots). \tag{35}$$

For the case of the Schrödinger Eq. (30) in which we set $Q(x) = V(x) - E$, this quantization condition gives accurate results for the energy eigenvalues E_n for small ϵ. Also, for *fixed* ϵ it gives accurate results for E_n when n is large.

For example, for the eigenvalue problem

$$y''(x) = (x^4 - E)y(x), \qquad y(\pm\infty) = 0, \tag{36}$$

the quantization (35) condition reads

$$\int_{-E^{1/4}}^{E^{1/4}} dt\sqrt{E - t^{1/4}} = \left(n+\frac{1}{2}\right)\pi \quad (n=0,1,2,\ldots), \tag{37}$$

and from evaluating the integral we get

$$E_n \sim \left[\frac{3\Gamma\left(\frac{3}{4}\right)\left(n+\frac{1}{2}\right)\sqrt{\pi}}{\Gamma\left(\frac{1}{4}\right)}\right]^{4/3} \qquad (n\to\infty). \tag{38}$$

This result is extremely accurate. For example, the exact value of E_{10} is $50.256\cdots$, while the WKB prediction gives 50.240 (relative error of -0.03%).

When the classically forbidden region lies between these two turning points at x_1 and x_2, we have the quantum phenomenon of tunneling through a potential barrier. WKB theory predicts that the transmission coefficient T for a beam of particles of energy E incident on a potential $V(x)$, where the maximum value of $V(x)$ is greater than E, is

$$T \sim \exp\left[-\frac{2}{\epsilon}\int_{x_1}^{x_2} dt\sqrt{V(t)-E}\right] \quad (\epsilon\to 0). \tag{39}$$

Note that this tunneling probability is exponentially small but *nonvanishing*.

The WKB quantization condition (35) and tunneling amplitude (39) are the leading-order perturbative predictions. These predictions can be refined by doing higher-order WKB calculations (Bender and Orszag, 1978). More advanced treatments of WKB theory involve the use of *complex* WKB matching techniques (Berry and Mount, 1972; Dingle, 1973).

IX. PERTURBATIVE CALCULATION OF EIGENVALUES

Enormous research effort has been spent in trying to solve eigenvalue problems using perturbation theory. Consider, for example, the Schrödinger equation eigenvalue problem

$$y''(x) = [V(x) + W(x) - E]y(x), \qquad y(\pm\infty) = 0. \tag{40}$$

To solve this problem perturbatively, we may insert the perturbation parameter ϵ so that it multiplies the function $W(x)$ so long as the simpler unperturbed eigenvalue problem

$$y_0''(x) = [V(x) - E_0]y_0(x), \qquad y(\pm\infty) = 0, \tag{41}$$

can be solved exactly and in closed form.

The eigenvalues and eigenfunctions of the perturbed problem

$$\begin{aligned} &y''(x,\epsilon) = [V(x) + \epsilon W(x) - E(\epsilon)]y(x,\epsilon),\\ &y(\pm\infty) = 0, \end{aligned} \tag{42}$$

can be expressed as perturbation series (power series in ϵ). The perturbation series for each of the eigenvalues has the form

$$y(x,\epsilon) = \sum_{n=0}^{\infty} y_n(x)\epsilon^n, \qquad E(\epsilon) = \sum_{n=0}^{\infty} E_n\epsilon^n. \tag{43}$$

These series are called a Rayleigh–Schrödinger perturbation series.

One particularly elegant method for obtaining the Rayleigh–Schrödinger perturbation series for the eigenvalue $E(\epsilon)$ is to represent the coefficients in the expansion in terms of Feynman graphs. The Feynman rules for constructing and evaluating these graphs are in wide use by theoretical physicists.

A simple example of an eigenvalue problem that one can solve in terms of a Rayleigh–Schrödinger perturbation series is

$$y''(x,\epsilon) = \left[\frac{1}{4}x^2 + \epsilon\frac{1}{4}x^4 - E(\epsilon)\right]y(x,\epsilon), \qquad y(\pm\infty) = 0. \tag{44}$$

For this example, called the *quartic anharmonic oscillator*, the perturbation expansion for the smallest eigenvalue is

$$E(\epsilon) \sim \frac{1}{2} + \frac{3}{4}\epsilon - \frac{21}{8}\epsilon^2 + \frac{333}{16}\epsilon^3 + \cdots. \tag{45}$$

The radius of convergence of this power series is 0. The series is asymptotic but not convergent.

The perturbation series (43) have been the object of intensive mathematical study for several decades. Among

the principal discoveries are that the radius of convergence of the series for $E(\epsilon)$, which in most cases is 0, is determined by singularities in the complex-ϵ plane which are square-root singularities. In general there are an infinite number of these singularities, and when the radius of convergence is 0, these singularities form a sequence in the complex-ϵ plane that has a limit point at the origin. The square-root singularities are points where pairs of eigenvalues become degenerate. Thus, analytically continuing around one of the square-root singularities exchanges one of the eigenvalues with another. This exchange process is called *level crossing*. (Bender and Wu, 1973; Le Guillou and Zinn-Justin, 1990; Reed and Simon, 1972; Simon, 1970).

In general, all of the energy levels are *analytic continuations of one another* and one can think of the entire spectrum as being represented by a single eigenvalue function $E(\epsilon)$ that is defined on a complex Riemann surface in the variable ϵ. Each of the eigenvalues is in one-to-one correspondence with the sheets of this Riemann surface. That is, we recover the kth eigenvalue by evaluating $E(\epsilon)$ on the kth sheet of the Riemann surface. Apparently then, quantization is a geometrical phenomenon and consists of nothing more than counting the sheets in a Riemann surface.

Furthermore, it can be shown that the function $E(\epsilon)$ is a *Stieltjes* function and, as a consequence, the Padé sequence formed from the power series (43) converges. So long as the coefficients in this series do not grow faster than $(2n)!$ this Padé sequence converges to the exact eigenvalue (see Le Guillou and Zinn-Justin, 1990). Many numerical and theoretical studies have contributed to the vast literature on the perturbation theory of eigenvalue problems.

X. MULTIPLE-SCALE PERTURBATION THEORY

The most general kind of perturbation theory, called *multiple-scale perturbation theory*, is based on the idea that physical problems exhibit various kinds of physical phenomena on different time scales (or distance scales). Multiple-scale perturbation theory was developed by astronomers who used it to calculate the effect of small perturbations on periodic systems. The objective in developing this perturbative approach was to calculate the small changes in the shapes of planetary orbits due to the weak influences of remote planets.

To illustrate the phenomenon of multiple scales we consider a very simple periodic system (a classical harmonic oscillator) subject to a small perturbing force. The model, which is described by the *Duffing* equation, consists of a particle attached to the end of a spring for which the restoring force is not just linear; rather, we consider the first nonlinear correction to Hooke's law. The position of such a particle obeys the equation

$$y''(t) + y(t) + \epsilon y^3(t) = 0. \tag{46}$$

The presence of the perturbation parameter ϵ multiplying the $y^3(t)$ term emphasizes that the deviation from Hooke's law is very small.

Let us solve the differential Eq. (46) subject to the initial conditions $y(0) = 1$ and $y'(0) = 0$. Since the perturbation problem is a *regular* perturbation problem (no abrupt changes occur as $\epsilon \to 0$), we may assume that the function $y(t)$ has a regular perturbation expansion with a nonzero radius of convergence:

$$y(t) = \sum_{n=0}^{\infty} y_n(t)\epsilon^n. \tag{47}$$

Note that the unperturbed problem $y''(t) + y(t) = 0$ is easy to solve and that the solution obeying the initial conditions is

$$y_0(t) = \cos t. \tag{48}$$

Substituting this result into the equation for $y_1(t)$ gives

$$y_1''(t) + y_1(t) = -\cos^3 t, \tag{49}$$

whose solution is

$$y_1(t) = \frac{1}{32}\cos(3t) - \frac{1}{32}\cos t - \frac{3}{8}t\,\sin t. \tag{50}$$

The third term in this equation is said to be *secular* because it grows with t. In this case it grows linearly with t because (49) represents a harmonic oscillator whose driving term, $\cos^3 t = \frac{1}{4}\cos(3t) + \frac{3}{4}\cos t$, contains the function $\cos t$, and this function oscillates at the *natural frequency* of the harmonic oscillator. Whenever a harmonic oscillator is driven on resonance, the solution grows rapidly with t.

The presence of a secular term tells us that the perturbation series (47) becomes inaccurate for large values of t, specifically $t > \frac{1}{\epsilon}$. This is because a finite number of terms in the perturbation series evidently contradict the property that the solution to the differential Eq. (46) is bounded. (To show that the solution is bounded we construct an energy integral by integrating the differential Eq. (46) once with respect to t.) The phenomenon that we observe here is quite remarkable: As is always the case in perturbation theory problems involving differential equations, when we collect powers of ϵ, we obtain a sequence of linear differential equations all of whose homogeneous parts are identical and whose inhomogeneous parts differ from order to order. What we see here is that higher orders of perturbation theory can be driven *on resonance* (that is,

they can be coupled resonantly) to lower orders of perturbation theory.

To treat the presence of secular terms in the conventional perturbation series we perform a rather fancy rearrangement of the terms in the perturbation series. In this case, we can calculate the most secular term in each order of the perturbation series. The most secular contribution to $y_n(t)$ has the form

$$\frac{1}{2n!}t^n\left(\frac{3i}{8}\right)^n e^{it} + \text{complex conjugate.} \qquad (51)$$

Substituting this term into the perturbation series and summing to all orders gives

$$y(t) = \cos\left[\left(1+\frac{3}{8}\epsilon\right)t\right] + \mathrm{O}(\epsilon). \qquad (52)$$

Note that the result is a *bounded* function of t. This calculation shows that while a finite number of terms in the perturbation expansion can exhibit secularity, an infinite rearrangement and summation of the perturbation series may no longer be secular.

More importantly, the rearrangement of the series demonstrates that various phenomena appear on different time scales. On the short time scale ($t \ll \frac{1}{\epsilon}$) the solution exhibits harmonic oscillation with frequency 1. On a longer time scale (t of order $\frac{1}{\epsilon}$) there is a frequency shift of order $\frac{3}{8}\epsilon$.

The notion that characteristic phenomena occur at different scales is apparent in all forms of perturbation theory. For example, we saw that in boundary-layer theory, there is a short-distance scale inside the boundary-layer in which the rapidly-varying inner solution occurs. The long-distance slowly-varying behavior of the outer solution is a distinctly different phenomenon. The same kind of rapidly-varying behavior (geometrical optics) and slowly varying behavior (modulation of the geometrical-optics solution to obtain the physical-optics solution) occur in WKB theory. Indeed, the techniques of boundary-layer theory may be used to derive boundary-layer theory and WKB theory from multiple-scale perturbation theory. Thus, multiple-scale perturbation theory may be regarded as the most general of all perturbation methods.

SEE ALSO THE FOLLOWING ARTICLES

AIRCRAFT AERODYNAMICS • BOUNDARY LAYERS • CELESTIAL MECHANICS • COMMUNICATION SATELLITE SYSTEMS • ELECTRON SPIN RESONANCE • GREEN'S FUNCTIONS • HYDROGEN BOND • WAVE PHENOMENA

BIBLIOGRAPHY

Baker, G. A., Jr. (1975). "Essentials of Padé Approximants," Academic Press, New York.

Baker, G. A., Jr., and Graves-Morris, P. (1996). "Padé Approximants," 2nd edition, Cambridge University Press, Cambridge.

Bender, C. M., Boettcher, S., and Milton, K. A. (1991). "A new perturbative approach to nonlinear partial differential equations," *J. Math. Phys.* **32,** 3031.

Bender, C. M., Milton, K. A., Pinsky, S. S., and Simmons, L. M., Jr. (1989). "A new perturbative approach to nonlinear problems," *J. Math. Phys.* **30,** 1447.

Bender, C. M., and Orszag, S. A. (2000). "Advanced Mathematical Methods for Scientists and Engineers," Springer, New York. (Originally published by McGraw-Hill, 1978.)

Bender, C. M., and Wu, T. T. (1969). *Phys. Rev.* **184,** 1231; *Phys. Rev. Lett.* **27,** 461 (1971); *Phys. Rev. D* **7,** 1620 (1973).

Berry, M. V., and Mount, K. E. (1972). "Semiclassical approximations in wave mechanics," *Rep. Prog. Phys.* **35,** 315–97.

Borel, E. (1928). "Leçons sur les Séries Divergentes," 2nd edition, Gautier-Villars, Paris, Reprinted by Éditions Jacques Gabay, Paris, 1988. Translated by C. L. Critchfield and A. Vakar, "Lectures on Divergent Series" (Los Alamos Scientific Laboratory, Translation LA-6140-TR, 1975).

Bowman, K. O., and Shenton, L. R. (1989). "Continued Fractions in Statistical Applications," Marcel Dekker, New York.

Brezinski, C. (1977). "Accélération de la Convergence en Analyse Numérique," Springer-Verlag, Berlin.

Brezinski, C. (1991). "A Bibliography on Continued Fractions, Padé Approximation, Extrapolation and Related Subjects," Prensas Universitarias de Zaragoza, Zaragoza.

Brezinski, C. (1991). "History of Continued Fractions and Padé Approximants," Springer-Verlag, Berlin.

Brezinski, C. (2000). *J. Comput. Appl. Math.* **122,** 1. Reprinted in (C. Brezinski, ed.), p. 1, "Numerical Analysis 2000 Vol. 2: Interpolation and Extrapolation," Elsevier, Amsterdam, 2000.

Cole, J. D. (1968). "Perturbation Methods in Applied Mathematics," Blaisdell, Waltham.

Dingle, R. B. (1973). "Asymptotic Expansions: Their Derivation and Interpretation," Academic, New York.

Gilewicz, J. (1973). In "Padé Approximants and Their Applications," (P. R. Graves-Morris ed.), p. 99, Academic Press, London.

Graves-Morris, P. R., Roberts, D. E., and Salam, A. (2000). *J. Comput. Appl. Math.* **122,** 51. Reprinted in (C. Brezinski, ed.), p. 51, "Numerical Analysis 2000 Vol. 2: Interpolation and Extrapolation," Elsevier, Amsterdam, 2000.

Hardy, G. (1949). "Divergent Series," Clarendon Press, Oxford.

Jones, W. B., and Thron, W. T. (1980). "Continued Fractions," Addison-Wesley, Reading, MA.

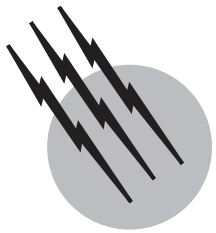
Kevorkian, J. (1981). "Perturbation Methods in Applied Mathematics," Springer-Verlag, New York.

Kevorkian, J., and Cole, J. D. (1996). "Multiple-Scale and Singular Perturbation Methods," Springer-Verlag, New York.

Lagerstrom, P. A. (1988). "Matched Asymptotic Expansions," Springer-Verlag, New York.

Le Guillou, J. C., and Zinn-Justin J. (eds.) (1990). "Large-Order Behaviour of Perturbation Theory," North Holland, Amsterdam.

Lorentzen, L., and Waadeland, H. (1992). "Continued Fractions with Applications," North-Holland, Amsterdam.
Nayfeh, A. H. (1973). "Perturbation Methods," Wiley, New York.
Nayfeh, A. H. (1981). "Introduction to Perturbation Techniques," Wiley, New York.
O'Malley, R. E. (1991). "Singular Perturbation Methods for Ordinary Differential Equations," Springer-Verlag, New York.
Pozzi, A. (1994). "Applications of Padé Approximation Theory in Fluid Dynamics," World Scientific, Singapore.
Reed, M., and Simon, B. (1972). "Methods of Modern Mathematical Physics," vols. I–IV, Academic Press, New York.
Shawyer, B., and Watson, B. (1994). "Borel's Method of Summability," Oxford U.P., Oxford.
Simon, B. (1970). *Ann. Phys.* **45,** 76.
Wall, H. S. (1973). "Analytic Theory of Continued Fractions," Chelsea, New York.



Probability

M. M. Rao
University of California, Riverside

I. Historical Evolution
II. Basic Notions of Modern Probability
III. Some Examples and Paradoxes
IV. Strong Law of Large Numbers and Its Uses
V. Conditional Probability and Dependence
VI. Stationary and Harmonizable Processes and Spectra
VII. Some Remarks on the Central Limit Problem
VIII. Statistical Inference Problems and Comments

GLOSSARY

Conditional expectation If X and Y are two random variables and if $B = [Y = b]$ is given and $P(B) > 0$, then the conditional expectation of X given $Y = b$ is the usual expectation defined below calculated with the conditional probability $P(\cdot \mid B)$ in place of the original probability.

Conditional probability If A and B are two events of positive probability, then the conditional probability of A given B is the ratio of the probability of the joint occurrence of A and B to that of just B and is denoted by $P(A \mid B)$. Difficulties arise if $P(B) = 0$.

Correlation Measure of relation (interdependence) between a pair of random variables.

Distribution Each numerical value of the random variable is assigned a probability, and a set of values has the sum of the probabilities. These values of the sets constitute the function that is sometimes called a (cumulative) distribution. All the values together receive probability 1.

Expectation Population average $E(\cdot)$ of the random variable over a distribution of its values with prescribed probabilities.

Independence A class of events is mutually independent if the occurrence (nonoccurrence) of any finite subset of them has no influence on the occurrence (nonoccurrence) of the others. A class of random variables is similarly independent if all the events they determine are mutually independent.

Random variable Rule to associate a numerical value for each point of the sample space.

Sample space Space of points representing all possible (distinct) outcomes of an experiment.

Stochastic process A collection of random variables, often indexed by integers as $X_1, X_2, \ldots$ (or by real numbers as $X_s, X_t, \ldots$). Also called a random process or time series when the indexing refers to the time at which these random variables can be observed.

Variance Measure of concentration of the values of the random variable about its expectation.

PROBABILITY theory is a branch of mathematics dealing with the analysis of the behavior of observable processes and experiments involving uncertainty. Its applications are to determine or predict the outcomes of experiments with a degree of assurance when numerical values are assigned to uncertainties. These applications include games of chance or gambling, fluctuations in economic behavior or stock markets, meteorology, gene combinations, biology, physics, and in fact most aspects of scientific activity, as well as to theoretical studies such as number theory and complex variables. Probability is the basis for statistical analysis and especially for inference theory. In each of the fields of application, first one has to formulate a model, analyze its theoretical implications, and then make the intended estimation or prediction to decide the suitability of the model by experimentation.

I. HISTORICAL EVOLUTION

The fact that the outcome of an experiment cannot always be predicted with certainty has been recognized since ancient times, and in most cases previous experience was used to gauge the amount of uncertainty. This numerical assignment of uncertainty is an art that has been developed to various degrees, and its measurement, in a refined form, is known as probability. For a proper appreciation of the topic, a somewhat detailed historical sketch will be presented in this section.

There was always a fascination among the ancients for guessing the unknown, on the basis of some accumulated knowledge, by observation and then verifying or modifying the guesses with further experimentation. This was especially true with games of chance, particularly dicing, which was often termed a vice but which was never given up. Indeed, gambling was well developed by the Greeks, and Aristotle condemned it as equivalent to robbery. The great Indian epic, the *Mahābhārata*, deals with the conflict growing out of a refined dicing game. The uncertainty was later analyzed in great detail, for instance, in the Jaina philosophy. It is of interest here to discuss this development briefly because it anticipates the logical foundations of our subject.

The Jaina philosopher Bhadrabāhu (ca. 433–357 B.C.) presents in one of his writings a highly developed system that had been in existence for more than a century before him, involving a classification of uncertainties into seven categories. This is termed the *syādvāda* system having *saptabhaṅginā ya* (seven types of analysis of prediction). The original (Sanskrit) classification is as follows:

1. *syādasti* (*syād* = "perhaps," *asti* = "exists"; "Perhaps it exists.")
2. *syātnāsti* ("Perhaps it is not.")
3. *syādastinā stica* ("Perhaps it is and is not.")
4. *syādavaktavyah* ("Perhaps it is indeterminate.")
5. *syādastica avaktavyaśca* ("Perhaps it is and also indeterminate.")
6. *syādatnāstica avaktavyaśca* ("Perhaps it is not and also indeterminate.")
7. *syādasti nāstica avaktavyaśca* ("Perhaps it is and it is not and also indeterminate.")

It can be shown that these divisions exhaust the possibilities, and uncertainties may always be treated as belonging to exactly one of them. These ideas seem to have been applied to some practical problems. For instance, there are references in medieval Hindu books to the practice of giving alms to religious mendicants without ascertaining whether or not they were deserving. The question was resolved in favor of continuing the practice since on observation it was found that "only ten out of a hundred were undeserving." The implication here, in present-day terminology, is that the probability is 0.9 that a deserving person receives alms. This system was discussed by the late P. C. Mahalanobis in 1954 with source material in *Dialectica* **8,** 95–111.

Gambling as well as the probabilistic aspects of trade and commerce appear to have provided an even stronger motivation to develop the subject during the Renaissance period. By about the middle of the 16th century the first combinatorial aspect of the subject was systematically described by G. Cardano (1501–1576) in a book on gambling. The binomial coefficient formulas were already known by the 12th and 13th centuries, so that they could be used. Thus, in his work, Cardano states that the probability of an event in repeated throws of an ideal die, or a pair of dice, is the ratio of the favorable outcomes to the possible ones. It will be illuminating to discuss some classical examples since this period may be regarded as the start of a new epoch in the development of probability, particularly regarding its combinatorial aspects.

If a die is tossed once, Cardano takes the probability of any one number of its six faces to be $\frac{1}{6}$, since for the ancients dice and coins must always be fair to play. Similarly, if f is the favorable and n the total number of outcomes, he takes the probability $p_1 = f/n$ for a toss; in two repeated tosses he states that the odds are f^2 to $n^2 - f^2$, or $p_2 = f^2/n^2$; and in k tosses he obtains the *product formula* $p_k = f^k/n^k$. Cardano also considers the average proportion in m tosses, if p is the probability of a single toss, as mp. If the experiment is repeated until a desired outcome appears, then he says that for an even chance the needed number m of trials is determined from $\frac{1}{2} = mp$. Thus, for a fair die, $p_1 = \frac{1}{6}$, so that $m = 3$, and if the experiment involves two dice, then $p_2 = \frac{1}{36}$, so that $\frac{1}{2} = mp_2$

gives $m = 18$. However, Cardano realizes that this application is not correct for large m since $mp > 1$ can happen and concludes that "this is most absurd." He could not solve this problem himself.

A very similar question was considered about a century later by the French nobleman and gambler Cavalier de Méré as follows. If p_1, p_2 are as above, then de Méré notes that $4p_1 = \frac{4}{6} = 24p_2 = \frac{24}{36}$ and concludes that the probability of at least a six in the first play should be the same as at least a double six in the second one. Both have an "equal chance." With experimentation, he found that the odds favor the first experiment and lost money in the betting. He wrote to B. Pascal (1623–1662) about his dilemma, and the latter in correspondence with P. Fermat (1601–1665) showed in 1654 that a correct computation validates the experimental evidence, truly reflecting the theory. Namely, the probability of not getting a six in four tosses is $(\frac{5}{6})^4$ (by Cardano's power law) and so at least one six has the probability $= 1 - (\frac{5}{6})^4 = 0.5177$, whereas a similar computation with the second play gives the correct probability $= 1 - (\frac{35}{36})^{24} = 0.4914 < 0.5177$. This simple calculation by Fermat and Pascal a century later gave further stimulus for considering sophisticated combinatorial probability problems.

It should be noted that a similar, but slightly easier problem was also posed to Galileo Galilei (1564–1642) by one of his friends about a half-century after Cardano raised it. If three fair dice are thrown, and one desires to get the sum of 9 in one play and 10 in the other, it was thought that the probabilities should be the same since the number of distinct ways 9 and 10 can be obtained is exactly six ($9 = 6+2+1 = 5+3+1 = 5+2+2 = 4+4+1 = 4+3+2 = 3+3+3$, and $10 = 6+3+1 = 6+2+2 = 5+4+1 = 5+3+2 = 4+4+2 = 4+3+3$). But the experiments favored 10 and not 9. Galileo showed that, whereas $3+3+3 = 9$ can occur in only one way, in the second configuration 10 can be obtained in more ways and calculated the respective probabilities for 9 and 10 to be $\frac{25}{216}$ and $\frac{27}{216}$.

These examples show the problems that were beyond the techniques of the ancients and motivated the further development of probability. The Cardano–de Méré problem was satisfactorily solved by Pascal and Fermat, but this was done independently of Cardano's important work "De Ludo Aleae" ("Games of Chance"), which was published almost a century after the author's death. Cardano was a well-known physician in Milan, but he wrote important books on mathematics ("Ars Magna" in 1545 is perhaps the most important) as well as medicine. Thus, from the ancients the next important advance should be attributed to Cardano and later independently to Galileo and Fermat–Pascal. Until recently Cardano's work was not properly appreciated since it was either misunderstood or termed unintelligible because of his Latin style. It took a contemporary mathematician of the caliber of Oystein Ore of Yale University to show that everything in Cardano's book is essentially correct and that his contributions to probability are important. (Thus, the history of science is well served if the historian is also accomplished in the subject.)

Here, one should note that Cardano briefly considered certain approximations to estimating an unknown probability and thought that, for a large number of repeated trials, the proportion of favorable to possible outcomes should stabilize. This is a form of the law of large numbers, but he did not have the tools to establish it. A similar observation was also found in an early 16th-century Sanskrit manuscript of an Indian writer (named Brahmagupta), saying that the precision of the arithmetic mean increases with the number of observations. The contemporary Russian mathematical sciences historian L. E. Maistrov finds this to be "a remarkable fact" because of the early date of its formulation.

A mathematically rigorous proof of the law of large numbers was established for the zero–one-valued random variables on repeated trials, with a careful enumeration, by James Bernoulli (1654–1705) at the end of the 17th century, and it appeared in his book "Ars Conjectandi," published posthumously in 1713. This was the next significant step in the evolution of probability. The work was systematically elaborated and extended by A. de Moivre (1667–1754) and P. S. Laplace (1749–1827) in the 18th and the 19th centuries. It involved both an extension of and new ways of looking into what became known as the central limit problem, still keeping the two-valued, or Bernoulli, random variables. Precise statements of laws of large numbers are given in Section IV.

Independently repeatable experiments having the same probability for each outcome, described as "success" or "failure," were the object of Bernoulli's work, and these are called the Bernoulli trials. One denotes by X_n a two-valued function taking 1 for "success" and 0 for "failure" on the nth trial. Thus, $S_n = X_1 + \cdots + X_n$ denotes the number of successes in n Bernoulli trials. The proportion S_n/n, which changes from one experiment to another, should "stabilize" as n increases. Let $P[X_n = 1] = p$ and $P[X_n = 0] = q = 1 - p$. If $p = q = \frac{1}{2}$, then one has a fair game. Bernoulli's theorem is that for each $\varepsilon > 0$, the event $\{|(S_n/n) - p| < \varepsilon\}$ has probability arbitrarily close to 1, if n is large enough. This is a statement about the behavior of the ratio S_n/n in *different sets* of games and *not* on the fluctuations of the proportions in a *single* game of trials. The latter is a stronger statement.

Before the work of Laplace, most calculations consisted of Bernoulli trials with $p = \frac{1}{2}$. By the time that de Moivre and especially Laplace started work on the

subject, the Leibniz–Newton theory of calculus was already available, so that a finer analysis of probabilities of the events discussed above was possible. But the Laplace–de Moivre result goes farther, and the first real limit statement, now called the central limit problem, was formulated. Thus, they were able to prove the first central limit theorem as

$$P\left[a < \frac{S_n - np}{\sqrt{npq}} < b\right] \sim \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx \qquad (1)$$

for large enough n, where $0 < p = 1 - q < 1$. The functions ϕ and Φ, where $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$ and $\Phi(x) = \int_{-\infty}^{x} \phi(t)\, dt$, appear for the first time in the works of de Moivre and Laplace, though the latter author used it more effectively in his analyses. Later C.-F. Gauss (1777–1855) also derived the same function but published his work in 1809 and noted its appearance in many different contexts. A lesser known mathematician in the United States, Robert Adrian (1775–1843), independently derived and published the function in 1808 in an obscure journal. Many people in France term Φ the Laplace distribution; in other places it is called the Gaussian distribution. In English publications, it is usually called a normal distribution. As a limit law, probabilists found it so common in physical applications that H. Poincaré (attributing it to Lippman) said that "there is something mysterious about this distribution since mathematicians think it is a law of nature and physicists are convinced that it is a mathematical theorem." It is generally regarded as the basic limit law of probability theory.

By the end of the 19th century random variables more complicated than those of Bernoulli appeared. Keeping the very appealing concept of mutual independence in mind, a generalization of the Bernoulli law of large numbers and the Laplace–de Moivre central limit theorem were obtained. Important research of P. L. Čebyšev (1821–1894), A. A. Markov (1856–1922), and A. M. Liapounov (1857–1918) contributed to this area. In another direction, E. Borel (1871–1956) proved a far stronger form of the law of large numbers for Bernoulli's random variables in 1909. He showed that $S_n/n \to p$ for almost all individual games, the exceptions constituting only zero probability. Thus, in the current language, for each $\varepsilon > 0$ and $\delta > 0$, one can find an $n_0[= n_0(\varepsilon, \delta)]$ such that for any $m \geq 1$, it is true that

$$P\left[\left|\frac{S_{n_0+k}}{n_0 + k} - p\right| < \varepsilon, k = 1, 2, \ldots, m\right] > 1 - \delta \qquad (2)$$

This is clearly a more precise result; it is called the *strong law of large numbers* and finally gives a frequency interpretation that justifies the intuitive feelings of the ancients. But these results only heightened the quest for a precise definition of probability, since the long-term average or the limiting frequency is not mathematically satisfactory. Developments of mathematical rigor by the end of the 19th century reached a stage at which an axiomatic basis of probability became necessary for it to be an accepted branch of mathematics. A respected mathematician, R. von Mises (1883–1953), worked on this problem intensely and proposed using a certain "collective" as a primitive notion to develop the frequency concept rigorously. However, this turned out to have serious defects. In despair, von Mises said in 1922 that "today probability theory is not a mathematical science." During that period a large number of important individual probabilistic limit theorems were obtained, although an acceptable axiomatic formulation was still missing. Fortunately, this was supplied by A. N. Kolmogorov in 1933, and his formulation is now universally accepted. Since then the subject has grown rapidly utilizing the abstract set functions and Lebesgue's integration. However, there are still problems, and we have no satisfactory agreement on the *interpretation* of all results. These and the salient features of modern probability will be discussed in the following sections.

It must be noted that, with the explosive growth of science in general and mathematics in particular during the last part of the 19th and early 20th centuries, a consolidation of the subject became imperative. Using the precise ideas of logic, some mathematicians tried to mold everything in the two-valued formulation so that each statement would be either true or false and the uncertainty eliminated. (Thus, according to such people, in the modern world, the *syādvāda* system could be retired.) It was soon realized, however, that this is not always the case. A simple illustration, usually attributed to H. Weyl, is as follows. Prove or disprove the statement: "If the irrational number π is expanded in decimals, then the numbers 0,1,2,3,4,5,6,7,8,9 will occur somewhere in their natural order." The truth of this assertion could not be determined for nearly a century, although π has been expanded for many millions of decimals, but probabilistically it can be ascertained quickly and positively. Only recently, in June and July of 1997, two Japanese computer scientists (Yasumasa Kanada and Daisuke Takahashi, at the University of Tokyo) using a massively parallel Hitachi machine with 2^{10} processors, have expanded and discovered the first occurrence of the above noted sequence after over 17 billion decimals (actually at 17, 387, 594, 880 th digit after the decimal point). The next five occurrences of the natural sequence are found respectively at 26, 852, 899, 245; 30, 243, 957, 439; 34, 549, 153, 953; 41, 952, 536, 161; and 43, 289, 964, 000 places. An interesting discussion of this matter is given by the Canadian number theorist J. M. Borwein (1998) where the difficulties of such a computation and precautions taken have been sketched. Thus

even from a pragmatic point of view, it is useful to have at hand, the Theory of Probability firmly founded and developed. This is precisely the purpose of the subject, and its development is proceeding at a fast pace.

II. BASIC NOTIONS OF MODERN PROBABILITY

An experiment, such as tossing a die or recording the test scores of students, is described by a probability model. All possible outcomes are represented by the distinct points of an abstract set Ω. Different combinations of these points, being subsets of Ω, are recognized as events that are of interest to the experimenter. Let these events be designated by $\mathcal{A}$. If Ω is finite, or at most countable, then $\mathcal{A}$ is taken to be the power set (the class of all subsets) of Ω. However, if Ω represents such experiments as the heights or weights of individuals or the fluctuations of an electric current, then Ω has uncountably many points and $\mathcal{A}$ has to be chosen more carefully, since taking it as a power set creates technical difficulties, for it includes many useless "events." Following the countable case and utilizing the experience of related mathematical constructs, one takes $\mathcal{A}$ to consist of all those sets of interest for the experiment and their combinations (unions, intersections, and differences) so as to create an algebraic structure. Thus, $\mathcal{A}$ will be a σ-algebra of subsets of Ω, meaning it is closed under countable unions and differences. Then it is also closed under countable intersections, and *only members of $\mathcal{A}$ will be called events*. This generally will be smaller than the power set and is the smallest class closed under the above countable operations containing all the basic collection of sets that are of interest to the experimenter. The problem here is to select the basic aggregate such that for each of its members the experimenter can prescribe a probability in a consistent manner. Let us give an example to clarify this discussion. If Ω corresponds to tossing a pair of dice, then it has 36 points representing each combination $\{(i, j): 1 \le i, j \le 6\}$, and if the dice are perfect, then each point is given the same probability, $\frac{1}{36}$. Here the consistency implies that the probabilities add on disjoint sets and the whole space gets unit weight. These values are given to the basic collection, that is, to each point in this case. The aim of the general study is that, if the experimenter can assign probabilities on the basic collection, it is possible to calculate probabilities of *all* events obtained by combining these fundamental sets with countable operations. Thus, a standard result of the subject is that it is possible to extend the probability P uniquely from the basic collection to all of $\mathcal{A}$, the σ-algebra determined by that collection. The triple $(\Omega, \mathcal{A}, P)$ is the model in Kolmogorov's setup that describes a given experiment completely.

With such a model in hand, random variables are defined as numerical functions on Ω, often called a sample space, to reflect the events relevant to the experiment. This means that, if $X : \Omega \to \mathbb{R}$, reals, $\{\omega \in \Omega: X(\omega) \in (a, b)\} \in \mathcal{A}$ (read "$\omega \in \Omega$" as ω in Ω); so $\{\ldots\}$ is an event of interest for *each* interval (a, b) of the range $\mathbb{R}$ of X. The condition is automatic if Ω is countable, since then $\mathcal{A}$ is the power set and even the weaker requirement that $\{\omega \in \Omega: X(\omega) = r\}$ be in $\mathcal{A}$, for each r in $\mathbb{R}$, suffices. Such a mapping X is a random variable, and one can calculate probabilities for all events determined by it for each interval (a, b). Thus, the basic elements of chance inherent in the model are carried over by this mapping to each interval, and hence X is called a random or chance variable. **When the basic model $(\Omega, \mathcal{A}, P)$ is not formulated carefully, people have difficulty in defining precisely what a chance variable is, and the resulting confusion sometimes leads to errors and paradoxes**. Indeed, without such a model a probability statement can have little significance.

Why does one want random variables when the basic probability model is already in hand? There are at least two reasons: (1) in order to achieve a finer mathematical analysis of the model and (2) to avoid the difficulties encountered by the ancients (e.g., the St. Petersburg paradox; see Section IV). Also, the analysis benefits by the inclusion of certain powerful techniques of calculus (e.g., Fourier transforms). These are discussed here since they aid in understanding some later developments. Thus, if $(\Omega, \mathcal{A}, P)$ is the model governing an experiment and $X: \Omega \to \mathbb{R}$ is a random variable, then with each such X, one can associate a real function F_X, called the distribution function of X, by the equation

$$F_X(x) = P[\{\omega: X(\omega) < x\}], \qquad -\infty < x < \infty \tag{3}$$

This is well defined since $\{\omega: X(\omega) < x\}$ is an event of $\mathcal{A}$, by definition of a random variable, and clearly $F_X(-\infty) = 0$, $F_X(x_1) \le F_X(x_2)$ if $x_1 \le x_2$, and $F_X(+\infty) = 1$, $\lim_{h \searrow 0} F_X(x - h) = F_X(x)$. For instance, if Ω is $\{H, T\}$ representing one toss of a coin, $\mathcal{A}$ being its power set, $P(\{H\}) = p$, $P(\{T\}) = 1 - p = q$, then $(\Omega, \mathcal{A}, P)$ is the underlying model. Let $X(H) = 1$, $X(T) = 0$; then X is called a Bernoulli random variable, and the F_X, called the Bernoulli distribution, is

$$F_X(x) = 0, \quad x \le 0; \quad F_X(x) = q, \quad 0 < x \le 1;$$
$$F_X(x) = 1, \quad x > 1$$

The following are some other examples of frequent appearance:

1. Let Ω be the set of points corresponding to all possible outcomes of tossing a coin n times, so that there are 2^n points in Ω. Taking $\mathcal{A}$ to be the power set of

Ω, define $S_n: \Omega \to \mathbb{R}$ as $S_n(\omega) =$ number of heads in $\omega \in \Omega$. Thus, $S_n(\omega) = k$ means that ω has k heads and $n-k$ tails in its representation. Then S_n is a random variable and F_{S_n}, called the binomial distribution first derived by J. Bernoulli, is given (with probability of heads $= p = 1-q$) by $F_{S_n}(x) = \sum_{k=0}^{[x]} f_{k,n}(p)$, where

$$f_{k,n}(p) = \begin{cases} 0, & k < 0 \\ \binom{n}{k} p^k q^{n-k}, & 0 \le k \le n \\ 0, & k > n \end{cases}$$

Here $\binom{n}{k}$ is the binomial coefficient $n!/k!(n-k)!$. If $n = 1$, this is the Bernoulli distribution F_X ($[x]$ is the integer part of x).

2. If $\Omega = \mathbb{R}$, $\mathcal{A} = \sigma$-algebra determined by all intervals of $\mathbb{R}$, and P is a probability on $\mathcal{A}$, $X: \Omega \to \mathbb{R}$, defined by $X(\omega) = \omega$ is a random variable, then $F_X(x) = P[X < x]$ given by

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt, \quad -\infty < x < \infty$$

is the normal distribution, already encountered in the last section.

3. If Ω, $\mathcal{A}$, and X are as in Example 2, then P or equivalently F_X given by the equation

$$F_X(x) = \frac{1}{\pi} \int_{-\infty}^{x} \frac{dt}{1+t^2}, \quad -\infty < x < \infty$$

is called the Cauchy distribution.

4. Let Ω be the set of integers, $\mathcal{A}$ the power set, and $X(\omega) = \omega$, $\omega \in \Omega$; also let $P(\{k\}) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, 2, \ldots$, and $= 0$ for other k. Then F_X is given by

$$F_X(x) = \begin{cases} 0, & x \le 0 \quad \text{or} \quad \lambda \le 0 \\ \displaystyle\sum_{0<k\le[x]} e^{-\lambda} \frac{\lambda^k}{k!}, & x > 0 \quad \text{and} \quad \lambda > 0 \end{cases}$$

This is called the Poisson distribution. It arises in the following simple manner from example 1 (and also in many other ways). If a coin is tossed repeatedly, the probability p of its landing heads changes slowly after each toss (due to impact on the floor or defective metal) but such that np is approximately a constant ($= \lambda$, say). Then the binomial probability changes to the above Poisson law. In fact from Example 1, $f_{0,n}(p) = q^n = (1-p)^n = (1-\lambda/n)^n \to e^{-\lambda}$, as $n \to \infty$ by a standard calculus result. But also

$$\frac{f_{k,n}(p)}{f_{k-1,n}(p)} = \frac{(n-k+1)p}{kq} = \frac{\lambda - [(k-1)/n]\lambda}{k(1-\lambda/n)} \to \frac{\lambda}{k} \quad \text{as} \quad n \to \infty$$

$k = 1, 2, \ldots$, so that by induction one has as $n \to \infty$:

$$f_{1,n}(p) \to \lambda e^{-\lambda}, \quad f_{2,n}(p) \to \frac{\lambda^2 e^{-\lambda}}{2!}, \ldots,$$

$$f_{k,n}(p) \to \frac{\lambda^k e^{-\lambda}}{k!}.$$

This distribution was invented by S. D. Poisson (1781–1840) in his investigations of civil and criminal matters in 1837. The same distribution was later found to arise in experiments involving radioactive emissions, telephone traffic, and so on. It can be used as a good approximation for binomial probabilities if p is small and n relatively large.

In a similar manner, Cauchy's distribution is useful in target-shooting and destruction problems and of course is the normal distribution in many areas, as noted in the introduction. One calls the random variables by the same names as their distributions. Thus, there are Bernoulli, binomial, normal, Cauchy, Poisson, and other random variables if their distributions have those names.

It is now clear how other distribution functions relating to possible games and experiments can be introduced. In fact with each increasing F, $0 = F(-\infty) \le F(x) \nearrow F(+\infty) = 1$, we can construct a probability model and a random variable on it having the given function as its distribution. Namely, one can take $\Omega = \mathbb{R}$, $\mathcal{A} =$ the σ-algebra determined by all the intervals, $X(\omega) = \omega \in \Omega$, and $P(A) = \int_A dF$, $A \in \mathcal{A}$. Then $(\Omega, \mathcal{A}, P)$ is a model, and $F = F_X$ is the distribution of X. This simple case can be extended to a family of distributions if they satisfy a natural compatibility condition. Such a result was established in 1933 by A. Kolmogorov when he presented the modern axiomatic setup giving a satisfactory basis for a rigorous theory of probability. We now discuss some intrinsic properties of this abstract model after introducing the fundamental concept of stochastic (or statistical) independence, which distinguishes this subject from other parts of mathematics and which was already used by the ancients.

In $(\Omega, \mathcal{A}, P)$, two events A, B are independent if the occurrence of one has no influence on the other, and hence the uncertainty of the joint occurrence must be greater than the individual uncertainties. This is now translated mathematically into the formula

$$P(A \text{ and } B) = P(A)P(B) \tag{4}$$

Taking $A = B$, this implies that A is independent of itself if and only if $P(A) = 0$ or 1, so that A is either an impossible or an absolutely certain event. The above concept for n events $A_1, \ldots, A_n$ must be applied with a little care. Thus, the events are *mutually independent* if for each $1 < m \le n$ the following statement holds for all $1 \le i_1 < i_2 < \cdots < i_m \le n$,

$$P\left(A_{i_1} \text{ and } A_{i_2} \ldots \text{and } A_{i_m}\right) = \prod_{k=1}^{m} P\left(A_{i_k}\right) \tag{5}$$

so that $2^n - n - 1$ such product equations must hold. If only every pair of n events satisfies Eq. (4), then there will be $\binom{n}{2}$ equations, and such a collection is called *pairwise independent*. Both the concepts are distinct generalizations of Cardano's power formula, and Eq. (5) is the most stringent. For random variables, these conditions take the following form. Let $X_1, \ldots, X_n$ be n random variables on $(\Omega, \mathcal{A}, P)$. Then they are said to be pairwise or mutually independent accordingly as for each pair X_i, X_j

$$\begin{aligned} &P[\omega\colon a_i < X_i(\omega) < b_i \quad \text{and} \quad c_j < X_j(\omega) < d_j] \\ &\quad = P[\omega\colon a_i < X_i(\omega) < b_j] \\ &\qquad \times P[\omega\colon c_j < X_j(\omega) < d_j] \end{aligned} \tag{6}$$

or for each $1 < m \le n$, $1 \le i_1 < i_2 < \cdots < i_m \le n$,

$$\begin{aligned} &P[\omega\colon a_j < X_{i_j}(\omega) < b_j, \quad j = 1, \ldots, m] \\ &\quad = \prod_{j=1}^{m} P\left[\omega\colon a_j < X_{i_j}(\omega) < b_j\right] \end{aligned} \tag{7}$$

for all real numbers $a_i < b_i, c_j < d_j$. Similarly, if $X_1, X_2, \ldots,$ is an infinite sequence of random variables, then it is called a mutually or pairwise independent collection if each of its finite subcollections satisfies Eq. (7) or (6). Sometimes a reader may wonder whether independent random variables always exist in sufficient numbers. This is a legitimate concern, but by the repetition of an experiment $(\Omega, \mathcal{A}, P)$ may be taken to be rich enough to support such collections of independent variables. This is merely a technical matter with a relatively easy solution. Thus, all these concepts are meaningful, and a nontrivial theory is built around the notion of independence.

To state these concepts with distribution functions, let $X_1, \ldots, X_n$ be random variables on $(\Omega, \mathcal{A}, P)$. Their joint (or multivariate) distribution is given by

$$\begin{aligned} &F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \\ &\quad = P[\omega\colon X_1(\omega) < x_1, \ldots, X_n(\omega) < x_n] \end{aligned} \tag{8}$$

and this is an extension of Eq. (3). If $x_2 = +\infty, \ldots, x_n = +\infty$, then the event in the right side becomes $[\omega\colon X_1(\omega) < x_1]$ and hence $F_{X_1}(x_1) = F_{X_1,\ldots,X_n}(x_1, \infty, \ldots, \infty)$ and F_{X_1} is called the marginal distribution. Similarly, F_{X_i} is defined. As an example, let $(\Omega, \mathcal{A}, P)$ be a model for tossing a die n times. Then $\Omega = A^n$, where $A = \{1, 2, \ldots, 6\}$, $\mathcal{A}$ is the power set, and $\omega \in \Omega$ is a point with n coordinates, each component being one of the integers $1, \ldots, 6$. If $X_i(\omega) = \omega(i) = i$th component of ω, then $X_i(\omega) \in A$, and X_i is a random variable. (Assuming the die to be perfect, we take $P[\omega\colon \omega(i) = 1] = (\frac{1}{6})^n$.) In the general case, $X_1, \ldots, X_n$ are mutually independent if their joint distribution is a product of the marginals:

$$\begin{aligned} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= \prod_{i=1}^{n} F_{X_i}(x_i) \\ &\text{for all} \quad -\infty < x_i < \infty \end{aligned} \tag{9}$$

It can be shown that Eqs. (7) and (9) are equivalent conditions. In case all $F_{X_i} = F$, the X_i are called identically distributed. Thus, for Bernoulli trials, $X_1, \ldots, X_n$ are independent and identically distributed $\{0, 1\}$-valued random variables.

For any random variable X on $(\Omega, \mathcal{A}, P)$, the kth moment is defined for any $k \ge 0$ as

$$E(X^k) = \int_{\mathbb{R}} x^k \, dF(x) \quad \left(= \int_{\Omega} X^k \, dP\right) \tag{10}$$

provided that $\int_{\mathbb{R}} |x|^k dF(x) < \infty$. If $k = 1$, $E(X)$ is the expectation (or the population mean or average) of X. The actual definition is the integral in parentheses, but its equivalence with that of the distribution F follows from Eq. (3), and this relation is termed *the fundamental law of probability*. If $Y = X - E(X)$, then Y is centered and $E(Y^2)$ is called the variance of X, denoted $\operatorname{Var} X = \sigma_x^2$, and its positive square root σ_x is termed the *standard deviation* of X. These serve a purpose analogous to the center of gravity and moment of inertia in mechanics. For the distributions introduced above, one has the following:

Distribution	Mean	Variance	Parameter
Bernoulli	p	$p(1-p)$	$0 < p < 1$
Binomial	np	$np(1-p)$	$0 < p < 1$
Poisson	λ	λ	$\lambda > 0$
Standard normal	0	1	
Cauchy	Does not exist	Does not exist	

With these ideas in hand, J. Bernoulli's law of large numbers can be given for random variables taking more than two values. His combinatorial enumeration is no longer possible. P. L. Čebyšev was the first to obtain a generalization for which he devised a simple but powerful inequality:

A. Proposition 1 (Čebyšev's Inequality)

Let X be a random variable on $(\Omega, \mathcal{A}, P)$ with mean μ and variance σ^2. Then for any $\varepsilon > 0$,

$$P[\omega\colon |X(\omega) - \mu| > \varepsilon] \le \sigma^2/\varepsilon^2 \tag{11}$$

B. Proposition 2

Let $X_1, X_2, \ldots$ be pairwise-independent random variables on $(\Omega, \mathcal{A}, P)$ with means $\mu_1, \mu_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$ such that $(1/n^2)\sum_{i=1}^n \sigma_i^2 \to 0$ as $n \to \infty$. If $S_n = \sum_{i=1}^n X_i$, then

$$P\left[\omega\colon \left|\frac{S_n(\omega) - E(S_n)}{n}\right| \geq \varepsilon\right] \to 0 \quad \text{as } n \to \infty \quad (12)$$

for each $\varepsilon > 0$; that is, the sequence obeys the law of large numbers. In particular if the X_n sequence is independent and identically distributed with mean μ and variance σ^2, it always obeys the law of large numbers.

Note that, in the Bernoulli case, the X_i have a common distribution with mean p and variance $p(1-p)$, so that $(1/n^2)\sum_{i=1}^n \sigma_i^2 = p(1-p)/n \to 0$ as $n \to 0$. Thus, Bernoulli's theorem is included in Čebyšev's result. If more generally all the X_i have a common distribution, then Var $X_i = \sigma^2$ so $(1/n^2)\sum_{i=1}^n \sigma_i^2 = \sigma^2/n \to 0$, so that Eq. (12) holds, and the second part follows. Since the demonstration is simple and instructive, it will be included here.

1. Proof of 1

The event $[|X - \mu| > \varepsilon]$ is the same as $[|X-\mu|^2 > \varepsilon^2]$. Hence, using Eq. (10) and setting $Y = X - \mu$, we obtain

$$\begin{aligned}\sigma^2 = E(Y^2) &= \int_\Omega Y^2\, dP \\ &= \int_{[|Y|>\varepsilon]} Y^2\, dP + \int_{[|Y|\leq\varepsilon]} Y^2\, dP \\ &\geq \varepsilon^2 \int_{[|Y|>\varepsilon]} dP = \varepsilon^2 P[|Y| > \varepsilon].\end{aligned}$$

This is the usual integral which is replaced by a sum in the discrete case. Thus, inequality (11) is established.

2. Proof of 2

Letting $Y_i = X_i - \mu_i$, note that the pairwise independence of X_i implies the same of Y_i so that

$$\text{Var } S_n = E\left(|S_n - E(S_n)|^2\right) = E\left(\left|\sum_{i=1}^n Y_i\right|^2\right)$$

$$= \sum_{i=1}^n E\left(Y_i^2\right) + \sum_{i\neq j} E(Y_i Y_j)$$

since expectation of the sum is the sum of the expectations by linearity of integral in Eq. (10),

$$\text{Var } S_n = \sum_{i=1}^n \sigma_i^2 + \sum_{i\neq j} E(Y_i)E(Y_j) = \sum_{i=1}^n \sigma_i^2$$

since $E(Y_i) = 0$. Here the pairwise independence is used to deduce, by Eqs. (9) and (10),

$$\begin{aligned}E(Y_i Y_j) &= \iint_{\mathbb{R}^2} y_i y_j\, dF_{yi}\, dF_{Y_i} \\ &= \int_{\mathbb{R}} y_i\, dF_{Y_i} \int_{\mathbb{R}} y_j\, dF_{Y_j} \\ &= E(Y_i)E(Y_j) = 0\end{aligned}$$

Hence by Eq. (11), applied to S_n, one has

$$\begin{aligned}&P[\omega\colon |S_n(\omega) - E(S_n)| > n\varepsilon] \\ &\leq \frac{\text{Var} S_n}{n^2\varepsilon^2} = \frac{1}{\varepsilon^2 n^2}\sum_{i=1}^n \sigma_i^2 \to 0\end{aligned}$$

as $n \to \infty$, by hypothesis. So Statement 2 holds.

In the previously mentioned result, one needs only the condition that $E(Y_i Y_j) = 0$, $i \neq j$, i.e., $\{X_i, or Y_i\}_1^\infty$ is a sequence of *uncorrelated* random variables, instead of (pairwise) independence. The former is weaker than the latter condition, but the contrast is not quite evident. This will now be explained, since it also has a deep and beautiful mathematical significance. The conclusion is through Fourier transforms, plus some properties of complex functions. Thus a pair of random variables X, Y are independent if and only if their joint and individual transforms, called traditionally here characteristic functions (ch.f.s), denoted $\varphi_X(t) = E(e^{itX})$ for X, and similarly for the others, satisfy:

$$\varphi_{X,Y}(t_1, t_2) = \varphi_X(t_1)\varphi_Y(t_2), \forall t_1, t_2 \in \mathbb{R}. \quad (13)$$

Now if X and $\tilde{X}$ have the same distribution, written $X \cong \tilde{X}$, then they will have the same ch.f., and if also $Y \cong \tilde{Y}$, from (13) one gets for any real $t_1, t_2 \in \mathbb{R}$, $t_1 X + t_2 Y \cong \tilde{X} + t_2\tilde{Y}$. Since this is true for all t_1, t_2 it means that the independence of X and Y is *equivalent* to having the same distribution of $t_1 X + t_2 Y$ and $t_1\tilde{X} + t_2\tilde{Y}$ for any $t_1, t_2 \in \mathbb{R}$. On the other hand a well-known and a good sufficient moment condition for a unique dtermination of the distribution of a random variable Z is:

$$\sum_{k=1}^\infty \frac{1}{(\alpha_{2k})^{\frac{1}{2k}}} = \infty,$$

where $\alpha_k = E(|Z|^k) < \infty$. These two properties can be combined to obtain a pleasant purely analytical condition explaining the equivalence of independence and a strengthened uncorrelatedness condition as follows.

Let $Q_1(X, a)$, $Q_2(Y, b)$ be a pair of real (measurable) functions of X and Y, depending also on some parameters $a \in A \subset \mathbb{R}^s$ and $b \in B \subset \mathbb{R}^{s'}$ where A, B are closed bounded parallelepipeds in the real Euclidean spaces. For instance, Q_1, Q_2 may be taken as some polynomials with a, b as coefficients. Let φ_m, ψ_n be continuous scalar functions on $\mathbb{R}$ for each m, n, to be specified below. Consider their expectations on the basic probability space:

$$E(\varphi_m(Q_1(X,a))) = \alpha_m(a); \quad E(\psi_n(Q_2(Y,b))) = \beta_n(b),$$

and the cross-product moments

$$E(\varphi_m(Q_1(X,a))\psi_n(Q_2(Y,b))) = \gamma_{m,n}(a,b), \quad m,n = 1,2,\ldots.$$

The function sequences φ_m, ψ_n are chosen to determine the distributions of $Q_1(X,a)$ and $Q_2(Y,b)$ uniquely. This is a substitute of the traditional moment condition recalled above. Finally suppose that there are simply connected bounded closed cylinders A_1, B_1 such that $A \subset A_1$, $B \subset B_1$ so that the moment functions α_m, β_n and $\gamma_{m,n}$ are holomorphic (of analytic) in A_1, B_1 and the product set $A_1 \times B_1$, respectively. This analyticity property of the first and second moments translates the independence hypothesis from being an abstract illusive concept to a property of (complex) function theory in the following manner.

In view of the holomorphy of the functions φ_m, ψ_n they determine the distributions of $Q_1(X,a)$ and $Q_2(Y,b)$ uniquely. Now suppose that the analytic function sequences α_m, β_n, $\gamma_{m,n}$ in the closed bounded polycylinders A_1, B_1 and $A_1 \times B_1$ satisfy the uncorrelatedness condition, namely:

$$\gamma_{m,n}(a,b) = \alpha_m(a)\beta_n(b), \quad a \in A_1, \quad b \in B_1,$$

for a finite set N of indexes, $1 \le m,n \le N < \infty$, where N depends only on the basic space on which X, Y are given. The conclusion is that *this finite set of relations on uncorrelatedness of* $\varphi_m(Q_1(X,a))$, $\psi_n(Q_2(Y,b))$ *implies, and hence is equivalent to, independence of* $Q_1(X,a)$, $Q_2(Y,b)$. The connection here depends *decisively* on some properties of complex functions of several variables, by forming the ideal generated by the infinite collection $\{\gamma_{m,n}(a,b) - \alpha_m(a)\beta_n(b), m,n \ge 1, a \in A_1, b \in B_1\}$ which then is also determined by just a **finite** set of generators of the same ideal. [Recall that an ideal is a sub ring of a ring which contains also the product of any element of a ring with an element of the ideal.] The above conclusion is a consequence of a renowned theorem of H. Cartan, the details of which are discussed, for instance, in Gunning and Rossi (1965, Chapter VIII), and a specialized version for the present purposes is given with clear exposition by Linnik (1968, pp. 83–87). In applications the functions ϕ_n, ψ_n can be chosen in different ways, for instance, as:

$$\varphi_n(x) = x^n e^{-g(x)}; \quad \psi_n(y) = y_n e^{-h(y)}, \quad n \ge 0,$$

where $g, h \ge 0$ are polynomials of degree at least 2, and $x > 0$, $y > 0$. Thus the unexpected relation between the fundamental concept of stochastic independence of random variables and its function theory counterpart with uncorrelatedness (or essential orthogonality) with a *finite set* of functions of the random variables is both an interesting and significant byproduct of some classical results in several complex variables. In the case of normal random variables the equivalence of these two concepts is well-known and simply verified. Note that if there are n random variables, in place of 2, the same conclusion extends since the basic Cartan theorem is valid in that case also.

III. SOME EXAMPLES AND PARADOXES

In this section two examples will be given to illustrate how the concepts and results of the preceding section can be used with precise descriptions, or abused if sufficient care in the description of a stochastic model is not exercised.

It is a common practice in probability to use gambling terminology to state several experiments and models, simply because no further detail of description is needed and the terms are universal. For two of the examples below the models are therefore based on coin tossing repeatedly until a kth head appears for the first time. The number of tosses can be infinite.

A. Example 1

Since the kth head may appear on the nth toss, $n \ge 1$ being any integer, our sample space Ω must consist of infinite sequences of labels "head" and "tail," which will be denoted by 1 and 0, respectively. Thus, $\omega \in \Omega$ if, for example, $\omega = (0, 1, 0, 0, \ldots, 1, 0, \ldots)$. Since every such ω can be put in a one-to-one relation with a point of the unit interval (using the dyadic expansion of reals), Ω has as (uncountably) many points, so that $\mathcal{A}$ has to be chosen with some care. Let $I_n \subset \Omega$ be an "interval," that is, $\omega \in I_n$, if and only if the first n components of ω have a prescribed pattern. For instance I_2 may have all $\omega \in \Omega$ whose first two components are (0, 1) so that $\Omega - I_2$ will have all ω with (1, 1), (1, 0), or (0, 0) as its first two coordinates. Let $\mathcal{A}$ be the smallest σ algebra containing all such I_n, $n \ge 1$. (Note that $I_n \cap I_m = I_{\min(m,n)}$ is an "interval," but $I_n \cup I_m$ or $I_n - I_m$ is not an interval, although $I_n - I_m$ can be expressed as a finite disjoint union of such intervals.) If I_n has k ones, and $n-k$ zeros, then define $P(I_n) = \binom{n}{k} p^k (1-p)^{n-k}$, where the probability of a head is p. This describes the model $(\Omega, \mathcal{A}, P)$, and, for instance, if $\tilde{I}_n$ is the set of ω's corresponding to the first head on the nth toss, then $\tilde{I}_n \in \mathcal{A}$ and $P(\tilde{I}_n) = (1-p)^{n-1} p$, since the order is fixed. Let $X(\omega)$ equal the number of zeros before the first head in ω. Then X is a countably valued random variable, since for each integer N, $\{\omega : 0 \le X(\omega) \le N\} = \tilde{I}_1 \cup \tilde{I}_2 \cup \cdots \cup \tilde{I}_{N+1} \in \mathcal{A}$. With this setup, all pertinent questions of the experiment can be answered. For instance.

$$E(X) = \sum_{n\geq 0} nP[\omega: X(\omega) = n]$$
$$= \sum_{n=0}^{\infty} nq^n p, \quad q = 1 - p$$
$$= pq \sum_{n=1}^{\infty} \frac{d}{dq}(q^n)$$
$$= pq \frac{d}{dq}\left(\frac{q}{1-q}\right) = \frac{q}{p}.$$

Similarly, Var X can be shown to be q/p^2. Higher moments are similar, but the computations are more invloved.

The same model $(\Omega, \mathcal{A}, P)$ can be used to present the following extension. If $Y(\omega)$ is the rth head after $(n-1)$ tails, then

$$P[Y = n] = \binom{n+r-2}{n-1} p^r q^{n-1}.$$
$$n = 1, 2, \ldots$$

If $r = 1$, this reduces to the previous case, called the *geometric distribution*. The general case is said to be the *negative binomial* or *Pascal distribution*. It can be shown that $E(Y) = nq/p$, Var $Y = nq/p^2$. This distribution has applications in many areas and also arises as a limit distribution for experiments describing a contagion or with "Bose–Einstein statistics" in physics.

B. Example 2

This example shows how *imprecise specification of an experiment can lead to difficulties*. The problem is known as Bertrand's paradox, and it is one of several such cases listed by J. Bertrand at the end of the nineteenth century exemplifying the difficulties with the subject at that time. Consider a circle of unit radius and inscribe an equilateral triangle in it so that its side is $\sqrt{3}$. Suppose a chord of the circle is chosen at random. What is the probability that its length exceeds the length of a side of the triangle, that is $\sqrt{3}$?

1. Solution 1

For Ω take the circle of radius 1 and $\mathcal{A}$ as the σ-algebra detemined by all concentric disks inscribed in Ω. For each $A \in \mathcal{A}$, let $P(A)$ be the proportion of the area of A contained in Ω. Then the required event is {chord of length $\geq \sqrt{3}$} $= A_0$ (say). The chord is fixed uniqely if it touches an inscribed circle. So the desired event is equivalent to the midpoint of the chord lying inside the circle of radius $\frac{1}{2}$ since that is the distance of the side of the equilateral triangle from the center. Hence, A_0 is the disk of radius $\frac{1}{2}$, and $P(A_0) = \pi(\frac{1}{2})^2/\pi = \frac{1}{4}$.

2. Solution 2

Let θ be the angle subtended by the chord at the center of the circle. Take $\Omega = [0, 2\pi]$, $\mathcal{A} = \sigma$-algebra determined by all the subintervals of Ω, and $P(A) =$ proportional length of A contained in Ω. In this model, the desired event A_0 holds if and only if $2\pi/3 < \theta < 4\pi/3$. Hence, $A_0 \in \mathcal{A}$, and $P(A_0) = (4\pi/3 - 2\pi/3)/2\pi = \frac{1}{3}$.

3. Solution 3

Since the position of the chord is uniquely determined by its distance from the center, one can take $\Omega = [0, 1]$, $\mathcal{A} = \sigma$-algebra determined by the subintervals, of Ω, and $P(A) =$ length of A as a legitimate model. Now the desired A_0 occurs if and only if the chord lies at a distance $\leq \frac{1}{2}$. Thus, $P(A_0) = \frac{1}{2}$.

Several other models can be constructed. Of the numbers $\frac{1}{4}, \frac{1}{3}, \frac{1}{2}$, which is the correct probability? Since the reasoning is correct in each case and the answers are different, one has a "paradox." Actually, all three models and hence all three answers are correct since one is conducting a *different experiment* each time. Here "chord is chosen at random" is ambiguous, since one can choose it in so many different ways, and it is necessary to indicate the method used in the "random selection." Thus, when such inconsistent results appear, it is necessary to examine the description of the problem carefully and eliminate multiple interpretations so as to express the experiment with a unique probability model $(\Omega, \mathcal{A}, P)$.

The model construction is made in detail to illustrate the basic ideas, but with experience one can visualize the procedure and reduce the calculations to a few lines. However, the underlying sample space and the resulting model should be clearly recognized. Only then should the theory be applied. Another type of paradox for evaluating conditional probabilities is given in Section V.C below.

IV. STRONG LAW OF LARGE NUMBERS AND ITS USES

In the preceding sections the laws of large numbers due to J. Bernoulli and P. L. Čebyšev were discussed. The general form says that, if $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables on $(\Omega, \mathcal{A}, P)$, with mean μ and variance σ^2, $S_n = \sum_{i=1}^{n} X_i$, then, omitting the display of $\omega \in \Omega$, one has

$$\lim_{n\to\infty} P[|S_n/n) - \mu| \geq \varepsilon] = 0,$$
$$\text{for each} \quad \varepsilon > 0. \tag{14}$$

To justify the frequency (or long-term-average) interpretation of probability, this statement is not adequate, not even for the Bernoulli variables. From the original publication

in 1713 of Bernoulli's work, the desired strong statement awaited many developments. It was only in 1909 that E. Borel succeeded in establishing the following result, as indicated in Section I, called the strong law of large numbers. If $X_1, X_2, \ldots$ are independent Bernoulli variables with the (constant) probability of success, $p, 0 < p < 1$, and $S_n = \sum_{i=1}^{n} X_i$, then

$$\lim_{n\to\infty} S_n/n = E(X_1) = p \qquad (14')$$

with probability 1. This statement means that there is a set $N \in \mathcal{A}$, $P(N) = 0$, and for each $\omega \in \Omega - N$, $S_n(\omega)/n \to p$ as $n \to \infty$. This is a much stronger statement than Eq. (14) and is also more difficult to prove. The result of Eq. (14′) initiated a new stage in probability. It was finally generalized by A. Kolmogorov in 1928 to all independent identically distributed random variables with *one moment existing* and shown to be the best result one can obtain. Let us explain the significance of this strong law.

First observe that, by Eq. (14), $|(S_n/n) - \mu|$ is likely to be small for large n, fixed, but this difference is not asserted to remain uniformly small as n is further increased. In contrast, Eq. (14′) says that the difference $|(S_n/n) - p|$ remains small for all n after a certain n_0, and this is the strength. Thus, for independent Bernoulli variables, the relative frequency S_n/n for large n gives practically the (unknown) probability p of success. This is the vindication of Cardano's feeling that these proportions "should stabilize."

The above discussion does not suggest the deeper implications of the strong law. To explain it further, we present three applications to (1) empirical distributions, (2) random walk, and (3) the St. Petersburg paradox. The last shows that the *assumptions* of Kolmogorov's generalization *should not be overlooked*.

The strong law of large numbers allows one to estimate the distribution of a random variable. Let $X_1, X_2, \ldots$ be independent random variables with a common distribution F. Typically, F is not known; at least one wants to test whether it is the assumed function. To do this consider, for each real x and positive integer n, the proportion $\nu_n(x)$ of the observations $X_1, \ldots, X_n$, which are less than x. Thus, $\nu_n(x) = (1/n)\{\text{no. of } X_i < x\}$. To see the meaning of this clearly, let $Y_i = 1$ if $X_i < x$; otherwise, $Y_i = 0$. Set $S_n(x) = \sum_{i=1}^{n} Y_i$. Then $\nu_n(x) = S_n(x)/n$. Since the Y_i are Bernoulli distributed and independent, $P[Y_i = 1] = P[X_i < x] = F(x)$, by Eq. (3). It follows by Eq. (14′) that $S_n(x)/n \to F(x)$ with probability 1. This is an application of the strong law. With some more work, one can show that this limit is uniform in x. In symbols,

$$P\left[\lim_{n\to\infty} \sup_{-\infty<x<\infty} |\nu_n(x) - F(x)| = 0\right] = 1 \qquad (15)$$

Here ν_n is called the sample or empirical distribution, and it gives a true picture of F when n is large enough. There are analogous results in the multidimensional case to estimate $F_{X_1,\ldots,X_n}$ of Eq. (8). The statement of Eq. (15) is called the Glivenko–Cantelli theorem, and its proof is not simple.

There is also a great deal of interest attached to the fact that $\nu_n(x) \to F(x)$ not only uniformly with probability one, as noted above, but that $P([\nu_n(x) \in A]) \to 0$, if $F(x) \notin A$, exponentially fast. It is a case of rare events, and the rate of decay of the probability of $[\nu_n(x) \in A]$ is of considerable interest for practical applications. [The jackpot in a national lottery is an example of a rare event and its great popular interest is evident.] In the 1930s, this aspect was investigated by N. V. Smirnov, H. Cramér and others, and as a consequence one finds that exponential decay indeed does occur. A comprehensive statement is as follows. If $Q_n(\cdot)$ is the probability distribution of the empiric function ν_n, then for each $A \subset \mathbb{R}$ (a Borel set), one has $Q_n(A)$ decreases exponentially. More precisely, $\lim_{n\to\infty} \log Q_n(A) = -\inf_{x\in A} I(x)$ where I is the *rate function* defined as complementary to the log moment (or cumulant) generating function of X_1, so that if μ is the probability distribution of X_1 (hence also of $X_i, i \geq 1$) and $\Lambda_\mu(t) = \log E(e^{tX_1})$, $x \in \mathbb{R}$, then $I(x) = \sup\{tx - \Lambda_\mu(t) : t \in \mathbb{R}\} \geq 0$. This result is also not simple, and one can consult Sanov (1957) for it and some generalizations. It leads to what is called the *large deviation theory*, and an outline of which is given in S.R.S. Varadhan (1984). There an account of how the ideas are applied to numerous problems including stochastic processes can be found.

Next a random walk example will be illustrated. Suppose that a particle moves on the x axis one step to the right or left according to whether a tossed coin falls heads or tails, respectively. When the coin is fair, the probability of a head is $\frac{1}{2}$ so that a step to the right or left is equally likely. This is called a symmetric random walk. The problem is to find the probability that the particle visits the origin or any other step. The same question can be asked if the particle moves in a plane in directions parallel to the coordinate axes. Let us consider the equiprobable (or symmetric) case for simplicity. Then with probability $\frac{1}{4}$ the particle moves one step in one of the four directions (i.e., up, down, left, or right). In three or higher dimensions, say k, the same thing takes place with probability $1/2k$, $k \geq 3$, for each step. This problem can be used as an approximation to the diffusion model in physics. The above random walk is unrestricted, but one can put barriers in one or more directions either to trap (or absorb) the particle or to turn back (or reflect) at the barrier. These are important but not simple questions.

Let us consider the one-dimensional case introduced above. Let Ω consist of all points ω, e.g., $(1, 0, 0, 1, \ldots)$, composed of ones and zeros corresponding to a head and tail in an unlimited sequence of tosses. Take $\mathcal{A}$ to

be the same class as that described in the first example of Section III, and P as the corresponding probability with $p = \frac{1}{2}$. Thus, $(\Omega, \mathcal{A}, P)$ will be the same model. Let $X_k(\omega) = +1$ or -1 depending on whether the kth coordinate is 1 or 0. So $P[X_k = 1] = P[X_k = -1] = \frac{1}{2}$. Then $X_1, X_2, \ldots$ are independent Bernoulli random variables. Let $S_n = \sum_{k=1}^{n} X_k$ with $S_0 = 0$, to denote the (random) position of the particle on the line at the nth step. If the particle starts at a point h on the x axis, then one takes $Y_k = X_k - h$, so that $P[Y_k = 1] = P[Y_k = -1] = \frac{1}{2}$, and the problem reduces to the preceding case. By Eq. (14′), $S_n/n \to 0$ with probability 1 ($\sum_{k=1}^{n} Y_k/n = (S_n/n) - h \to 0$ in the general case). This implies, at least intuitively, that S_n takes 0 (or h) infinitely many times. However, a rigorous proof of the statement requires more work. Actually in 1921, G. Pólya showed that the symmetric random walk visits the origin infinitely often in one and two dimensions (so that "all roads lead to Rome" in these cases) but that in three dimensions there is positive probability that it may not return to the origin. The latter statement holds in k dimensions, $k \geq 3$. It was also found in 1940 that the probability of returning to the initial position in three dimensions is ~ 0.35. Thus, the particle drifts to infinity with greater probability than a return to the origin.

Even when the X_k's are only independent real (not necessarily Bernoulli) random variables with a common distribution, $\{S_n, n \geq 0\}$ is called a general random walk, where $S_n = \sum_{k=1}^{n} X_k$, $S_0 = 0$. If $X \in \mathbb{R}$ and if $P[|S_n - x| < \varepsilon$, for infinitely many $n] = 1$ for each $\varepsilon > 0$, then S_n is said to visit x infinitely often, and x is called a recurrent point of the random walk. A relatively deeper fact is that the origin is a recurrent point of the random walk S_n whenever $P[|S_n/n| > \varepsilon] \to 0$ as $n \to \infty$, for any $\varepsilon > 0$. In particular if the X_k have means zero, then the law of large numbers implies this. Probability theory studies the behavior (or fluctuation phenomena) of partial sums of different types of random variables. The sequence $\{S_n, n \geq 0\}$ is an instance of a random or stochastic process in which the variables are no longer Bernoulli, and hence the counting methods are insufficient. New techniques and theory are essential. In discussing the classical period and teaching of mostly simpler problems, J. L. Doob observed in 1942 (in explaining the need for a general study of stochastic processes) the following: "This calculation [with permutations and combinations] looms large in elementary courses, however, only because calculations need little mathematical preparation either for student or teacher. It is considered easier to perform long calculations than to develop a general point of view, bolstered by theory. This may be good pedagogy, up to a certain point, but gives students a false picture of the subject." This admonition is generally heeded in the later developments of the subject, and at present probability theory involves the study of genuinely broader phenomena, being motivated by special cases.

The fluctuation phenomenon of partial sums noted above is at the heart of the single- and multiple-server queuing problems, sequential analysis, and other applications. They cannot be considered here.

As a final example, we show how overlooking the moment condition of the strong law leads to a difficulty, called the St. Petersburg paradox. Suppose that a game is played with a fair coin until it falls heads. If a head falls on the nth toss, the player gets 2^n dollars and nothing otherwise. Here again the probability model $(\Omega, \mathcal{A}, P)$ is the same as in the preceding example. Let $X(\omega)$ represent the gain for the player on ω. So if the first 1 in ω is at the nth place, then $X(\omega) = 2^n$ and $P[X = 2^n] = 2^{-n}$. Let $X_1, X_2, \ldots$ be independent random variables with the same distribution as X. Thus, X_n represents the nth play. If $S_n = \sum_{k=1}^{n} X_k$, then $S_n/n \not\to$ a finite limit since $E(X) = \sum_{n=1}^{\infty} 2^n(1/2^n) = +\infty$ and the strong law of large numbers is not applicable. If $E(X) = \mu$ is finite, then $S_n/n - \mu \to 0$ with probability 1 and if μ is the entrance fee for the player per game, then it is fair. However, in the problem at hand, for any α, $S_n - n\alpha$ will be positive for large n, and hence to make the game fair, the entrance fee α must be infinite, which is a paradoxical situation. A number of modifications have been proposed from the time this "paradox" was noted by Daniel Bernoulli in 1730 in St. Petersburg. Since the hypothesis of the existence of one moment is needed for the strong law, that is not applicable here and one should not try to "bend" the result to the present case. Actually, it can be shown, by the method of truncation, that $S_n/n \log_2 n \to 1$ with probability 1, so that $(1/n)(S_n - n \log_2 n) \to 0$ and the game may be called "fair." Here, $\log_2 n$ is the logarithm to the base 2. This shows that the entrance fee should be variable from game to game ($log_2 n$ dollars), and the paradox disappears.

In reality, it turns out that if a fixed entrance fee α is charged and α is large, one has to play an extraordinarily large number n of games in order to have $S_n > n\alpha$ with probability 0.99, for instance. If the player is allowed only one game, then the situation, of course, changes. Thus, the probability that there is at least one head before N tosses is $= 1 - 2^{-N+1}$, and hence the player will suffer a loss if N is large, even though the gain is 2^N dollars if the first head appears on the Nth trial.

V. CONDITIONAL PROBABILITY AND DEPENDENCE

Thus far only independent random variables and experiments have been discussed. The next decisive step in the development is to consider classes of dependence.

Even the random walk process $\{S_n, n \geq 0\}$, with $S_n = \sum_{k=1}^{n} X_k$, $S_0 = 0$, has already involved a dependent sequence since $S_{n+1} = X_{n+1} + S_n$, and S_n's are not independent. Here a new concept is needed and it is introduced as follows.

A. Conditioning

To consider the simplest problem, let A, B be two events and, if B is known to have occurred, find the probability of A given B, denoted $P(A \mid B)$. Since the knowledge of B should be used with A, one considers $A \cap B$ (=A and B together). Assuming $P(B) > 0$, one has to look at the new class of events $\mathcal{A}(B) = \{A \cap B\colon A \text{ in } \mathcal{A}\}$ and consider $P(A \cap B)$. Since $0 \leq P(A \cap B) \leq P(B)$ and B is fixed, to make $P(A \mid B)$ a probability function [so $P(B \mid B) = 1$] one normalizes $P(A \mid B)$ as $P(A \cap B)/P(B)$ and calls $P(\cdot \mid B)$ the conditional probability on $\mathcal{A}$, given B. Note that, if A, B are independent so that intutively the occurrence of B has no "influence" on A, one gets $P(A \mid B) = P(A)$ by Eq. (4), and hence this definition extends the earlier concept. It is useful to express it as

$$P(A \text{ and } B) = P(B)P(A \mid B) \tag{16}$$

and this can be employed to calculate the left side if the right side is known. As long as $P(B) > 0$ this is well defined, but if $P(B) = 0$, then the left side is clearly zero also, so that $P(A \mid B)$ is indeterminate but Eq. (16) always makes sense. Let us set aside the indeterminate case first, and suppose that both $P(A) > 0$ and $P(B) > 0$. Then one can define $P(B \mid A)$ in the same way, and Eq. (16) implies interchanging A, B:

$$P(B \mid A) = P(B)P(A \mid B)/P(A). \tag{17}$$

This simple formula is of considerable interest in applications. A useful form is to take Ω as the union $\bigcup_{k=1}^{n} \Omega_k$, $P(\Omega_k) > 0$ for each k, and $\Omega_k \cap \Omega_{k'} = \emptyset$.Thus, Ω occurs if and only if exactly one event Ω_k occurs. Since $A = A \cap \Omega = \bigcup_{k=1}^{n} (A \cap \Omega_k)$, and $P(A) = \sum_{k=1}^{n} P(A \cap \Omega_k)$ for any event A, one has with Eq. (16)

$$P(A) = \sum_{k=1}^{n} P(\Omega_k)P(A \mid \Omega_k). \tag{18}$$

Taking $B = \Omega_k$ in Eq. (17) and using Eq. (18), one gets the *Bayes formula*:

$$P(\Omega_k \mid A) = \frac{P(A \mid \Omega_k)P(\Omega_k)}{\sum_{k=1}^{n} P(A \mid \Omega_k)P(\Omega_k)}. \tag{19}$$

This result has the following interesting interpretation. If an event takes place in one of n mutually exclusive and exhaustive ways with probability $p_k [= P(\Omega_k) > 0]$, then after observing an event A, one can calculate by Eq. (19) the conditional probability of the hypothesis Ω_k given A if $P(A \mid \Omega_k)$ and p_k are known. The practical applications of this interpretation, which are quite useful in real-life problems, will be indicated later in the section. When $P(B) = 0$, then evaluation of $P(A \mid B)$ is difficult as shown in Section C.

Let us consider now the profound significance of conditional probability. The concept, in fact, was already employed in the 17th and 18th centuries. Actually formulae (17) [and (19)] were used by the clergyman T. Bayes (1702–1761) himself. His paper was published 2 years after his death.

Here two examples are considered since they lead to the important Markovian and martingale dependencies:

1. Example 1

Suppose that an urn contains r red and b blue balls of the same physical characteristics. With each random drawing of a ball, it is returned together with c balls of its type and d balls of the other type, where c, d are integers (positive or not). The process is stopped when the number of balls of either color is likely to be negative at a drawing. This is a variation of the classical Pólya urn scheme. The probability of a sequence RB (red, blue) is given by Eq. (17) as

$$P(\text{RB}) = \frac{r}{b+r}\,\frac{b+d}{b+c+r+d}$$

and similarly

$$P(\text{BR}) = \frac{b}{b+r}\,\frac{r+c}{b+c+r+d}$$

which are generally different, since the order matters. If $c = -1$, $d = 0$, and one does not need to recognize the order of the sequence, then

$$P(\text{a red and blue}) = \frac{2br}{(b+r)(b+r-1)} = \binom{b}{1}\binom{r}{1}\Big/\binom{b+r}{2}.$$

The last formula, the more familiar one, can also be obtained without using the notion of conditional probability. On the other hand, suppose that there are two urns having $b_1 + r_1$ and $b_2 + r_2$ balls of the types described. If one flips a coin and chooses urn I if the coin falls heads and II if it is tails, with probabilities p and $q = 1 - p$, then

$$P(\text{RB}) = pP(\text{RB} \mid \text{I}) + qP(\text{RB} \mid \text{II}) \qquad \text{by formula (18)}$$

$$= p\frac{r_1}{b_1+r_1}\,\frac{b_1+d_1}{b_1+c_1+r_1+d_1} = q\frac{b_2}{b_2+r_2}\,\frac{c_2+r_2}{b_2+c_2+r_2+d_2}.$$

These results have applications in, for example, accident insurance or contagion. Using the fact that $P(B) = pP(B \mid \mathrm{I}) + qP(B \mid \mathrm{II})$, one has the conditional probability $P(\mathrm{R} \mid \mathrm{B}) = P(\mathrm{RB})/P(\mathrm{B})$, which will be different from $P(\mathrm{R})$.

2. Example 2

Let us go on to another set of ideas. We start with the classical gambler's ruin problem, which was originally proposed by the Dutch mathematician C. Huygens (1629–1695) in his book published in 1657 and was solved by A. de Moivre in 1711. The solution was also obtained by P. Montmort (1678–1719) and Nicholas Bernoulli independently at about the same time. To state it, suppose players I and II have initial fortunes of a and b dollars, and one plays against the other by tossing a coin. If the coin falls heads, I gets a dollar from II and, if if it is tails, gives a dollar to II. The game stops when I is ruined (i.e., has zero dollars) or II is ruined (so I gets $a+b$ dollars). The problem is to find the probability that I is ultimately ruined if p is the probability of a head in a toss.

Let $p_a(q_a)$ be the probability of ultimate win (ruin) of I if the starting capital is a. After the first toss, I has either $a+1$ or $a-1$ dollars. Hence Eq. (18) implies

$$q_a = pq_{a+1} + qq_{a-1}. \tag{20}$$

If $a=1$, then clearly $q_1 = pq_2 + q$, since $q_0 = 1$, and similarly $q_{a+b} = 0$ so that $q_{a+b-1} = qq_{a+b-2}$. Hence, for solving the problem, Eq. (20) can be rewritten in a more familiar form,

$$q_x = pq_{x+1} + qq_{x-1} \qquad x = 1, 2, \ldots, a+b-1 \tag{21}$$

with $q_0 = 1$, $q_{a+b} = 0$. This is a second-order difference equation, and the standard calculus method is to consider its characteristic polynomial:

$$pt^2 - t + q = 0$$

or

$$pt^2 - t + 1 - p = (t-1)(pt - q) = 0.$$

The roots are thus $t_1 = 1$, $t_2 = q/p$. The general solution of Eq. (21) is

$$q_x = At_1^x + Bt_2^x \quad \text{or} \quad q_x = A + Bx \tag{22}$$

depending on whether $t_1 \neq t_2$ or $t_1 = t_2$. Using the boundary conditions $q_0 = 1$ and $q_{a+b} = 0$, one gets, if $q \neq p$, then $A + B = 1$, $A + Bt_2^{a+b} = 0$, so that

$$B = [1 - (q/p)^{a+b}]^{-1}, \quad A = 1 - B$$

and if $q = p = \frac{1}{2}$, then $A = 1$, and $B = -(a+b)^{-1}$. Hence,

$$q_a = \frac{(q/p)^a - (q/p)^{a+b}}{1 - (q/p)^{a+b}} \qquad a = 1, 2, \ldots, a+b-1, \tag{23}$$

and when $p = q = \frac{1}{2}$,

$$q_a = 1 - \frac{a}{a+b} = \frac{b}{a+b} \qquad a = 1, 2, \ldots, a+b-1. \tag{24}$$

By interchanging p and q, as well as b and a, one gets the probability p_a of winning from Eqs. (23) and (24), satisfying $p_a + q_a = 1$, as expected. If Y denotes the ultimate gain for I, it is a random variable such that $P[Y = b] = p_a$ and $P[Y = -a] = q_a$. Thus, the expected gain for I is

$$E(Y) = -aq_a + bp_a = (a+b)p_a - a \quad \text{if } p \neq q$$
$$= -a\frac{b}{a+b} + b\frac{a}{a+b} = 0 \quad \text{if } p = q = \frac{1}{2}.$$

Thus, the second case represents a "fair" game.

B. Markov Dependence

A reason for giving the detailed description above is that the ruin problem has many variations and has motivated several important developments in probability. We indicate two of these. Let $A_1, A_2, \ldots, A_n$ be events such that $P(A_i) = a_i > 0$, for all i. Then Eq. (16) admits an obvious extension:

$$P(A_1, \ldots, A_n) = P(A_1 \text{ and } A_2 \text{ and } \ldots \text{ and } A_n)$$
$$= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1, A_2)$$
$$\times \cdots P(a_n \mid A_1, \ldots, A_{n-1}).$$

In 1906, A. A. Markov, a pupil of P. L. Čebyšev, introduced the fecund idea of a chain dependence; namely, the conditional probability of A_k, $k > 1$, given $A_1, \ldots, A_{k-1}$, depends only on the preceding event A_{k-1}. So a sequence $A_1, A_2, \ldots, A_n$ of events, with $P(A_k) > 0$ all k, is a Markov chain if

$$P(A_1, \ldots, A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_2) \times \cdots P(A_n \mid A_{n-1}). \tag{25}$$

To use a picturesque language, setting $p_{ij} = P(A_j \mid A_i) \geq 0$, one says that a system forms a stationary Markov chain if p_{ij} is the probability of being in state j, given that in the preceding instant it was in state i; then for $i_1 < i_2 < \cdots < i_n$ one has

$$P\left(A_{i_1}, \ldots, A_{i_n}\right) = a_{i_1} p_{i_1 i_2} p_{i_2 i_3} \cdots p_{i_{n-1} i_n}, \tag{26}$$

where $a_i = P(A_i)$ is the initial probability of being in state i. The numbers p_{ij} are called stationary transition probabilities. Thus, the system is completely determined by

the set of initial probabilities and the transition probabilities. The word *stationary* is often omitted, but its significance is that the a_i's and p_{ij}'s do not depend on the particular instant and are the same starting at any time and moving in unit steps. In the contrary case, they could depend on the time variable and nonstationarity results, as shown below.

If the set of all possibilities for the system is finite, it is called a finite Markov chain, and the possibilities are states (the complete set being the state space). If the states are labeled $1, 2, \ldots, n$, and the initial probability vector is $(a_1, a_2, \ldots, a_n)$, so that $\sum_{i=1}^{n} a_i = 1$, then this and the transition

$$Q = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & & \cdot \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \tag{27}$$

describe the Markov chain. If the number of states is infinite, then $n = +\infty$, and one has a denumerable Markov chain. Note that one always has $\sum_{j=1}^{n} p_{ij} = 1$, for each $i = 1, 2, \ldots, n$, since from each state the system moves into one of the other states. In matrix theory, such square matrices of nonnegative elements whose row sums add to 1 are called stochastic matrices. The general theory says that, conversely, with every stochastic matrix it is possible to associate a Markov chain. Thus, such a special matrix theory merges with this part of probability.

The gambler's ruin example noted above will now be expressed as a Markov chain. The possible states are $0, 1, \ldots, a+b$ where players I and II have initial fortunes of a and b dollars and the game continues between the states $1, 2, \ldots, a+b-1$, with probabilities $p_{i,i+1} = p$ and $p_{i,i-1} = q$. Also, $p_{0,0} = 1$, $p_{a+b,a+b} = 1$, and $p_{ij} = 0$ if $| i - j | \geq 2$, $p_{ii} = 0$, $1 \leq i \leq a+b-1$ since the system moves in these states. Thus, Q is a matrix of size $(a+b+1)$ by $(a+b+1)$ given by

$$Q = \begin{bmatrix} 1 & 0 & 0 & & & \cdots & 0 \\ q & 0 & p & 0 & & \cdots & 0 \\ 0 & q & 0 & p & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & & & \vdots \\ & & & & \ddots & & \\ \vdots & \vdots & & & q & 0 & p \\ 0 & 0 & & \cdots & 0 & 0 & 1 \end{bmatrix}$$

The initial distribution is $a_i = 1$ if $i = a$, and $a_i = 0$ if $i \neq a$. This completes the description of the ruin problem as a Markov chain. There are many other applications to random walks, diffusion, birth–death processes, and the like.

A transition matrix need not have its columns sum to 1, but if it has this additional property, it is called doubly stochastic. Such chains admit special properties. Clearly, a symmetric transition matrix is doubly stochastic, and a doubly stochastic 2-by-2 matrix is symmetric, but it is easily seen that there are doubly stochastic non-symmetric matrices of order k by k for each $k \geq 3$. Symmetric transition matrices correspond to "path-reversible" chains.

In terms of random variables, these concepts are as follows. If $X_1, X_2, \ldots$ is a sequence of random variables on $(\Omega, \mathcal{A}, P)$ taking values in a fixed finite set $(\alpha_1, \ldots, \alpha_n)$, each value with positive probability, then it is termed a finite stationary Markov process if, for $k \geq 2$, one has

$$\begin{aligned} &P[X_k = \alpha_j \mid X_1 = \beta_1, \ldots, X_{k-1} = \beta_{k-1}] \\ &\quad = P[X_k = \alpha_j \mid X_{k-1} = \beta_{k-1}], \end{aligned} \tag{28}$$

where $\beta_1, \ldots, \beta_{k-1}$ are a not necessarily distinct subset of $(\alpha_1, \ldots, \alpha_m)$. Let $p_{ij} = P[X_k = \alpha_j \mid X_{k-1} = \alpha_i]$, $k = 2, 3, \ldots, n$. If p_{ij} depends also on k, then one calls $X_1, X_2, \ldots$, a (finite) nonstationary Markov process. The general theory of Markov processes, for which the range, or state space, may be uncountables, constitutes an important branch of probability with numerous applications.

A familiar example of a Markov process is the sequence $\{S_n, n \geq 1\}$, where $S_n = \sum_{k=1}^{n} X_k$ with X_k independent, and this will be stationary if the X_k all have the same distribution. This is easy to verify if all the X_k take values in the fixed finite set $(\alpha_1, \ldots, \alpha_m)$ considered above, but it is also true for any independent random variables, and then one needs to define the conditioning concept in a more general context using advanced techniques. This shows how the subject is spreading in different directions.

In the previous discussion the system was restricted to move from one state to another in one step, but higher transitions are possible and they can be easily determined. Thus, a transition from a state j to another k in n steps is the conditional probability of the system moving from j to k in n steps, denoted $p_{jk}^{(n)}$. Between j and k the chain can visit all possible neighboring states. This n-step transition probability is calculated inductively as

$$p_{jk}^{(n)} = \sum_{l} p_{jl} p_{lk}^{(n-1)}, \quad p_{jk}^{(1)} = p_{jk}, \quad n \geq 2,$$

where l runs through all-states of the state space. A further induction gives the *Chapman–Kolmogorov formula*:

$$p_{jk}^{(m+n)} = \sum_{l} p_{lk}^{(m)} p_{lk}^{(n)}, \quad m \geq 1, \quad n \geq 1. \tag{29}$$

A symbolic matrix multiplication can be used by setting $Q^{(1)} = Q = (p_{lk}^{(1)})$. Then Eq. (29) becomes

$$Q^{(m+n)} = Q^{(m)} Q^{(n)}, \tag{30}$$

and if $Q^{(0)}$ is an identity matrix, then $Q^{(n)}$ is symbolically the nth power of Q. For the general (not necessarily stationary) Markov process, set

$$p_{jk}(r, s) = P[X_s = \alpha_k \mid X_r = \alpha_j], \quad 1 \le r < s$$

which is the transition probability of the system to move from state j at the epoch r to state k at epoch s. Then Eq. (29) becomes

$$p_{jk}(r, t) = \sum_l p_{jl}(r, s) p_{lk}(s, t), \quad 1 \le r < s < t \qquad (31)$$

and writing $Q(r, s) = (p_{jk}(r, s), 1 \le j, k \le m)$, one has the general form of the **Chapman–Kolmogorov equation** as

$$Q(r, t) = Q(r, s)Q(s, t), \qquad 1 \le r < s < t \qquad (32)$$

with $Q(r, r)$ an identity. In the stationary case $Q(r, s) = Q^{(s-r)}$, which thus depends on only the time elapsed and not the instants r, s individually. The functional equations (30) and (32) have important connections in analysis. Thus, Eq. (30) is the semigroup property for stochastic matrices, whereas Eq. (32) is called the evolutionary property. This shows the far-reaching nature of the Markovian concept and suggests how probability theory creates and helps settle problems in many other areas.

There are further results in which the above Markovian dependence is replaced by kth-order Markovian dependence in the sense that the present depends not just on the immediate past but on k-preceding past values. These processes arise as solutions of certain kth-order stochastic differential equations and in other problems. They can not be discussed here.

C. Martingales

Another important random process is called a martingale. To introduce this, it is necessary to state the concept of conditional expectation. In Eq. (10) the ordinary expectation of a random variable is defined using its probability (or distribution). The same is true of conditinal expectations wherein the probability function is replaced by the conditional probability function. Thus, if X, Y are random variables with values in $(\alpha_1, \alpha_2, \ldots)$ and $P[Y = \alpha_j] > 0$, then $P[X = \alpha_i | Y = \alpha_j]$ is well defined, and the conditional expectation of X given $Y = \alpha_{j_0}$, denoted $E(X \mid Y)(\alpha_{j_0})$, is obtained as

$$E(X \mid Y)(\alpha_{j_0}) = \sum_{i \ge 1} \alpha_i P[X = \alpha_i \mid Y = \alpha_{j_0}] \qquad (33)$$

provided that the series converges absolutely for each $j_0 \ge 1$. Similarly, if $Y_1, Y_2, \ldots, Y_k$ are fixed at $\beta_1, \ldots, \beta_k$, then

$$E(X \mid Y_1, Y_2, \ldots, Y_k)(\beta_1, \ldots, \beta_k) = \sum_{i \ge 1} \alpha_i P[X = \alpha_i \mid Y_1 = \beta_1, \ldots, Y_k = \beta_k] \qquad (34)$$

when the series is absolutely convergent and the conditional probabilities are well defined. In the cases where $P[Y = \alpha_{j_0}] = 0$, there will be problems and the conditional probability is indeterminate, as seen in Eq. (16). If it is given a fixed value, the advanced theory shows that Eqs. (33) and (34) are still defined, but there will be some nonuniqueness. This does not do any harm, and there are methods of dealing with them, such that Eq. (34) can be used for most applications as though it were well defined. With this provision, one can express Eq. (33) symbolically as

$$E(X \mid Y) = \int_\Omega X(\omega) P(d\omega \mid Y),$$

and this holds if $E(|X|) < \infty$. Thus, $E(X \mid Y)$ exists and satisfies the rules of the ordinary expectation $E(X \mid X) = X$[like $E(1) = 1$] and

$$E(X_1 + aX_2 \mid Y) = E(X_1 \mid Y) + aE(X_2 \mid Y), \quad a \in \mathbb{R}.$$

Also if X, Y are independent, $E(X \mid Y) = E(X)$, and $X \ge 0$ implies $E(X \mid Y) \ge 0$ whatever the random variable Y is.

However, if $P(Y = \alpha) = 0$, then $P(d\omega \mid Y)$ is not defined by Eq. (33). So it [and hence the conditional expectation $E(X \mid Y)$] must be defined differently! To see how this may be done, rewrite Eq. (33) after multiplying both sides with $P(Y = \alpha_j)$, $\alpha_j \in B$. Then one has

$$\sum_{\alpha_j \in B} E(X \mid Y)(\alpha_j) P(Y = \alpha_j) = \sum_{\alpha_j \in B} \sum_{i \ge 1} \alpha_i P(X = \alpha_i \mid Y = \alpha_j) P(Y = \alpha_j) = \sum_{i \ge 1} \alpha_i P(X = \alpha_i, Y \in B),$$

or

$$\int_B E(X \mid Y)\, dP_Y = \int_\Omega X\, P(d\omega, B).$$

In particular $X = \chi_A$, writing $P(A \mid Y) = E(\chi_A \mid Y)$, as noted already, one deduces

$$\int_B P(A \mid Y)\, dP_Y = \int_A P(d\omega, B) = P(A \cap B). \qquad (*)$$

This equation defines $P(A \mid Y)$ uniquely for all events A and all $B = [a < Y \le b]$, by a standard result in Real Analysis. So A. Kolmogorov proposed such a $P(\cdot \mid Y)$ as the definition of a conditional probability which needs no further restrictions, and coincides with the earlier definition in case Y takes at most countably many values. However, an evaluation of this general object is not always simple and no recipe exists. For instance, if $P(B) = 0$ and if $B_n \downarrow B$

with $P(B_n) > 0$, an approximating sequence, then (*) may be expressed as

$$\frac{1}{P_Y(B_n)} \int_{B_n} P(A \mid Y)\, dP_Y = \frac{P(A \cap B_n)}{P_Y(B_n)}, \qquad (+)$$

and then let $n \to \infty$. Now one may employ a form of the generalized Fundamental Theorem of Calculus, to get $P(A \mid Y)$ as the "derivative" $dP(A \cap \cdot)/dP_y$. Unfortunately the answer depends on the approximating sequence $\{B_n, n \geq 1\}$, unless they are selected in some sophisticated manner. This will now be illustrated by the following two examples, sometimes called the Borel-Kolmogorov and Kac-Slepian paradoxes, to emphasize the underlying deeper problem.

1. Borel-Kolmogorov paradox

Let X, Y be independent random variables with a common exponential distribution $F(x) = 1 - e^{-x}$, $x > 0$, and $= 0$ otherwise. Let $Z = \frac{(X-a)}{Y}$, $a > 0$. If $\beta \in \mathbb{R}$, $B = [Z = \beta]$, and $A = [Y < y]$, it is desired to calculate $P(A \mid B)$. Now the right side of (+) will be evaluated with two different approximations. Elementary calculations show that the joint density $f_{Y,Z}$ of Y and Z is given by

$$f_{Y,Z}(y, z) = \begin{cases} ye^{-(yz+a)-y}, & \text{for } y > 0 \text{ and } yz > -a \\ 0, & \text{otherwise.} \end{cases}$$

Then the traditional conditional density $f_{Y|Z}$ for $Z = \beta$ is $f_{Y,Z}(y, \beta)/f_Z(\beta)$ and is seen to be

$$f_{Y|Z}(y \mid \beta) = \begin{cases} y(1+\beta)^{-2}e^{-y(1+\beta)}, & \text{for } y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

So $P(A \mid B) = \int_0^y f_{Y|Z}(u \mid \beta)\, du = (1+\beta)^4[1 - e^{-y(1+\beta)}(1 + y(1+\beta))]$. However, one can consider another approximation. Since $B = [Z = \beta] = [X - Y\beta = a]$ if $U = X - Y\beta$ so that $B = [U = a]$, one has the joint density of Y, U as

$$f_{Y,U}(y, u) = \begin{cases} e^{-y(1+\beta)-u}, & \text{for } u > 0, \text{ and } \beta y + u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Now the above formula for a conditional density becomes

$$f_{Y|U}(y \mid a) = \begin{cases} (1+\beta)e^{-y(1+\beta)}, & \text{for } y > 0, \text{ and } \beta y > -a \\ 0, & \text{otherwise.} \end{cases}$$

Thus $P(A \mid B) = 1 - e^{-y(1+\beta)}$, $y > 0$, which is different from the previous value, resulting in a paradox. The actual question of E. Borel and an attempt with an unsatisfactory solution by A. Kolmogorov in the 1920s, included in his Foundations (1933), is quite similar.

2. Kac-Slepian Paradox

A more natural problem arising in a concrete application has been studied by M. Kac and D. Slepian (1959), who showed that the usual approximation procedure is again unsatisfactory. A slightly specialized version of it is as follows. Let $\{X_t, t \geq 0\}$ be a Gaussian process with $E(X_t) = 0$ and covariance $E(X_s X_t) = e^{-(s-t)^2}$ so that X_t is differentiable in mean for each t, i.e., there is a Y_t such that $E([X_{t+h} - X_t]/h - Y_t)^2 \to 0$ as $h \to 0$. Then $\{Y_t, t \geq 0\}$ is also Gaussian and X_t, Y_t are independent. The Y_t process is "tangent" to the X_t process at each t. The problem is to find the conditional probability of $A_y = [Y_0 < y]$ given that the process started at a, i.e., $X_0 = a$. Since $B = [X_0 = a]$, $P(B) = 0$, the evaluation of $P(A \mid B)$ is not a simple computation from definition, and one can approximate it as in (+). Thus for each $m \in \mathbb{R}$ and $\delta > 0$ consider $B_\delta^m = \{\omega\colon X_t(\omega) = a + mt$ for some $0 \leq t \leq \delta(1+m^2)^{-1}\}$ so that B_δ^m is the event that X_t passes through the line $y = a + mt$ at a distance δ units from a, at some time t and that $P(B_\delta^m) > 0$. It is clear that $B_\delta^m \downarrow B$ as $\delta \downarrow 0$ for each m. Then one finds after some computation that, since Y_t is distributed as $N(0, 4)$,

$$\lim_{\delta \searrow 0} P\left(A_y \mid B_\delta^m\right) = \int_{-\infty}^{y} f_{Y_0|X_0}^m(u \mid a)\, du,$$

where

$$f_{Y_0|X_0}^m(u \mid a) = \frac{|u - m|\, e^{-\frac{u^2}{32}}}{32e^{-\frac{m^2}{32}} + m\int_{-m}^{m} e^{-\frac{u^2}{32}}\, dv},$$

which is different for each different $m \in \mathbb{R}$, so that there are uncountably many answers to the problem resulting in a strong paradox. The omitted computations can be found in the paper by Kac and Slepian (1959), and an extended treatment of conditioning as well as the details of the above problem in the book (Rao (1993)). Indeed this computational problem is not satisfactorily solved in the literature. A few available (satisfactory) procedures and suggestions for investigations are discussed in the last reference. However using an unambiguous abstract definition of conditional probability and expectation, it is possible to develop the general theory and this will now be briefly illustrated by introducing the concepts of martingales.

If $X_1, X_2, \ldots$ is a sequence of random variables on $(\Omega, \mathcal{A}, P)$ each having at least one moment, representing, for instance, fortunes of a gambler at epochs 1, 2,..., respectively, then it is called a *martingale* or a fair game provided that $E(X_{n+1} \mid X_1, \ldots, X_n) = X_n$; that is, the whole past being known (or given), the expected fortune on the next game is the same as on the last (or present) game. It is called a *submartingale* (or a favorable game to the gambler) if the expected fortune is at least as much as the present one, that is, $E(X_{n+1} \mid X_1, \ldots, X_n) \geq X_n$. If

the inequality is reversed, then one has a *supermartingale* (or unfavorable game). Thus, a martingale is both a sub- and a supermartingale.

A simple example of a martingale is given by $\{S_n, n \geq 1\}$, where $S_n = \sum_{k=1}^{n} X_k$, the X_n being independent random variables with zero means. Indeed,

$$\begin{aligned} E(S_{n+1} \mid S_1, \ldots, S_n) &= E(X_{n+1} + S_n \mid S_1, S_2, \ldots, S_n) \\ &= E(X_{n+1} \mid S_1, \ldots, S_n) + E(S_n \mid S_1, \ldots, S_n) \\ &= E(X_{n+1}) + S_n = 0 + S_n = S_n, \end{aligned}$$

since X_{n+1} is independent of $S_1, \ldots, S_n$, $E(X_{n+1}) = 0$ and the rules of conditional expectation given above are used. It should be observed that the sequence $\{S_n, n \geq 1\}$ is also a Markov process. In general, however, for a Markov process no moments need exist, and for a martingale only the conditioning on the first moment is required to satisfy the fair-game property. For Markov processes, the property is related to the structure of the whole (conditional) distribution. Thus, there are processes that belong to both martingale and Markov classes (the symmetric random walk is an example, and the above one is also a walk), but these classes are quite distinct and have extensive individual theories.

One of the basic questions here is the behavior of a martingale for large n. Does the game have a limit? We state the result, which applies to all the three types of martingales. The proof is somewhat intricate and cannot be discussed here.

3. Theorem

Let $\{X_n, n \geq 1\}$ be a sequence of random variables on a probability space, such that $E(|X_n|) \leq K_0 < \infty$; that is, the expected values are all bounded by K_0. If the sequence is a martingale or a sub- or supermartingale, then $X_n \to X_\infty$ with probability 1, and $E(|X_\infty|) \leq K_0$.

An application of this result is presented to illustrate its potential use. Suppose that a sequence of observations $Y_1, Y_2, \ldots$ is taken on an experiment. The observations may be dependent. It is assumed here, from prior knowledge, that these are governed by one of two probability measures P_1 and P_2 and it is desired to know from the observations which measure is more likely to govern the phenomenon. For instance, P_1 may be normal and P_2 Cauchy. In fact, both of these spread on the whole space and have single humped densities, but the normal one has thin tails, whereas the Cauchy has relatively thick tails. The above theorem helps to settle the correct underlying probability.

Let F_n and G_n be the n-dimensional distributions of the first n observations $(X_1, \ldots, X_n)$ for the probabilities P_1 and P_2, respectively, having positive densities so that

$$\begin{aligned} P_1(X_1 < x_1, \ldots, X_n < x_n) &= F_n(x_n, \ldots, x_n) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_n(t_1, \ldots, t_n)\, dt_1 \ldots dt_n \\ P_2(X_1 < y_1, \ldots, X_n < y_n) &= G_n(y_1, \ldots, y_n) \\ &= \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_n} g_n(u_1, \ldots, u_n) du_1 \cdots du_n \end{aligned}$$

and

$$f_n > 0, \quad g_n > 0.$$

Consider

$$\begin{aligned} Y_n(\omega) &= L_n(X_1(\omega), \ldots, X_n(\omega)) \\ &= \frac{g_n(X_1(\omega), \ldots, X_n(\omega))}{f_n(X_1(\omega), \ldots, X_n(\omega))}. \end{aligned}$$

The function Y_n (or L_n) is called the likelihood ratio for the probability densities of P_1 and P_2, and it plays an essential role in statistical decision theory. Since $f_n > 0$, $g_n > 0$, Y_n is well defined and is a random variable. Also by Eq. (10),

$$\begin{aligned} Ep_1(Y_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} & \frac{g_n(X_1, \ldots, x_n)}{f_n(x_1, \ldots, x_n)} \\ & \times f_n(x_1, \ldots, x_n)\, dx_1 \cdots dx_n = 1 \quad n \geq 1. \end{aligned}$$

Similarly $E_{p_2}(1/Y_n) = 1$. If we show that $\{Y_n, n \geq 1\}$ is a martingale sequence, then the above theorem implies that $Y_n \to Y$ with P_1 probability 1, and similarly $1/Y_n \to \tilde{Y}$ with P_2 probability 1. Hence, if $(\Omega, \mathcal{A}, P_1)$ is the correct model, then Y_n tends to be small for large n. The relevant statistical procedure, which is an extension by U. Grenander of a classical result due to J. Neyman and E. S. Pearson, is to reject the model with P_1 (hence, accept P_2) if, for large n, $Y_n(\omega) > \alpha$, where α is chosen such that $P_1[Y_n > \alpha] \leq 0.05$ (or 0.01). Thus, our problem hinges on the fact that $\{Y_n\}_1^\infty$ forms a martingale sequence. This property is verified as follows:

$$\begin{aligned} & E_{p_1}(Y_{n+1} \mid Y_1, \ldots, Y_n)(\omega) \\ &= E_{p_1}(Y_{n+1} \mid X_1, \ldots, X_n)(\omega) \\ &= \int_{-\infty}^{\infty} \frac{g_{n+1}(x_1, \ldots, x_n, u)}{f_{n+1}(x_1, \ldots, x_n, u)} \\ &\quad \times \tilde{P}_1(du \mid X_1, \ldots, X_n)(\omega) \\ &= \int_{-\infty}^{\infty} \frac{g_{n+1}(x_1, \ldots, x_n, u)}{f_{n+1}(x_1, \ldots, x_n, u)} \\ &\quad \times \frac{f_{n+1}(x_1, \ldots, x_n, u)}{f_n(x_1, \ldots, x_n)}\, du, \\ &= \frac{1}{f_n(x_1, \ldots, x_n)} \int_{-\infty}^{\infty} g_{n+1}(x_1, \ldots, x_n, u)\, du \end{aligned}$$

$$= \frac{g_n(x_1, \ldots, x_n)}{f_n(x_n, \ldots, x_n)} = Y_n(\omega).$$

Here we used a form of conditional probability:

$$\tilde{P}_1(u \mid X_1, \ldots, X_n)(\omega) = (\text{defined as}) \int_{-\infty}^{u} \frac{f_{n+1}(x_1, \ldots, x_n, t)}{f_n(x_1, \ldots, x_n)}\, dt,$$

where $x_k = Y_k(\omega)$. Thus, $E(Y_{n+1} \mid Y_1, \ldots, Y_n) = Y_n$ and the martingale property is verified. If f_n is not strictly positive, one can show that the Y_n sequence forms a supermartingale, and the same conclusion holds. This is an indication of uses of the martingale concept, and the theory is rich in applications to sequential analysis, differentiation theory, mathematical finance, and others. We next consider another set of dependence classes.

VI. STATIONARY AND HARMONIZABLE PROCESSES AND SPECTRA

The preceding dependence class demanded the existence of one moment for the random variables. Here another family, which assumes two moments for each random variable, is discussed. It plays an important role in electrical and communication engineering problems, among others. We consider $\{X_n, n = 0, \pm 1, \pm 2, \ldots\}$ as a doubly infinite sequence capable of representing all the past and present and observing the future of an experiment as it progresses. These random variables X_n have two moments but need not be independent. Some terminology should be introduced here.

For simplicity, let $E(X_n) = 0$. Then $\sigma_n^2 = E(|X_n|^2)$ is the variance, and $r(m, n) = E(X_m \bar{X}_n)$ is the covariance. We admit the possibility of complex values, motivated by the above-stated applications. Thus, $X_n = U_n + iV_n$, where U_n, V_n are real random variables with means zero and finite second moments. Also $\rho(m, n) = r(m, n)/\sigma_m \sigma_n$ is the correlation between X_m and X_n and is a measure of the dependence. If $\rho(m, n) = 0$ or equivalently $r(m, n) = 0$, then X_n and X_m are uncorrelated; when they have zero means, as here, they are also *orthogonal*. If $r(m, n) = \tilde{r}(m - n)$ so that the covariance is a function only of the difference of their positions, then the sequence is called *stationary*, a concept introduced by A. Khintchine in 1934. A simple example of a stationary sequence is as follows. If $Y_1, Y_2, \ldots, Y_n$ are independent with zero means and unit variances, the random variables (for a fixed n),

$$X_m = \sum_{k=1}^{n} Y_k e^{ikm}, \qquad m = 0, \pm 1, \pm 2, \ldots$$

form a stationary sequence, since $E(X_m) = 0$, and

$$r(m, m + l) = E(X_m \bar{X}_{m+l}) = \sum_{k=1}^{n} e^{-ikl} \quad [= \tilde{r}(l)]$$

depending only on l. Note that $\tilde{r}(l) \neq 0$ so that the X_m sequence is correlated. This class is again large, since one can have stationary normal (or Gaussian) processes and even a stationary Markov normal process and others. It also has an interesting and useful structure theory, some of which will be explained here.

If $\{X_n, n = 0, \pm 1, \pm 2, \ldots\}$ is a stationary process, as defined above, with a covariance function $r(\cdot)$, it is positive definite in the sense that for each set $\{\alpha_1, \ldots, \alpha_n\}$ of complex numbers

$$\sum_{k=1}^{n} \sum_{j=1}^{n} \alpha_k \bar{\alpha}_j r(k - j) = \sum_{k=1}^{n} \sum_{j=1}^{n} \alpha_k \bar{\alpha}_j E(X_k \bar{X}_j)$$
$$= E\left(\left(\sum_{k=1}^{n} \alpha_k X_k \right) \left(\sum_{j=1}^{n} \alpha_j X_j \right) \right)$$
$$= E\left(\left| \sum_{k=1}^{n} \alpha_k X_k \right|^2 \right) \geq 0. \tag{35}$$

Such positive-definite functions $\{r(l), l = 0, \pm 1, \pm 2, \ldots\}$ have been studied in mathematical analysis for a long time. They can be characterized by a classical 1911 result of G. Herglotz. Thus, r is uniquely representable as

$$r(k) = \int_{-\pi}^{\pi} e^{ik\lambda}\, dG(\lambda), \qquad k = 0, \pm 1, \ldots, \tag{36}$$

where G is a nonnegative, nondecreasing, left-continuous, and bounded function on $(-\pi, \pi]$. For this it is essential that the index set be *all* integers. Then r is the Fourier coefficient of such a G, the integral being the usual Stieltjes type. Here G is called the *spectral function* of the stationary process, and it has other physical interpretations in electrical engineering problems. Representation (36) allows one to use many classical results. We present an application to a filtering problem.

Let $\{X_n, Y_n; n = 0, \pm 1, \pm 2, \ldots\}$ be a pair of stationary series connected by the difference (also termed an autoregressive) equation

$$\Lambda X_n = X_n + a_1 X_{n-1} + \cdots + a_k X_{n-k} = Y_n, \tag{37}$$

where the a_j are constants. The X_n series is called an input and the Y_n series an output, and the linear operation Λ is termed a polynomial filter. A natural problem here is as follows: If the output is known, for what types of filters can the input be recovered? A method for solving this problem is indicated below:

If G_x and G_y are the spectral functions of the X_n and Y_n series, then Eqs. (36) and (37) give, for the covariances r_x and r_y of the series,

$$r_y(l) = E(Y_{n+l}\bar{Y}_n) = E[(\Lambda X_{n+l})(\overline{\Lambda X_n})]$$
$$= \sum_{j=0}^{k}\sum_{j'=0}^{k} a_j \bar{a}_{j'} E(X_{n-j+l}\bar{X}_{n-j'}), \quad a_0 = 1$$
$$= \sum_{j=0}^{k}\sum_{j'=0}^{k} a_j \bar{a}_{j'} r_x(j'-j+l)$$
$$= \int_{-\pi}^{\pi} \left(\sum_{j=0}^{k} a_j e^{-ij\lambda}\right)\left(\sum_{j=1}^{k} \bar{a}_j e^{ij\lambda}\right) e^{il\lambda} dG_x(\lambda). \tag{38}$$

If $p_k(z) = \sum_{j=0}^{k} a_j z^j$, the kth-order polynomial determined by the a_j's, then Eq. (38) becomes

$$\int_{-\pi}^{\pi} e^{il\lambda}\, dG_y(\lambda) = \int_{-\pi}^{\pi} \left|p_k(e^{-i\lambda})\right|^2 e^{il\lambda}\, dG_x(\lambda). \tag{39}$$

Here $p_k(z)$ is called the *characteristic polynomial* of the filter equation (37) and $\phi(\lambda) = p_k(e^{-i\lambda})$ is termed the *filter characteristic*, because the zeros of ϕ govern the behavior of Λ. To explain this, it is necessary to state a relatively advanced result established independently in 1940 by H. Cramér and A. Kolmogorov. According to this, one has

$$X_n = \int_{-\pi}^{\pi} e^{-in\lambda}\, dZ_x(\lambda), \tag{40}$$

where $\{Z_x(\lambda), -\pi < \lambda \le \pi\}$ is a random process such that for each $-\pi < \lambda_1 < \lambda_2 < \lambda_3 \le \pi$, the increments $Z_x(\lambda_3) - Z_x(\lambda_2)$ and $Z_x(\lambda_2) - Z_x(\lambda_1)$ are uncorrelated, and $E(|Z_x(\lambda_3) - Z_x(\lambda_2)|^2) = G_x(\lambda_3) - G_x(\lambda_2)$. The integral in Eq. (40) is defined with a procedure of the classical Riemann method and is called a *stochastic integral*. If $Z_y(\cdot)$ is the corresponding process for the Y_n series, Eq. (37) becomes

$$\int_{-\pi}^{\pi} e^{in\lambda}\, dZ_y(\lambda) = Y_n = \Lambda X_n = \int_{-\pi}^{\pi} \phi(\lambda) e^{in\lambda} dZ_x(\lambda), \tag{41}$$

where $\phi(\lambda)$ is the filter characteristic. Using this and the location of the zeros of $\phi(\lambda)$ or of $p_k(z)$, we shall present the solution of the problem without giving all the mathematical details.

If the roots w_j of $p_k(z) = 0$ are distinct and $|w_j| > 1$ or if some $w_j = e^{i\lambda_j}$ and at such λ_j the spectral function $G_y(\cdot)$ is continuous, then the input X_n can be expressed as

$$X_n = \int_{-\pi}^{\pi} \phi(\lambda)^{-1} e^{in\lambda}\, dZ_y(\lambda) = \sum_{j\ge 0} b_j(a_1, \ldots, a_k) Y_{n-j}, \tag{42}$$

where the b_j are constants determined by $a_1, \ldots, a_k$ [or equivalently the roots of $p_k(z)$], and the series converges in a well-defined (called a mean-square) sense. The point of the location of the roots is that X_n is obtainable completely from the past and present of the output, if they are outside the unit circle. In that case Λ is termed *physically realizable*. If the roots do not satisfy these restrictions, then, when the solution is obtainable, it involves future values in addition so that the filter is *not* physically realizable.

Most of the above questions, detailed for stationary sequences, can be considered for a closely related nonstationary class called harmonizable processes. A brief account of the latter will be included since the formulas obtained for the stationary case hold, essentially in the same form, for the general setup, thereby giving a sort of "robustness" for these results.

Thus, a sequence $\{X_n, n = 0, \pm 1, \pm 2, \ldots\}$ is termed *strongly* [*weakly*] *harmonizable* if each X_n admits a representation as a stochastic integral (40), with a fixed $\{Z_x(\lambda), -\pi < \lambda \le \pi\}$, which is a random process as before but which need not have orthogonal increments. It satisfies (i) $E(Z_x(\lambda)) = 0$, and (ii) if $E(Z_x(\lambda)\overline{Z_x}(\lambda')) = F_x(\lambda, \lambda')$ then F_x has finite Vitali [Fréchet] variation in the following sense. For any (standard) subsets A, B of $(-\pi, \pi]$

$$\text{Vitali var}(F_x)(A, B) = |F_x|(A, B)$$
$$= \sup\left\{ \sum_{i=1}^{n}\sum_{j=1}^{n} |\, F_x(A_i, B_j)\, | : \{A_i\}_1^n, \{B_j\}_1^n, \right.$$
$$\left. \text{disjoint subsets of } A \text{ and } B,\ n < \infty \right\} < \infty.$$

and

$$\text{Fréchet var}(F_x)(A, B) = \|F_x\|(A, B)$$
$$= \sup\left\{ \left| \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \overline{b_j} F_x(A_i, B_i) \right| : \{A_i\}_1^n, \{B_j\}_1^n \right.$$
$$\left. \text{as above, and } |a_i| \le 1, |b_j| \le 1 \right\} < \infty.$$

It is clear that $\|F_x\|(A, B) \le |F_x|(A, B)$. There is strict inequality for the weakly harmonizable case since then $|F_x|(T, T) = +\infty$, but $\|F_x\|(T, T) < \infty$ always where $T = (-\pi, \pi]$. Thus, a (weakly or strongly) harmonizable process is the Fourier transform of a certain process with two moments finite. One can verify that the covariance $r_x(m, n) = E(X_m \overline{X_n})$ admits a representation

$$r_x(m, n) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{im\lambda - in\lambda'}\, dF_x(\lambda, \lambda'),$$
$$m, n = 0, \pm 1, \ldots. \tag{36'}$$

The function F_x is called the *spectral bi-measure* of the X_n sequence in both the weak and strong cases. In the stationary case, F_x concentrates on the diagonal of the square

$T \times T$, so that the harmonizability concept reduces to the stationary one. [For the weak harmonizability work, some care is needed regarding the integral in Eq. (36). The relevant theory of this case and its comparison with the stationary aspects, in all its detail, can be found in Chang and Rao (1986). Leaving these technicalities aside, one can proceed with the analogs of the above case.]

It is now possible to obtain the corresponding (polynomial) filter equations. Let us sketch the formulas related to Eq. (38) for the (strongly) harmonizable case. If the autoregressive or polynomial filter equation is given by Eq. (37), then Eq. (41) holds without change. If $\phi_n(\lambda)$ is the filter characteristic, then the analog of Eq. (38) becomes

$$r_y(m,n) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \phi_m(\lambda)\overline{\phi_n}(\lambda')e^{im\lambda - in\lambda'}\, dF_x(\lambda,\lambda').$$

When all the roots of $\phi_n(\lambda) = 0$ lie outside of the unit disc, then X_n can be represented as a series (42) in which the $Z_y(\cdot)$ of the Y_n sequence need not have orthogonal increments any longer.

The preceding discussion for harmonizable and stationary sequences has analogs for continuous parameter processes. Instead of the sequence $\{X_n, n = 0, \pm 1, \pm 2, \ldots\}$, one considers $\{X_t, t \in \mathbb{R}\}$, and $T = (-\pi, \pi]$ is replaced by $\mathbb{R}$ (the dual additive group of $\mathbb{R}$, T being the dual of the integers $\mathbb{Z}$). Thus when $E(X_t) = 0$, the covariance $r(s,t) = E(X_s\bar{X}_t)$ is given by $r(s,t) = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{isu-itv}\, dF(u,v)$ and this reduces in the stationary case with $r(s,t) = r(s-t) = \int_{\mathbb{R}} e^{i(s-t)u}\, d\tilde{F}(u)$ where $\tilde{F}$ concentrates on the diagonal of $\mathbb{R} \times \mathbb{R}$, the spectral function. Then there are the corresponding statements of all the results noted above for the sequences, whose details are, for instance, in Change and Rao (1986).

Every harmonizable process is a projection of some stationary process. Thus each such process has a "stationary dilation." This intrinsic property is useful for the structural analysis of harmonizable classes. The weakness of this result is that the dilated process exists but is not usually available to the experimenter, and hence cannot be implemented in applications. However, an important consequence of this property is that each continuous linear transform of a stationary or harmonizable process is always (weakly) harmonizable, so that the last class is closed under such mappings. This fact is important for the theoretical study of these processes. [For details, see Chang and Rao (1986).]

VII. SOME REMARKS ON THE CENTRAL LIMIT PROBLEM

The preceding discussion shows that a major part of probability theory is devoted to the asymptotic analysis of various sequences of random variables. This actually started with the Bernoulli law of large numbers and the Laplace–de Moivre central limit theorem. From an applicational point of view, these results can be used only as an approximation, using a finite but large number n of observations. Then one should know the possible errors of approximation in such applications. It was P. L. Čebyšev who first emphasized the importance of this problem and obtained some useful bounds for the probabilities of these errors with his inequality. These results were refined and extended to more general types of (still independent) random variables by his students, A. A. Markov and A. M. Liapounov, the latter obtaining the most comprehensive from by 1901 of the central limit theorem, which as noted before is usually regarded as the basic limit theorem of probability theory. For the first time Liapounov employed the then new technique of Fourier transforms in his work, and it quickly became a standard tool. Before this, Markov had introduced a useful truncation method but it was not as powerful as the former. However, progress on the error estimation problem has been slow. Here we present a general result on the error estimation in the central limit problem, which was obtained independently by A. C. Berry in 1941 and in a somewhat more detailed manner by G. C. Esséen in 1945, as follows.

A. Theorem (Berry–Esséen)

Let $X_1, X_2, \ldots$ be independent random variables with three moments finite. If μ_n, σ_n^2, and β_n^3 are the mean, variance, and third absolute central moment, respectively, of X_n, and

$$S_n = \sum_{k=1}^{n} X_k; \quad \sigma^2(S_n) = \sum_{k=1}^{n}\sigma_k^2; \quad \beta_n^3(S_n) = \sum_{k=1}^{n}\beta_k^3$$

then we have ($\beta_k^3 = E(|X_k - \mu_k|^3)$)

$$\left| P\left[\frac{S_n - E(S_n)}{\sigma(S_n)} < x\right] - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\, dt \right| \le C_0[\beta_n(S_n)/\sigma_n(S_n)]^3 \tag{43}$$

for all $-\infty < x < \infty$ with an absolute constant $0 < C_0 < \infty$ and $n \ge 1$. If all X_n have the same distribution so that $\mu_n = \mu, \sigma_n^2 = \sigma^2$, and $\beta_n^3 = \beta^3$, then Eq. (43) solves the classical central limit problem,

$$\left| P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} < x\right] - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\, dt \right| < \frac{C_1}{\sqrt{n}}\left(\frac{\beta}{\sigma}\right)^3 \tag{44}$$

for all $-\infty < x < \infty$ and some constant $0 < C_1 < \infty$, $n \ge 1$.

A considerable effort has been made to determine the best constants C_0, C_1 in Eqs. (43) and (44). In 1949

H. Bergström showed that $C_0 \leq 4.8$, and after many improvements by various people, V. M. Zolotarev showed in 1966 that $C_1 \leq 1.321312$. These are not yet the best possible bounds. The importance of this theorem lies in the fact that the size of the right-side bounds, which are uniform in x, can be made smaller than a prescribed level by choosing a suitably large n. Then one can use the normal distribution for $[S_n - E(S_n)]/\sigma(S_n)$ as a satisfactory approximation.

It is clear that in Eq. (43) we have to demand that $(\beta(S_n)/\sigma(S_n)) \to 0$ as $n \to \infty$ for the limit theorem. This is automatic in Eq. (44). It can be shown that the rate of the right side of Eq. (44) is optimal. Also, as Liapounov himself showed, in bound (43), $2 + \delta$ moments suffice for some $\delta > 0$ but not $\delta = 0$. The proof of this theorem is intricate, and this is one reason the error estimation problem for other results is slow, and computer simulations are often utilized when no such theorem as above is available.

VIII. STATISTICAL INFERENCE PROBLEMS AND COMMENTS

In most work on probability, distributions of random variables are assumed to be known from prior knowledge. These functions typically contain some parameters that are not always precisely known. For instance, in a binomial distribution the probability of success p is usually not completely known since one can not be certain of the amount of bias a coin has. Similarly, in a normal distribution the parameters are the mean μ and variance σ^2, which specify the distribution. They may not be known exactly. In the case of linear filters described by Eq. (37) the structural coefficients $a_1 \ldots, a_n$ are not always available to the experimenter, and the covariance functions or equivalently their spectral functions in Eq. (36) need not be known before hand. All these parameters have to be estimated from the observed process. Even if one can assume these structural parameters, it will often be necessary to ascertain or test the validity of the prior assumptions. These problems are nontrivial, and they are essential for a successful application of probability. The aspect of probability theory dealing with these questions is inference theory, and the questions have to be answered from sampled data, for which several new criteria should be formulated. All of this belongs to what is called statistical inference theory, which forms a second step in the analysis after the basic modeling aspect of probability theory is formulated. A very brief sketch of the new questions will be given here.

The estimation of parameters can often be accomplished using the principle of least squares. This principle was first published by A. M. Legendre in 1806 but apparently had been known to C.-F. Gauss since 1795, although the latter did not publish it until 1809. Several other principles suitable for other occasions have since been devised. Some of these are maximum likelihood, minimum chi-square, minimax, and regret principles. These ideas, belonging to statistical inference, and more generally to decision theories, grew into a chapter in that subject. Thus, after finding estimators by means of the various methods, one must compare them on the basis of some other principles of "goodness" characteristics. Except in simple situations, here again asymptotic analysis and hence limit theorems of probability theory play a critical part. Often such questions lead to new studies in probability, and in fact many of the extensions of the work from mutually independent random variables to the dependent classes discussed in earlier sections arose from these applications. General limit theorems for such dependent classes as Markov processes, martingales, and stationary processes have achieved significant successes.

There are other problems in which the unknown parameters of a distribution cannot be regarded as absolute constants, especially when the repeatability of an experiment is not feasible because of cost and other considerations. For instance, testing of a space rocket is not feasible for several repeated trials because of cost. In such cases the distributions can be taken as conditional, and the parameters become the values of the conditioned random variables. Indeed, A. Rényi proposed in 1955 a conditional model in lieu of the $(\Omega, \mathcal{A}, P)$ considered thus far. This may be used in some problems. In treating the parameters as values of random variables with completely known distributions of their own, one can use the Bayes formula [Eq. (17)] and convert the problem into one of absolute probabilities. A price paid for this is to turn the independent observations of the conditional model into certain dependent ones in the resulting (higher-order) model. For some of the problems of the type noted above, this Bayesian analysis gives satisfactory results. Asymptotically, however, many of these formulations give the same results. These questions properly belong to inference theory.

At the classical finite sampling level a reference to Lehmann (1959) suffices, and at the stochastic processes level, one has to go further and refer to Grenander (1981), Hall and Heyde (1980), Liptser and Shiryayev (1977), and Rao (2000).

We shall conclude this article by pointing out that probability theory is nurtured by the most interesting problems of natural sciences and other human activities that are precisely describable but generally depend on chance. The measurement of uncertainty is partly an art, as is a construction of the underlying probability model. The twin questions of randomness and interpretation of probability statements have been set aside in the modern axiomatic

foundations of the subject. Practical implications of these questions are important. Attempts to define randomness are not so far satisfactory, since perhaps they are somewhat similar to defining "nature." The difficulty and the unsuccessful efforts can be seen from the recent (non probabilistic) popularizing account in Beltrami (1999). But practioners of the subject can and do proceed with the firmly founded axiomatic setup of Kolmogorov's with confidence in its theoretical conclusions.

Further details and proofs of all the statements made in the preceding sections may be found in the specialized literature. A representative cross-section mainly since the 1950s from introductory to relatively advanced levels, treating these subjects extensively, together with specific sources noted above, is given in the bibliography that follows.

SEE ALSO THE FOLLOWING ARTICLES

NUMBER THEORY, ELEMENTARY • STATISTICAL ROBUSTNESS • STATISTICS, FOUNDATIONS • STATISTICS, MULTIVARIATE • STATISTICS, NON-PARAMETRIC • STOCHASTIC PROCESSES

BIBLIOGRAPHY

Beltrami, E. (1999). "What is Random? (Chance and Order in Mathematics and Life)," Copernicus-Springer, New York.

Borwein, J. M. (1998). "Brouwer-Heyting sequences converge," *The Mathematical Intelligencer* **20,** 14–15.

Chang, D. K., and Rao, M. M. (1996). Bimeasures and nonstationary processes. *In* "Real and Stochastic Analysis," pp. 7–118, Wiley-Interscience, New York.

Chung, K. L. (1979). "Elementary Probability Theory and Stochastic processes," 3rd ed. Springer, New York.

De Groot, M. H. (1986). "Probability and Statistics," 2nd Ed., Reading, Massachusetts.

Doob, J. L. (1953). "Stochastic Processes," Wiley, New York.

Galambos, J. (1984). "Introductory Probability Theory," Marcel Dekker, New York.

Grenander, U. (1981). "Abstract Inference," Wiley-Interscience, New York.

Gunning, R. C., and Rossi, H. (1965). "Analytic Functions of Several Complex Variables," Prentice-Hall, Englewood Cliffs, NJ.

Hall, P., and Heyde, C. C. (1980). "Martingale Limit Theory and Its Applications," Academic Press, New York.

Kolmogorov, A. N. (1933). "Foundations of the Theory of Probability," Springer, Berlin (English Translation, 1956), Chelsea, New York.

Lehmann, E. L. (1959). "Testing Statistical Hypotheses," Wiley, New York.

Linnik, Ju. V. (1965). "An application of a theorem of H. Cartan to mathematical statistics," *Soviet Math. Dokl.* **6,** 291–293.

Linnik, Ju. V. (1968). "Statistical Problems with Nuisance Parameters," American Math. Soc, Providence, RI.

Liptser, R. S., and Shiryayev, A. N. (1977). "Statistics of Random Processes-I,II," Springer, New York.

Priestley, M. B. (1982). "Spectral Analysis and Time Series," Academic Press, London.

Rao, M. M. (1981). "Foundations of Stochastic Analysis," Academic Press, New York.

Rao, M. M. (1984). "Probability Theory with Applications," Academic Press, New York.

Rao, M. M. (1993). "Conditional Measures and Applications," Marcel Dekker, New York.

Rao, M. M. (1995). "Stochastic Processes: General Theory," Kluwer Academic, Dordrect, The Netherlands.

Rao, M. M. (2000). "Stochastic Processes: Inference Theory," Kluwer Academic, Dordrect, The Netherlands.

Rozanov, Yu. A. (1982). "Markov Random Fields," Springer, New York.

Sanov, I. N. (1957). "On the probability of large deviations of random variables," *Mat. Sbornik* **42**(86), 11–44. English Translation in "Selected Translations in Mathematical Statistics and Probability," (1961). AMS **1,** 213–224.

Varadhan, S. R. S. (1984). "Large Deviations and Applications," CBMS-NSF, SIAM Publications, Philadelphia.



Queueing Theory

H. M. Srivastava
University of Victoria

I. Fundamental Concepts
II. Poisson and Non-Poisson Queues
III. Queues with Variable Parameters
IV. Double-Ended Queues
V. Birth-and-Death Queueing Systems
VI. Imbedded Markov Chain Queueing Systems
VII. Equations for the M/M/1 Queueing System
VIII. Examples of Applications
IX. Further Developments in Related Areas

GLOSSARY

Balking Queueing phenomenon in which a prospective customer decides not to join a queue (and is lost to the system) after looking at the size of the queue and estimating the time he or she may have to wait before receiving service.

Generating function (or *Z* transform) Function $P(z, t)$ of two variables defined usually by

$$P(z, t) = \sum_{n=0}^{\infty} P_n(t) z^n$$

where z is a dummy variable and $P_n(t)$ denotes, for example, the probability that there are n units in the system at time t.

Laplace transform Integral transform $p_n(s)$ of the probability distribution function $P_n(t)$, defined by

$$p_n(s) = \int_0^{\infty} e^{-st} P_n(t)\, dt$$

whenever the improper integral exists, s being a dummy parameter.

Markov chain Markov process with a discrete state space.

Markov process Process with no memory extending before the previous instant (i.e., a process whose future probabilistic evolution after any time t depends only on the state of the system at time t and is independent of the history of the system before the time t).

Poisson distribution Probability distribution with density function $\mu^n e^{-\mu}/n!\,(n = 0, 1, 2, \ldots)$.

Poisson queue Queueing model in which the number of arrivals in a given duration of time follows the Poisson distribution and the service times follow the negative-exponential distribution (i.e., a probability distribution with density function $\mu e^{-\mu W}$).

Probability density function (or frequency function) Derivative $dF(x)/dx$ of a probability distribution function $F(x)$.

Probability distribution function (or cumulative probability) Function $F(x)$ of a random variable X defined, for all possible real numbers x ranging from $-\infty$ to ∞, to be the probability that $X \leqq x$, that is, $F(x) = \text{Prob}\{X \leqq x\}$.

Random variable Variable whose value depends on the outcome of a random experiment, such as the tossing of a coin, one play of a game of blackjack in Las Vegas, and so on.

Regeneration point Time epoch (instant) t at which the characteristic Markov property holds true; that is, the future probabilistic evolution of the given process depends only on the state of the process at time t, a statement of the history of the process having no predictive value.

Reneging Queueing phenomenon in which a customer, after having joined a queue, becomes impatient with waiting and leaves the queue before service starts (and is thus lost to the system).

Stochastic process Possible actual (e.g., physical) process of the real world that has some random (or stochastic) element involved in its structure.

QUEUEING THEORY (or waiting-line theory) is an important branch of applied probability theory and is used to study service systems that are prone to congestion. The subject was nonexistent 70 years ago, and it had no name 50 years ago. Today the theory of queues has an honored place among various mathematical models that represent and predict operations; indeed, queueing theory is one of the first to have been developed systematically. Some of the common applications of queueing theory range from day-to-day bank, department store, airport, and hospital queues to such industrial queues as inventories, assembly (or production) lines, and computer time-sharing systems.

I. FUNDAMENTAL CONCEPTS

The subject of queueing theory has grown out of the pioneering work done in this field by Agner Krarup Erlang (1878–1929), a Danish telephone engineer. In the first decade of the 20th century, Erlang investigated the effects of fluctuations in demand on the operation of Copenhagen's early telephone system and thus developed the first queueing model ever recorded (and also the foundation of the mathematical theory of traffic). It may be of historical interest that Erlang's queueing model antedates Lanchester's model of air warfare by more than 10 years and precedes von Neumann's model for competition by more than 30 years.

Queueing theory is a standard mathematical tool for workers in various areas of operations research. The range of applications of queueing theory has widened as operations research has extended its coverage. Typical queueing theory applications in operations research include those dealing with (1) problems of air transport, for example, the stacking of aircrafts waiting to land or the complications of passenger reservations; (2) problems in public service operations, such as hospital operation; (3) problems involving inventories, such as the spare-parts inventory of the U.S. Army; (4) problems involving the storage of water behind dams that have the potential to improve the planning and operation of hydroelectric power systems; (5) problems involving the maintenance and servicing of equipment and machines; (6) problems of rail, truck, and sea transport, such as the marshaling of yards in rail or truck transport; (7) problems in automobile traffic involving queues of cars in front of traffic lights or tollbooths, at entrances to major highways, in tunnels or behind slow cars on highways, at international border crossings, and so on; (8) queueing problems associated with banks, department stores, law courts, libraries, post offices, telephone and telex companies, and so on; and (9) problems in experimental sciences, such as experimental physics in which, for instance, one is required to correct the observed readings of elementary-particle detectors when counting at a high speed. There are examples of numerous other queueing problems that one naturally encounters in one's daily life. Thus, queueing theory has vast applications, ranging from day-to-day department store, airport, and hospital queues to industrial queues such as inventories, assembly (or production) lines, and computer time-sharing systems. In some cases we even encounter nonphysical queues like the names of candidates, say, for a job, which do indeed constitute a queue on paper.

The proliferation of queueing theory in the 75 years of its development has been rather extensive. Not only has it been immensely enriched by the opening of a wide variety of applications, but it has also been extended considerably in order to handle situations in which no queues exist. Obviously, therefore, these important extensions of queueing theory have rendered its present name somewhat anachronistic, and but for the verbal simplicity of the name *queueing theory,* the subject should be accorded a more appropriate title, such as the theory of stochastic service systems.

A. Queueing Systems

A queueing system can be described as a system having a service facility at which units of some kind (generically called "customers") arrive for service; whenever there are more units in the system than the service facility

can handle simultaneously, a queue (or waiting line) develops. The waiting units take their turn for service according to a preassigned rule, and after service they leave the system. Thus, the input to the system consists of the customers demanding service, and the output is the serviced customers. A queueing system is usually characterized by the following terms:

1. *The input process.* Let the customers arrive at the instants $t_0, t_1, t_2, \ldots$; then the interarrival times are

$$u_r = t_r - t_{r-1}, \qquad r = 1, 2, 3, \ldots. \tag{1}$$

The random variables u_r are, in general, assumed to be statistically independent, and their probability distribution $A(u)$ is called the interarrival time distribution or, simply, the arrival distribution or input distribution. The customers may come from an infinite source, as in the case of telephone calls and as is assumed in many queueing studies, or from a finite source. They may arrive singly or in groups of fixed or various sizes. The queueing system may have an upper limit on the number that can be admitted into the system, as in the case of finite waiting space.

2. *The queue discipline*. This can be described as the rule determining the formation of a queue or queues and the manner in which a customer or customers are selected for service from those waiting. The most common queue discipline is "first come, first served," according to which the units enter service in order of their arrival. Other possibilities are random selection for service, a priority rule, or even the "last come, first served" rule. In the case of priority, there may be two classes, namely, the priority class and the nonpriority class, or there may be several priority classes representing different levels of priority. Furthermore, there may be a preemptive priority discipline according to which a lower-priority unit is taken out of service whenever a higher-priority unit arrives, the service on the preempted unit resuming only when there are no higher-priority units in the system. Contrary to this is the nonpreemptive, or the head-of-the-line, priority rule, in which priorities are taken into account only at the commencement of service, and once started, the service is continued until completion. It is customary to include under this heading queueing phenomena, such as balking and reneging, depicting the behavior of the waiting customers. Customers are said to balk when, looking at the size of the queue and then estimating the time they may have to wait before service, they do not join the queue. After joining a queue, customers are said to renege if they become impatient with waiting and leave the queue before service starts.

3. *The service mechanism.* The time that elapses while a unit is being served is called its service time. The service times $v_1, v_2, v_3, \ldots$ of the successive units are assumed to be independent of one another and of the input distribution, and their probability distribution $B(v)$ is called the service-time distribution or, briefly, the service distribution. The specification of service mechanism includes the number of servers. Thus, there are single-server and multiserver systems (sometimes called single-channel and multichannel from telephone parlance). Again, in a bulk service system, service may be provided in batches of fixed or various sizes.

As is common in probability theory, if we can write

$$dA(u) = a(u)\,du; \qquad dB(v) = b(v)\,dv, \tag{2}$$

then $a(u)$ and $b(v)$ are, respectively, the interarrival time and service-time density functions.

B. Queue Parameters

For an understanding of various queueing systems and for their efficient management, the following three queue parameters are generally investigated.

1. *Queue length*. This is the number of units waiting in a queue or present in a system. In the latter case it is sometimes called the system size. The determination of the probability distribution of this discrete random variable is important for the design of the system—for example, for providing an extra server or a larger waiting space in the case of large queues.
2. *Waiting time*. This is the time a customer has to wait in a queue before being served. Sometimes the service time is also included in the waiting time. The terms *queueing time* and *waiting time* are also used depending on whether the service time is excluded or included. This evidently is a continuous variable, and its probability distribution is important from the customer's point of view. Actually, for some queueing systems (e.g., M/M/1), the queueing time is *not* continuous for $t = 0$.
3. *Busy period*. A busy period starts with the arrival of a customer who finds the system empty and ends when the system next becomes empty. However, the initial busy period may start with any number of units in the system, and for multiserver systems we can define a busy period to be the time interval during which a given number k are busy throughout. The probability distribution of the duration of a busy period is important from the server's point of view.

C. Notation

We shall follow Kendall's notation and specify a queueing system by a triple $\cdot/\cdot/\cdot$, in which the first two constituents are letters representing the form of input and

service-time distributions, respectively, and the third is a number denoting the number of servers. The following letters are used for signifying distributions:

1. M (standing for Markov) for the negative exponential interarrival or service-time distribution:

$$A(u) = 1 - e^{-\lambda u}, \qquad 0 \leqq u < \infty \tag{3}$$

or

$$B(v) = 1 - e^{-\mu v}, \qquad 0 \leqq v < \infty \tag{4}$$

according to whether M occurs in the first or the second place in the triple. From Eq. (3), it follows that the mean interarrival time is

$$\int_0^\infty u \, dA(u) = \frac{1}{\lambda}. \tag{5}$$

Similarly, from Eq. (4), the mean service time is $1/\mu$. Alternatively, we may say that λ and μ are, respectively, the mean arrival and service rates per unit time. It can easily be seen, and is commonplace in elementary probability theory, that Eq. (3) is equivalent to saying that the number of arrivals in a given duration of time follows the Poisson distribution:

$$\text{Prob}\{n \text{ arrivals in time } t\} = (\lambda t)^n e^{-\lambda t}/n! \qquad n = 0, 1, 2, \ldots. \tag{6}$$

Consequently, very often Eq. (3) is spoken of as Poisson arrivals or Poisson input. Similarly, Eq. (4) implies that the number of services rendered in a given time duration is Poisson distributed (only while the server is busy).

2. D (standing for deterministic) for constant interarrival or service times:

$$A(u) = \begin{cases} 0, & \text{if } u < 1/\lambda \\ 1, & \text{if } u \geqq 1/\lambda \end{cases} \tag{7}$$

or, equivalently, the density function

$$a(u) = \delta(u - 1/\lambda), \tag{8}$$

where $\delta(x)$ is the Dirac delta function. The situation is similar for $B(v)$ if D is in the second place.

3. E_k for the k-Erlang distribution:

$$a(u) = \{\lambda^k/(k-1)!\}u^{k-1}e^{-\lambda u}, \quad 0 \leqq u < \infty \tag{9}$$

and similarly for $B(v)$. When $k = 1$, it reduces to the negative-exponential distribution, As $k \to \infty$ such that k/λ is finite, we have the constant interarrival time case, as we shall see later.

From Eq. (9), the Laplace transform of $a(u)$ is given by [see Eq. (18)]

$$\mathcal{L}\{a(u) : s\} = [\lambda/(\lambda + s)]^k \to e^{-\Lambda s} \tag{10}$$

as $k \to \infty$ if we set $k/\lambda = \Lambda$, a finite quantity. But $e^{-\Lambda s}$ is the Laplace transform of the Dirac delta function $\delta(u - \Lambda)$, which evidently completes the proof.

4. G for general (or arbitrary) interarrival or service-time distribution. Some authors write GI for general independent input to stress independence.

Thus, for example, we write $M/E_k/1$ to denote a single-server queueing system with Poisson arrivals and k-Erlang service-time distribution, and GI/G/s for the most general case of general independent input, general service times, and s servers (channels).

Whenever it is desirable, we may use an extended notation, writing for example, $M/E_k/1{:}N$ to denote that N is the maximum allowable number in the system at any time.

D. Queue Equations

From the viewpoint of mathematical tractability, the negative-exponential interarrival or service-time distribution is the simplest, and fortunately these distributions serve as a good fit in many practical problems. We give below a mathematical model for the simplest queueing system: M/M/1. Let

$$P_n(t) = \text{Prob}\{n \text{ units in the system at time } t\}. \tag{11}$$

Now, with Poisson arrivals, we have from Eq. (6)

Prob{one arrival during the time interval

$$(t, t + \Delta t)\} = \lambda \, \Delta t \, e^{-\lambda \Delta t} = \lambda \, \Delta t + o(\Delta t). \tag{12}$$

Prob{no arrival during the time interval

$$(t, t + \Delta t)\} = 1 - \lambda \, \Delta t + o(\Delta t) \tag{13}$$

Prob{more than one arrival during the time interval

$$(t, t + \Delta t)\} = o(\Delta t), \tag{14}$$

where $o(\Delta t)$ stands for terms negligible in comparison with Δt. Similar equations, with μ in place of λ, hold true for the number of services completed during the time interval $(t, t + \Delta t)$. Thus, considering the possible transitions during the time interval $(t, t + \Delta t)$, we have, using the multiplication and addition theorems of probability,

$$P_n(t+\Delta t) = P_n(t)\{1 - \lambda \, \Delta t + o(\Delta t)\}\{1 - \mu \, \Delta t + o(\Delta t)\} + P_{n-1}(t)\{\lambda \, \Delta t + o(\Delta t)\} + P_{n+1}(t)\{\mu \, \Delta t + o(\Delta t)\} \tag{15}$$

from which, on proceeding to the limits as $\Delta t \to 0$,

$$(d/dt)P_n(t) = -(\lambda + \mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t), \qquad n = 1, 2, 3, \ldots. \tag{16}$$

It is easy to see that, for $n = 0$, we have

$$(d/dt)P_0(t) = -\lambda P_0(t) + \mu P_1(t). \tag{17}$$

Equations (16) and (17) constitute a system of differential–difference equations. Applying Laplace transforms of the form

$$\mathcal{L}\{P_n(t) : s\} = p_n(s) = \int_0^\infty e^{-st} P_n(t)\, dt$$

$$\mathrm{Re}(s) \geqq 0, \qquad n = 0, 1, 2, \ldots \tag{18}$$

the system can be reduced to a set of pure difference equations, whereas by using the generating function (or, as it is sometimes called, the Z transform)

$$P(z, t) = \sum_{n=0}^{\infty} P_n(t) z^n \tag{19}$$

it is reduced to a differential equation. Very often, we use both the transforms (18) and (19) and thus reduce the system of differential–difference Eqs. (16) and (17) to an algebraic equation.

II. POISSON AND NON-POISSON QUEUES

Most of the initial work on queueing theory was done under steady-state assumption, under which the probabilities P_n are independent of time, and therefore the derivatives on the left-hand sides of Eqs. (16) and (17) vanish, thus giving rise to difference equations that are usually easier to solve. Also, most of the work was initially done for Poisson queues, that is, queues with Poisson arrivals and negative-exponential service times. The negative-exponential distribution has the important Markovian or "forgetfulness" property: If, for instance, the service times have a negative-exponential distribution, then at any time the residual service time also has the same distribution. That is why in Eqs. (16) and (17) we did not have to take into account the time since the service on the unit being served began, which is called the elapsed service time. This is not the case with other distributions. However, for mathematical convenience, most of the other distributions have also been approached through the negative-exponential distribution.

We give below some well-known methods of queueing theory by means of which general non-Poisson queues (i.e., queues with general interarrival or service-time distributions) are analyzed.

A. Imbedded Markov Chain Technique

Suppose that we are considering an M/G/1 system. If we study the system only at the epochs when customers depart after service, then between any two such consecutive epochs the only possible transitions are customer arrivals for which the Markovian property is available. Such epochs or time points are called regeneration points, and the process considered at these epochs is called a Markov chain imbedded in the total process, which is non-Markovian. For a GI/M/1 system, for example, the arrival epochs are the regeneration points. A detailed discussion of non-Poisson queueing systems that are analyzed by the imbedded Markov chain technique is presented in Section VI.

B. Supplementary Variable Technique

According to this method, we study the system in continuous time but enhance the description of the system by including supplementary variables denoting the time since the last regeneration point. For example, for an M/G/1 queueing process we define $P_n(x, t)\, dx$ as the probability that at time t the number in the system is n and the elapsed service time of the unit in service lies between x and $x + dx$. Similarly, the GI/M/1 system can be dealt with by the inclusion of the elapsed time since the last arrival.

C. Phase Technique

This method consists of dividing the interarrival time or the service time into a number of phases, such that the time occupied by each phase has the negative-exponential distribution. Erlang used this method by regarding service to be completed in k fictitious phases, each of which has the same negative-exponential distribution, thus obtaining what is called a k-Erland distribution, which is, in fact, the k-fold convolution of the negative-exponential distribution with itself. For arrival distributions, we can in the same way postulate an arrival timing channel consisting of a number of phases. We may also consider the generalization, in which the number of phases is itself a random variable. One may, of course, exploit the concept of phases to study more complex queueing systems through the use of Markov chains.

III. QUEUES WITH VARIABLE PARAMETERS

In the theory of queues one often encounters studies of queueing systems in which both the interarrival and service times follow the same distribution, such as the negative-exponential distribution, the parameters of either or both of these distributions being variables. Some notable examples of queues with variable parameters are given below.

A. Queues with Time-Dependent Parameters

Consider a queueing system with Poisson arrivals and negative-exponential service times, the parameters of the

arrival and service-time distributions being, respectively, $\lambda(t)$ and $\mu(t)$, which are functions of the time t. If we let

$$P_n(t) \equiv P_{i,n}(t) = \text{Prob}\{N(t) = n \mid N(0) = i\} \qquad n, i = 0, 1, 2, \ldots, \tag{20}$$

where

$$N(t) = \text{number in the system at time } t, \tag{21}$$

then we shall obtain the following set of equations (known as Kolmogorov equations):

$$(d/dt)P_n(t) = -[\lambda(t) + \mu(t)]P_n(t) + \lambda(t)P_{n-1}(t) + \mu(t)P_{n+1}(t), \quad n \geqq 1 \tag{22}$$

$$(d/dt)P_0(t) = -\lambda(t)P_0(t) + \mu(t)P_1(t) \tag{23}$$

together with the initial conditions,

$$P_{i,n}(0) = \delta_{i,n}, \qquad n, i = 0, 1, 2, \ldots, \tag{24}$$

where $\delta_{m,n}$ is the Kronecker delta defined by

$$\delta_{m,n} = \begin{cases} 1, & \text{if} \quad m = n \\ 0, & \text{if} \quad m \neq n \end{cases}. \tag{25}$$

By suitably rescaling the time axis and on introducing a new set of dependent variables, this queueing problem can be solved by using a generating function like Eq. (19).

B. Queues with State-Dependent Parameters

In order to consider a queueing process with state-dependent parameters, let

$$P_{i,j}(t) = \text{Prob}\{\text{system state at time } t \text{ is } j \mid \text{state was } i \text{ at } t = 0\} \tag{26}$$

so that the process can be represented by the following infinite set of differential–difference equations:

$$(d/dt)P_{i,j}(t) = \lambda_{j-1}P_{i,j-1}(t) - (\lambda_j + \mu_j)P_{i,j}(t) + \mu_{j+1}P_{i,j+1}(t), \qquad j = 1, 2, 3, \ldots \tag{27}$$

$$(d/dt)P_{i,0}(t) = -\lambda_0 P_{i,0}(t) + \mu_1 P_{i,1}(t) \qquad i = 0, 1, 2, \ldots \tag{28}$$

subject to the initial conditions

$$P_{i,j}(0) = \delta_{i,j}, \quad i, j = 0, 1, 2, \ldots \tag{29}$$

in terms of the Kronecker delta $\delta_{i,j}$ defined by Eq. (25).

The initial-value problem consisting of the system of Eqs. (27) and (28), together with the initial condition (29), is usually solved by appealing to spectral theory.

C. Queues with Balking and Reneging

As a special case of queues with state-dependent parameters, we may consider an M/M/1 queueing system with further assumptions about the balking and reneging behavior of the customers. Suppose that the mean arrival and service rates are λ and μ, respectively, and that the queue discipline is first come, first served. An arriving customer balks with probability n/N and, therefore, joins the system with probability

$$e_n = 1 - n/N, \qquad n = 0, 1, 2, \ldots, N, \tag{30}$$

where n is the number in the system and N the maximum number allowed in the system. After joining the queue, each customer waits for service a certain length of time, which is evidently a random variable with the density function

$$d(t) = \alpha e^{-\alpha t}, \qquad \alpha > 0, \tag{31}$$

and if service does not begin by then, he or she departs and is lost to the system. Denoting, as usual, the transient probability of n in the system at time t by $P_n(t)$, we have the following differential–difference equations:

$$(d/dt)P_0(t) = -\lambda P_0(t) + \mu P_1(t) \tag{32}$$

$$(d/dt)P_n(t) = -[(1 - n/N)\lambda + \mu + (n-1)\alpha]P_n(t) + [1 - (n-1)/N]\lambda P_{n-1}(t) + (\mu + n\alpha)P_{n+1}(t), \qquad n = 1, 2, 3, \ldots, N-1 \tag{33}$$

$$(d/dt)P_N(t) = -[\mu + (N-1)\alpha]P_N(t) + (\lambda/N)P_{N-1}(t). \tag{34}$$

Furthermore, if P_n denotes the steady-state probability of n in the system and if

$$\gamma = \mu/\alpha \quad \text{and} \quad \delta = \lambda/(N\alpha) \tag{35}$$

then the steady-state equations are

$$0 = -N\delta P_0 + \gamma P_1 \tag{36}$$

$$0 = -[\gamma + (N-n)\delta + n - 1]P_n + (N - n + 1)\delta P_{n-1} + (\gamma + n)P_{n+1}, \quad N = 1, 2, 3, \ldots, N-1 \tag{37}$$

$$0 = -(\gamma + N - 1)P_N + \delta P_{N-1}. \tag{38}$$

These last difference equations [Eqs. (36)–(38)] can be solved in a straightforward manner. Thus, in terms of the incomplete beta function $B_x(\alpha, \beta)$ defined by

$$B_x(\alpha, \beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\,dt \tag{38a}$$

we find that

$$P_0 = \frac{1 + NB_z(\gamma, N)}{z^{\gamma-1}(1-z)^N}, \tag{38b}$$

where, for convenience, $z = \delta/(1+\delta)$. Furthermore, the mean number in the queue is

$$L_q = (1+\delta)^{-1}[N\delta - (\gamma+\delta)(1-P_0)] \tag{38c}$$

the mean number in the system is

$$\begin{aligned} L &= \sum_{n=1}^{N} nN = L_q + 1 - P_0 \\ &= (1+\delta)^{-1}[N\delta - (\gamma-1)(1-P_0)] \end{aligned} \tag{38d}$$

and the probability that there are R or more in the system is

$$\begin{aligned} Q_R &= \sum_{n=R}^{N} P_n \\ &= P_0\left\{\frac{N!\Gamma(\gamma)B_z(\gamma+R-1, N-R+1)}{(N-R)!\Gamma(\gamma+R-1)z^{\gamma-1}(1-z)^N}\right\}, \end{aligned} \tag{38e}$$

where z is given as in Eq. (38b).

IV. DOUBLE-ENDED QUEUES

As an example of double-ended queues, consider the queues of taxis and passengers at a taxi stand, where at times there are passengers waiting for taxis and at other times there may be taxis waiting for passengers. Such queueing systems, sometimes called double queues, in which either the units wait in queue for service or idle servers queue up for customers, have varied applications in inventory and production processes. Obviously, in these systems there are two types of units, and interchangeably either may be called passengers (customers) and the other taxis (servers).

A. A Random Walk with Partially Reflecting Barriers

An interesting example of a double-ended queue is provided by a random walk of a particle on the entire real axis, where the particle moves one step to the right or left, respectively, with probabilities λ and μ per unit time, and there are partially reflecting barriers at positions—M and N, say. Suppose that the first barrier reflects the particle with probability r_1 and absorbs it with probability $1-r_1$ and that the second barrier reflects and absorbs the particle with probabilities r_2 and $1-r_2$, respectively. Defining

$$\begin{aligned} P_n(t) = \text{Prob}\{&\text{particle is at position } n \text{ at time } t\} \\ &-M \leqq n \leqq N \end{aligned} \tag{39}$$

we have the differential–difference equations

$$(d/dt)P_{-M}(t) = -r_1\lambda P_{-M}(t) + \mu P_{-M+1}(t), \tag{40}$$

$$\begin{aligned} (d/dt)P_{-M+1}(t) = &-(\lambda+\mu)P_{-M+1}(t) + r_1\lambda P_{-M}(t) \\ &+ \mu P_{-M+2}(t), \end{aligned} \tag{41}$$

$$\begin{aligned} (d/dt)P_n(t) = &-(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) \\ &+ \mu P_{n+1}(t) \\ &-M+2 \leqq n \leqq N-2, \end{aligned} \tag{42}$$

$$\begin{aligned} (d/dt)P_{N-1}(t) = &-(\lambda+\mu)P_{N-1}(t) + r_2\mu P_N(t) \\ &+ \lambda P_{N-2}(t), \end{aligned} \tag{43}$$

$$(d/dt)P_N(t) = -r_2\mu P_N(t) + \lambda P_{N-1}(t), \tag{44}$$

which can be solved by using the Laplace transform [Eq. (18)] and certain generating functions of the type given by Eq. (19).

By setting r_1 and r_2 equal to 0 or 1, we obtain the special cases when both of the barriers are absorbing (reflecting) and when one is absorbing and the other is reflecting. In particular, the case $r_1 = r_2 = 1$, with reflecting barriers at $-M$ and N, corresponds to the double-ended queueing model with limited waiting space for N passengers or M taxis if we define

$$P_n(t) = \begin{cases} \text{Prob}\{n \text{ passengers waiting at time } t\}, \\ \qquad\qquad n \geqq 0 \\ \text{Prob}\{-n \text{ taxis waiting at time } t\}, \\ \qquad\qquad n < 0 \end{cases} \tag{44a}$$

and we thus find that

$$\begin{aligned} P_n(t) = \beta^{n-i}\Bigg[& e^{-\alpha t}\Bigg\{ I_{i-n}(2t\sqrt{\lambda\mu}) \\ & + \sum_{j=0}^{\infty}\big[I_{2(M+N+1)(j+1)+i-n}(2t\sqrt{\lambda\mu}) \\ & + \big[I_{2(M+N+1)(j+1)-i+n}(2t\sqrt{\lambda\mu})\big]\Bigg\} \\ & + \lambda\sum_{j=0}^{\infty}\int_0^t e^{-\alpha\tau}\big\{I_{2(M+N+1)j+2N-i-n}(2\tau\sqrt{\lambda\mu}) \\ & - 2\beta^{-1}I_{2(M+N+1)j+2N-i-n+1}(2\tau\sqrt{\lambda\mu}) \\ & + \beta^{-2}I_{2(M+N+1)j+2N-i-n+2}(2\tau\sqrt{\lambda\mu}) \\ & + I_{2(M+N+1)j+2M+i+n+2}(2\tau\sqrt{\lambda\mu}) \\ & - 2\beta^{-1}I_{2(M+N+1)j+2M+i+n+1}(2\tau\sqrt{\lambda\mu}) \\ & + \beta^{-2}I_{2(M+N+1)j+2M+i+n}(2\tau\sqrt{\lambda\mu})\big\}\,d\tau\Bigg], \end{aligned}$$

$$n = -M, -M+1, \ldots, N, \tag{44b}$$

where $I_\nu(z)$ denotes the modified Bessel function of the first kind defined by

$$I_\nu(z) = \sum_{m=0}^{\infty} \frac{(\frac{1}{2}z)^{\nu+2m}}{m!\Gamma(\nu+m+1)}, \qquad |z| < \infty \tag{44c}$$

and it is assumed that initially (at $t=0$) there are i passengers waiting.

B. Double-Ended Queues with Time-Dependent Rates

Consider a double-ended queueing problem in which the mean arrival rates of taxis and passengers are, respectively, $\alpha(t)$ and $\beta(t)$, which are functions of the time t. Let

$$P_n(t) = \begin{cases} \text{Prob}\{n \text{ taxis waiting at time } t\}, \\ \qquad\qquad n \geqq 0 \\ \text{Prob}\{-n \text{ passengers waiting at time } t\}, \\ \qquad\qquad n < 0 \end{cases} \tag{45}$$

be defined for $n=0, \pm1, \pm2, \ldots$. Then we have the differential–difference equations

$$\begin{aligned}(d/dt)P_n(t) &= [P_{n-1}(t) - P_n(t)]\alpha(t) \\ &\quad - [P_n(t) - P_{n+1}(t)]\beta(t), \\ &\qquad n = 0, \pm1, \pm2, \ldots,\end{aligned} \tag{46}$$

which are usually solved by employing a Laurent generating function

$$F(z,t) = \sum_{n=-\infty}^{\infty} P_n(t) z^n. \tag{47}$$

Indeed, if we define

$$A(t) = \int_0^t \alpha(\tau)\, d\tau \tag{47a}$$

and

$$B(t) = \int_0^t \beta(\tau)\, d\tau, \tag{47b}$$

we shall obtain the explicit representation

$$\begin{aligned}P_n(t) &= \left[\frac{A(t)}{B(t)}\right]^{(n-i)/2} \exp(-A(t) - B(t)) \\ &\quad \times I_{n-i}(2\sqrt{A(t)B(t)}), \\ &\qquad n = 0, \pm1, \pm2, \ldots,\end{aligned} \tag{47c}$$

where $I_\nu(z)$ is defined by Eq. (44c) and i denotes the initial state of the system.

C. Double-Ended Queues with Bulk Service

We now consider a double-ended queueing system by assuming a general interarrival time distribution, with probability density $A(x)$, for customers (passengers), while the taxis arrive in a Poisson stream with mean rate μ. Suppose that there is a limited waiting space for M taxis and N passengers and that a taxi takes a fixed number S of customers from the queue, or the whole queue, whichever is less. We use the probabilities

$$P_n(x,t)\, dx, \qquad -M \leqq n \leqq N, \tag{48}$$

the meaning of which is similar to that given in Eq. (44) with the rôles of taxis and passengers reversed and with the addition that the time since the last customer arrival lies in the interval $(x, x+dx)$. Let $\eta(x)\Delta$ be the first-order probability that a customer arrives in time $(x, x+\Delta)$, given that there was no customer arrival up to x, so that

$$A(x) = \eta(x) \exp\left(- \int_0^x \eta(u)\, du \right). \tag{49}$$

Following the usual probability arguments, we can represent the queueing system by the differential–difference equations

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x} + \eta(x)\right) P_{-M}(x,t) = \mu P_{-M+1}(x,t) \tag{50}$$

$$\begin{aligned}&\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x} + [\mu + \eta(x)]\right) P_n(x,t) \\ &\quad = \mu P_{n+1}(x,t), \qquad -M+1 \leqq n \leqq -1\end{aligned} \tag{51}$$

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x} + [\mu + \eta(x)]\right) P_0(x,t) = \mu \sum_{j=1}^{S} P_j(x,t) \tag{52}$$

$$\begin{aligned}&\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x} + [\mu + \eta(x)]\right) P_n(x,t) \\ &\quad = \mu P_{n+S}(x,t), \qquad 1 \leqq n \leqq N - S\end{aligned} \tag{53}$$

$$\begin{aligned}&\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x} + [\mu + \eta(x)]\right) P_n(x,t) = 0, \\ &\qquad N - S + 1 \leqq n \leqq N,\end{aligned} \tag{54}$$

subject to the following boundary conditions:

$$P_{-M}(0,t) = 0 \tag{55}$$

$$\begin{aligned}P_n(0,t) &= \int_0^\infty P_n(x,t)\eta(x)\, dx, \\ &\qquad M+1 \leqq n \leqq N-1\end{aligned} \tag{56}$$

$$P_N(0,t) = \int_0^\infty [P_{N-1}(x,t) + P_N(x,t)]\eta(x)\, dx. \tag{57}$$

Assuming further that the initial state i of the system is so constrained that $0 < i < S$, we can obtain a closed-form solution of this double-ended queueing problem involving S-seated taxis, again by using a set of suitable generating functions and Laplace transforms.

V. BIRTH-AND-DEATH QUEUEING SYSTEMS

The limiting behavior of a many-server queue as the number of servers becomes large is usually studied by considering it as a birth-and-death process. Such a queueing process $X(t)$ can be defined as a continuous-time Markov chain whose state space is the set of nonnegative integers and whose transition probabilities

$$P_{i,j} = \text{Prob}\{X(t+\tau) = j | X(t) = i\} \quad i, j = 0, 1, 2, \ldots \tag{58}$$

are given by the matrix differential equation

$$(d/dt)\mathbf{P}(t) = \mathbf{A}\mathbf{P}(t) \tag{59}$$

subject to the initial condition

$$\mathbf{P}(0) = \mathbf{I}, \tag{60}$$

where $\mathbf{P}(t)$ is the infinite matrix

$$\mathbf{P}(t) = [P_{i,j}(t)], \quad i, j = 0, 1, 2, \ldots \tag{61}$$

$$\mathbf{A} = \begin{bmatrix} -(\lambda_0 + \mu_0) & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix} \tag{62}$$

$\mathbf{I}$ is the infinite identity matrix

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ \vdots & 0 & 1 & \cdots \\ & \vdots & 0 & \cdots \\ & & \vdots & \end{bmatrix} \tag{63}$$

and λ_i and μ_i denote, respectively, the birth and death rates per unit time when the size of the population is i.

For the n-server queue M/M/n with mean arrival rate λ and mean service rate per customer μ, the process $X_n(t)$ representing the number of customers in the system at time t is a birth-and-death queueing process of the type described above with, of course,

$$\lambda_i \equiv \lambda, \quad i = 0, 1, 2, \ldots \tag{64}$$

$$\mu_i = \begin{cases} i\mu, & \text{if } 0 \leqq i \leqq n \\ n\mu, & \text{if } i \geqq n+1 \end{cases}. \tag{65}$$

For such a queueing process, it can be shown that the parameters $\{\lambda_i\}$ and $\{\mu_i\}$ uniquely determine $P_{i,j}(t)$ defined by Eq. (58) and that the sample paths of $X_n(t)$ are jump functions with probability 1. Furthermore, in terms of the Poisson–Charlier polynomials $c_n(x, a)$ defined by

$$c_n(x; a) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{x}{k} k! a^{-k} a > 0; \quad x = 0, 1, 2, \ldots \tag{65a}$$

we have

$$P_{i,j}(t) = \pi_j \int_0^\infty e^{-xt} Q_i(x) Q_j(x)\, d\Psi(x) \tag{65b}$$

where, for convenience,

$$Q_k(x) = c_k(x/\mu, \lambda/\mu), \quad k \leqq n \tag{65c}$$

the sequence $\{\pi_j\}$ is defined by

$$\pi_0 = 1 \tag{65d}$$

$$\pi_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}, \quad n \geqq 1 \tag{65e}$$

and $\Psi(x)$ is a positive regular measure on $0 \leqq x < \infty$.

VI. IMBEDDED MARKOV CHAIN QUEUEING SYSTEMS

An important characteristic of a birth-and-death process is that, if at any time t the system state E_i is known, then (in principle) the state probabilities for all $t + \tau$ ($\tau \geqq 0$) are also known, irrespective of the route the system may have followed in reaching the state E_i at time t. When this holds true for a system, we may write equations relating the state probabilities at any time t to those at any other time $t + \tau$ ($\tau \geqq 0$). In fact, it can be shown that a necessary and sufficient condition for a system to possess this property, called the Markov property, is that the time between successive changes of state be negative-exponentially distributed (but not necessarily identically distributed). A process that has this characteristic Markovian property is called a Markov process.

As already pointed out in Section I, the main characteristics of a queueing system are the input process, the queue discipline, and the service mechanism, and the three main parameters of a queueing system are the queue length, the waiting time, and the busy period. In a wide variety of queueing systems, such as the Poisson queues (i.e., queues with Poisson arrival and negative-exponential service times), the queue length $N(t)$ is a Markov process because the specification of the queue length at time t is

adequate to predict its length at time $t + \tau$ ($\tau \geqq 0$). Indeed, this is true on account of the Markovian property of the Poisson distribution [Eq. (6)].

We now turn to the birth-and-death process with a view to ascertaining why, for example, an M/G/1 queue, unlike the M/M/s queue, cannot be modeled as a birth-and-death process. Denoting by $N(t)$ the number of customers in the system at time t and applying the law of total probability, we have

$$\begin{aligned}&\text{Prob}\{N(t+\tau) = j\}\\ &\quad = \sum_{i=0}^{\infty} \text{Prob}\{N(t+\tau) = j \mid N(t) = i\}\\ &\qquad\qquad \times \text{Prob}\{N(t) = i\},\\ &\qquad\qquad t \geqq 0; \quad \tau \geqq 0; \quad j = 0, 1, 2, \ldots. \end{aligned} \tag{66}$$

Equations (66) are valid for any queueing system. All that is required is first to evaluate the conditional probabilities

$$\text{Prob}\{N(t+\tau) = j \mid N(t) = i\}, \tag{67}$$

which are called the transition probabilities, and then to solve Eqs. (66) for the state probabilities

$$\text{Prob}\{N(\zeta) = k\}, \quad \zeta \geqq 0; \quad k = 0, 1, 2, \ldots. \tag{68}$$

While analyzing queueing systems that are birth-and-death processes, we tacitly assume τ to be small and let $\tau \to 0$. In view of the fact that all the transition probabilities except those representing a transition of step size 0 or ± 1 are $o(\tau)$ as $\tau \to 0$, we can avoid laborious calculations of the transition probabilities, since most of them become irrelevant as $\tau \to 0$. In principle, however, we do calculate the transition probabilities first, simply because the birth-and-death process is a Markov process; that is, the time separating successive transitions is negative-exponentially distributed, obviating the need to consider how long the system had been in state E_i at time t, or exactly how and in what way state E_i at time t evolved. This situation does not hold true in the case of a non-Poisson queue, such as the M/G/1 queue in which the service-time distribution is not assumed to be the negative-exponential distribution. Consequently, in order to specify the transition probabilities correctly we must know not only the precise state in which the system is at time t, but also exactly how long the system had been in that state. Equation (66), though not incorrect, is no longer applicable in the analysis of the M/G/1 queue.

Thus, when the input and service-time distributions assume general forms, as in various (non-Poisson) special cases of the s-server queueing system GI/G/s, the process $N(t)$ is no longer Markovian. The correct specification of such a queueing system is then provided by the vector process $[N(t), \xi(t), \eta(t)]$, where $\xi(t)$ denotes the time that has elapsed since the last arrival and $\eta(t)$ is the expended service time of the customer being served at time t. A systematic study of this vector space is theoretically possible, but the mathematical analysis to be applied is much too involved. A simpler method of analysis is therefore required for the study of these queueing models. A powerful method for the analysis of such queueing systems is provided by Palm's concept of regeneration points and Kendall's technique of imbedded Markov chains, both of which are introduced below.

A set $\mathcal{S}$ of time epochs is said to be a set of regeneration points for a stochastic process $X(t)$ if and only if, for all $t > t_0$, we have

$$\begin{aligned}&\text{Distribution}\{X(t) \mid X(t_0)\}\\ &\quad = \text{Distribution}\{X(t) \mid X(\tau) \text{ for all } \tau \leqq t_0\}\end{aligned} \tag{69}$$

whenever t_0 is known to belong to $\mathcal{S}$. Clearly, the condition (69) implies that the development of the process $X(t)$ during $t > t_0$ is independent of the history of the process during $0 < t \leqq t_0$. In a Markov process, the entire range of t values is a set of regeneration points. By definition, a process that has a set of regeneration points is called a regenerative process.

Suppose now that there exists a denumerable $\mathcal{S}$-set of regeneration points

$$\{t_n, n = 0, 1, 2, \ldots\}$$

such that

$$0 < t_0 < t_1 < t_2 < \cdots$$

and let $X_n = X(t_n)$. Then it follows readily from the above definition that the sequence of random variables $\{X_n\}$ forms a Markov chain (which is, by definition, the discrete analog of a Markov process, so that the characteristic Markov property holds true only at discrete values of the parameter t, i.e., at a discrete set of regeneration points). This Markov chain is said to be "imbedded" in the given stochastic process $X(t)$.

In a single-server queueing process, the set of points at which the counter is free is clearly an $\mathcal{S}$-set. Another example of an $\mathcal{S}$-set is the set $\mathcal{S}_x$ consisting (for a given x) of those time epochs at which the customer at the counter has already been served for a period x.

In the case of the M/G/1 queue, with Poisson input and a general service-time distribution, the service completion points (i.e., those epochs at which the customers complete service and leave the system) are regeneration points. This is so because, whenever a customer leaves the system, either the system becomes empty or a previously waiting customer starts service; in either event the transition probabilities (and hence also the future evolution of the system) depend only on the number of customers

in the system at the service completion point. Thus, the system state at this discrete set of regeneration points is an imbedded Markov chain.

Finally, in the general case of a non-Poisson queueing system, if there exists an increasing sequence $\{t_n\}$ of regeneration points such that we can easily compute the transition probabilities associated with

$$\{N_{n+1} \mid N_n\}; \quad N_n = N(t_n),$$

where $N(t)$ denotes the queue length at time t, then the analysis using the imbedded Markov chain $\{N_n\}$ will yield valuable information concerning the process $N(t)$.

VII. EQUATIONS FOR THE M/M/1 QUEUEING SYSTEM

The M/M/1 queue is the simplest (and most interesting) of the various queueing systems presented in the preceding sections. Indeed, it is the classic example of a queueing system, and the analytical tools required for its study are rather elementary. Even though these analytical techniques do not easily carry over into more complex queueing systems, the general behavior of an M/M/1 queue is, in many respects, similar to that observed in the more complex cases. For this reason, we present here the following equations, which summarize the important results for an M/M/1 queueing system (where ρ is the traffic intensity or utilization factor and assuming steady state, wherever appropriate):

$$\begin{aligned} P_0 &= \text{Prob\{no customers in the system\}} \\ &= 1 - (\lambda/\mu) = 1 - \rho \\ P_n &= \text{Prob}\{n \text{ customers in the system}\} \\ &= \rho^n P_0(t) = \rho^n(1-\rho) \\ L &= \text{average (or mean) number of customers in the system} \\ &= \rho/(1-\rho) = \lambda/(\mu-\lambda) \end{aligned}$$

with variance

$$\begin{aligned} \sigma_L^2 &= \rho/(1-\rho)^2 \\ L_{\mathrm{q}} &= \text{average (or mean) number of customers in the queue} \\ &= \rho^2/(1-\rho) = \lambda^2/[\mu(\mu-\lambda)] \\ W &= \text{average (or mean) waiting time in the system} \\ &= (1/\mu)/(1-\rho) = 1/(\mu-\lambda) \\ W_{\mathrm{q}} &= \text{average (or mean) waiting time in the queue} \\ &= (\rho/\mu)/(1-\rho) = \lambda/[\mu(\mu-\lambda)] \\ L &= \lambda W; \quad \mathrm{L_q} = \lambda \mathrm{W_q} \end{aligned}$$

$$\begin{aligned} &\text{Prob\{interdeparture time} \leqq t\} \\ &\quad = \text{Prob\{idle period duration} \leqq t\} = 1 - e^{-\lambda t} \\ \psi(y) &= \text{probability density function of the waiting time spent by a customer in the queue before service} \\ &= \lambda(1-\rho)e^{-\mu(1-\rho)y} \\ &= (1-\lambda/\mu)\lambda e^{(\lambda-\mu)y}, \quad 0 \leqq y < \infty \\ g(y) &= \text{probability density function of the busy period (i.e., the interval of time between two successive idle periods)} \\ &= (1/y\sqrt{\rho})e^{-(\lambda+\mu)y} I_1(2y\sqrt{\lambda\mu}), \end{aligned}$$

where $I_v(z)$ is the modified Bessel function of the first kind defined by Eq. (44c):

$$\begin{aligned} f_n &= \text{Prob}\{n \text{ customers are served during a busy period}\} \\ &= \frac{1}{n}\binom{2n-2}{n-1}\rho^{n-1}(1+\rho)^{1-2n}. \end{aligned}$$

VIII. EXAMPLES OF APPLICATIONS

In this section we briefly consider a few examples of applications of queueing theory in today's world. Indeed, as already indicated in the previous sections, queueing theory finds applications in a wide variety of fields that may be in need of optimization. Essentially, therefore, queueing theory is used to determine the optimum number of servers by striking a balance between the cost of waiting and the cost of an idle server.

A. Typical Telephone Booth Problem

Consider a typical telephone booth problem in which customers arrive at the booth in Poisson stream with an average interarrival rate of 10 min; that is, the average time between two successive customer arrivals is 10 min. Therefore, the parameter λ is equal to 0.1 customer arrival per minute, or 6 customer arrivals per hour. Suppose also that the length of an average telephone call is exponentially distributed, with an average service time of 3 min, so that there are $\frac{1}{3}$ service activities per minute, and we may set μ equal to 20 services per hour.

In order to find the probability that a customer arriving at the booth will have to wait for service, we proceed as follows:

$$\begin{aligned} &\text{Prob\{an arriving customer must wait for service\}} \\ &\quad = 1 - P_0 = 1 - (1 - \lambda/\mu) \\ &\quad = \lambda/\mu = 6/20 = 0.3. \end{aligned}$$

Thus, there is (approximately) one chance in three that a customer arriving at the telephone booth must wait for the booth to become empty.

Next we compute the average queue length, that is, the average number of customers waiting in the queue for service.

$$\begin{aligned}&\text{Average number of customers in the queue}\\&= L_q = \lambda^2/[\mu(\mu-\lambda)]\\&= 36/(20\times 14) = 0.13\end{aligned}$$

so that an arriving customer will find (on the average) 0.13 customer ahead of him or her on arriving at the telephone booth.

Suppose now that the telephone company proposes to install a second telephone booth next to the existing one, provided that the company is convinced that an arriving customer would expect to wait at least 3 min for service (i.e., for the telephone booth to be empty). By how much does the flow of arriving customers have to increase in order to justify the installation of a second telephone booth? To answer this question, we compute λ so that the average waiting time in the queue is 3 min, or $\frac{1}{20}$ hr. Thus,

$$W_q = \lambda/[20(20-\lambda)] = 1/20,$$

which yields

$$\lambda = 10 \text{ arrivals per hour}$$

It follows that the flow of arriving customers must increase from 6 per hour to 10 per hour in order to justify the additional telephone booth.

Finally, we calculate the probability that an arriving customer will have to wait at least 10 min for the telephone booth to be empty:

$$\begin{aligned}&\text{Prob}\{\text{waiting time} \geqq \tfrac{1}{6}\text{ hr}\}\\&= \int_{1/6}^{\infty}\left(1-\frac{\lambda}{\mu}\right)\lambda e^{(\lambda-\mu)y}\,dy = -\frac{\lambda}{\mu}e^{(\lambda-\mu)y}\Big|_{1/6}^{\infty}\\&= -\frac{6}{20}e^{-14y}\Big|_{1/6}^{\infty} = (0.3)e^{-2.3} = 0.03\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\text{Prob}\{\text{total time (waiting time + service time)}\\&\text{in the system} \geqq \tfrac{1}{6}\text{hr}\}\\&= \int_{1/6}^{\infty}(\mu-\lambda)e^{(\lambda-\mu)y}\,dy = -e^{(\lambda-\mu)y}\Big|_{1/6}^{\infty}\\&= -e^{-14y}\Big|_{1/6}^{\infty} = e^{-2.3} = 0.10.\end{aligned}$$

B. Queueing Problem Involving Traffic Lights

Consider an M/E_k/1 queueing system that involves the characteristics displayed by a series of k traffic lights. To study the system as a queueing problem, we let λ denote the customer (vehicle) arrival rate per minute and let μ denote the service rate per minute (i.e., the traffic lights will turn green for a period of time, allowing μ vehicles to pass through the system over a 1-min period). Then, assuming an infinite source population, an infinite queue length, and an FIFO (first-in-first-out) queue discipline, the following relationships are easily derived for this M/E_k/1 queueing system:

$$\begin{aligned}L &= \text{average number of vehicles in the system}\\&= \frac{k+1}{2k}\frac{\lambda^2}{\mu(\mu-\lambda)} + \rho\\L_q &= \text{average number of vehicles in the queue}\\&= \frac{k+1}{2k}\frac{\lambda^2}{\mu(\mu-\lambda)}\\W &= \text{average waiting time in the system}\\&= \frac{k+1}{2k}\frac{\lambda}{\mu(\mu-\lambda)} + \frac{1}{\mu}\\W_q &= \text{average waiting time in the queue}\\&= \frac{k+1}{2k}\frac{\lambda}{\mu(\mu-\lambda)}\end{aligned}$$

where, as usual, $\rho = \lambda/\mu$ and the Erlang stage parameter $k = 1, 2, 3, \ldots$.

Each of the output parameters L, L_q, W, and W_q is important in the analysis of the queueing problem associated with k traffic lights. For the sake of simplicity, let $k = 2$ (i.e., a queueing system consisting of two traffic lights) and $\mu = 7$ (i.e., seven vehicles are allowed to pass through the system over a 1-min period). The following tabulation illustrates a comparative behavior of the system when the customer (vehicle) arrival rate per minute doubles from $\lambda = 3$ to $\lambda = 6$ ($k = 2$; $\mu = 7$):

	$\lambda = 3$	$\lambda = 6$
L	1.3929	16.2857
L_q	0.9643	15.4286
W	0.4643	2.7143
W_q	0.3214	2.5714

Thus, a transition from $\lambda = 3$ to $\lambda = 6$ arrivals per minute (in a system with two traffic lights and $\mu = 7$ services per minute) has resulted in a rather significant increase in the customer's waiting time. More precisely, a customer who waits in the system (with $\lambda = 3$) for a period of ~28 sec will now have to wait in the system (with $\lambda = 6$) for ~2 min and 34 sec, that is, a waiting-time increase by a factor of

$\sim$5 as a result of the doubling of customer arrivals from 3 to 6 arrivals per minute in a system with two traffic lights and 7 services per minute.

C. Baggage Claim Problem

Many queueing systems are equipped with certain special rules for service that are markedly different from those of the classic queueing situations described in this and the preceding sections. Consider, for example, the queue of airline passengers who wish to claim their luggage from a baggage carousel.

Suppose that there are N_b bags and N_p passengers having one or more bags to claim. Note that the passengers leave an airplane and walk to the baggage claim area while the bags are removed from the airplane and transported to a conveyor belt (or some other device), where they are handled at a finite rate; therefore, the bags are not all available simultaneously. Let

$$N_b S_b(t) = \text{cumulative number of bags to become available to customers by time } t$$

so that

$$\begin{aligned} S_b(t) &= \text{fraction of bags to arrive by time } t \\ &= \text{Prob\{some arbitrarily chosen bag arrives by time } t\} \end{aligned}$$

Also let

$$N_p S_p(t) = \text{cumulative number of passengers to arrive at the baggage claim area by time } t$$

so that

$$\begin{aligned} S_p(t) &= \text{fraction of passengers to arrive by time } t \\ &= \text{Prob\{some arbitrarily chosen passenger arrives by time } t\}. \end{aligned}$$

Clearly, there are two distinct queues: a queue of bags waiting to be claimed by the owners and a queue of passengers waiting for their bags. The service facility is intended, in some fictitious manner, to unite a passenger (customer) with his or her bag or bags and to send them both on their way. Neglecting the time a passenger takes to remove his or her bag or bags from the carousel (or rack), it is easily seen that

$$\begin{aligned} &\text{Prob\{an arbitrary bag is removed from the carousel (or rack) by time } t\} \\ &= \text{Prob\{the bag has arrived by time } t \text{ and its owner has also arrived by time } t\}. \end{aligned}$$

Indeed, it is reasonable to assume that these last two events are statistically independent.

The expected cumulative number of bags to depart by time t is given by

$$\begin{aligned} Q_b(t) &= N_b \text{ Prob\{a bag has left}\} \\ &= N_b S_b(t) S_p(t) \end{aligned}$$

so that the expected number of bags in the queue of bags is

$$R_b(t) = N_b S_b(t)[1 - S_p(t)]$$

These results hold true even if some passengers have more than one bag to claim, provided that the passenger removes his or her bag from the carousel immediately even though he or she must still wait for a second bag.

IX. FURTHER DEVELOPMENTS IN RELATED AREAS

This last section is devoted to a brief exposition of some further developments in various areas related rather closely to that of queueing theory. Of the numerous such areas, we choose to briefly describe here six important ones with a view to stimulating the interested reader to further studies and research in these areas.

A. Metamodeling in Complex Computer Systems

In recent years we have encountered a remarkably significant increase in the number of approximate methods which are available for analyzing nonhomogeneous (nonproduct form) queueing network models occurring in complex computer systems. Generally speaking, these are *ad hoc* methods which usually focus on specific aspects of the involved system operation and which appear to be different from one another. Thus, it is a difficult task to clearly see the underlying principles of model development, to understand the relationship between different queueing models of the same system, or to apply the existing approximate methods to other seemingly new situations.

Metamodeling in complex computer systems is a systematic study of approximate methods in queueing network modeling and (more importantly) of the way these methods are developed; indeed, it identifies the underlying modeling process and provides tools and techniques for model development which will enable one to realistically sort through the many different methods available in the literature, understand their bases and shortcomings, and (if and when needed) apply them profitably to new and significant problems.

B. Dynamic Flows in Communication Network Models

In the design of communication networks, the fundamental problem consists in finding efficient traffic control

rules. In the case of a slightly loaded communication network, which is subjected to a stationary load, this problem is rather simple, since each user can choose a route that best serves its own interests. In the situation when the offered traffic increases, however, the problem turns out to be significantly more complex. Competition for limited network resources results in various nonlinear phenomena which can manifest themselves in deadlocks or overall network congestion. Furthermore, the procedures that did serve the interests of individual users in the slightly loaded case may now provide a poor service for *all* users when the traffic is rather heavy.

The dynamic flow theory (that is, the modeling and optimization of dynamic flows) is applicable to the analysis and synthesis of control structures for numerous types of networks; it is especially appropriate for systematically treating communication networks such as circuit and packet-switched networks.

C. Queueing Theory in Polling Systems

A polling system provides a way of serving requests by several (possibly buffered) users in a cyclic order, with generally nonzero switchover times. Polling schemes have already been applied in numerous computer-terminal communication systems, including (for example) such standard data link protocols as BSC, SDLC, and HDLC, and in local-area computer networks (e.g., token ring). In the language of queueing theory, one needs here to consider a multiple queue, cyclic service system whose congestion analysis has indeed been the subject of many investigations.

D. Scheduling Sequences and Queues

The main objective in resource scheduling is to construct a work schedule which *either* (i) completes the work at the earliest possible date by making use of the available resources, *or* (ii) makes the fullest and most efficient use of resources to complete the work by a fixed date.

One of the simplest examples of a work schedule under the aforementioned approach (i) (that is, of a work schedule constructed within the resource restraints) is provided by the case where there is a fixed labour force available, the duration and labour content of each network activity is prescibed (fixed), and it is required to determine the period of completion of the project within these limitations.

In examples of resource scheduling under the previously mentioned approach (ii), the project completion date is to be treated as an important objective. Thus it is necessary to construct a work schedule within a particular time span. Since, in this case, activities on the critical path cannot be moved without affecting the project completion date, the scheduling of noncritical activities at alternative dates (within their earliest and latest cycles) provides some flexibility (latitude) which can indeed be deployed to produce a work schedule with a view to satisfying a particular criterion such as smoothest work load, least overtime operations, etc.

Because of the obvious importance of the sequence in which activities are presented for scheduling, such activities are deliberately arranged in a priority-sequenced queue. This scheduling queue is so formed as to first consider the most important activities (that is, activities which are likely to be the most difficult to schedule). And, in order to qualify for a place at all in the scheduling queue, such activities must be available for scheduling at that point of the calculation, that is, when all preceding activities are completed.

E. Queueing Theory and Statistical Performance Evaluation of Computer Networks and Systems

The various analyses and techniques involving continuous- and discrete-time queueing models, stochastic Petri networks (which have emerged as a remarkably promising high-level performance modeling tool for systems that exhibit concurrency, synchronization, and randomness), and so on, are found to apply in the problem of statistical performance evaluation of computer networks and systems. Such problems are becoming increasingly important as we are continually seeking to design more and more sophisticated (and efficient) communication and information processing systems. The ability to predict the performance of a proposed system without actually having to construct the system is undoubtedly an extremely cost-effective design tool.

F. Boundary-Value Problems in Queueing Theory

In the analysis of certain queueing models which differ only slightly from the classical M/G/1 model (for instance, if two types of customers have to be distinguished), one is often led to the problem of solving a functional equation of the form:

$$K(p_1, p_2)\,\Phi(p_1, p_2) = A(p_1, p_2)\,\Phi(p_1, 0) + B(p_1, p_2)\Phi(0, p_2) + C(p_1, p_2)\,\Phi(0, 0) \qquad (|p_1| \le 1;\ |p_2| \le 1), \quad (70)$$

where the kernel $K(p_1, p_2)$ as well as the functions $A(p_1, p_2)$, $B(p_1, p_2)$, and $C(p_1, p_2)$ are prescribed, and $\Phi(p_1, p_2)$ is an unknown function which should be a

bivariate generating function in p_1 and in p_2 of a proper probability distribution with the set

$$\mathbb{N}_0 \times \mathbb{N}_0 \quad (\mathbb{N}_0 := \mathbb{N} \cup \{0\}; \quad \mathbb{N} := \{1, 2, 3, \ldots\})$$

as its support. Functional equations of the type given by (70) are usually investigated by first solving some Riemann–Hilbert boundary-value problems which are fomulated for the unique determination of the functions $\Phi(p_1, 0)$ and $\Phi(0, p_2)$.

SEE ALSO THE FOLLOWING ARTICLES

COMPUTER NETWORKS • MATHEMATICAL MODELING • OPERATIONS RESEARCH • PROBABILITY • STOCHASTIC PROCESSES • TELECOMMUNICATIONS • Z-TRANSFORM

BIBLIOGRAPHY

Agrawal, S. C. (1985). "Metamodeling: A Study of Approximations in Queueing Models," MIT Press, Cambridge, MA.

Baccelli, F., and Brémaud, P. (1987). "Palm Probabilities and Stationary Queues," Springer-Verlag, Berlin, Heidelberg, and New York.

Bartlett, M. S. (1980). "An Introduction to Stochastic Processes," 3rd ed., Cambridge University Press, New York.

Bhat, U. N., and Basawa, I. V. (eds.) (1992). "Queueing and Related Models," Oxford University Press, Oxford, New York, and Toronto.

Blanc, J. P. C. (1982). "Application of the Theory of Boundary Value Problems in the Analysis of a Queueing Model with Paired Services," Mathematical Centre Tracts 153, Mathematisch Centrum, Amsterdam.

Borovkov, A. A. (1984). "Asymptotic Methods in Queuing Theory" (trans. from the Russian by D. Newton), Wiley, New York.

Brémaud, P. (1981). "Point Processes and Queues," Springer-Verlag, New York.

Bruell, S. C., and Balbo, G. (1980). "Computational Algorithms for Closed Queueing Networks," Elsevier/North-Holland, New York.

Bunday, B. D. (1986). "Basic Queueing Theory," Edward Arnold (Publishers) Limited, London.

Cohen, J. W., and Boxma, O. J. (1983). "Boundary Value Problems in Queueing System Analysis," Elsevier/North-Holland, Amsterdam, New York, and Oxford.

Cooper, R. B. (1981). "Introduction to Queueing Theory," 2nd ed., Elsevier/North-Holland, New York.

Disney, R. L., and Kiessler, P. C. (1987). "Traffic Processes in Queueing Networks: A Markov Renewal Approach," Johns Hopkins University Press, Baltimore and London.

Filipiak, J. (1988). "Modeling and Control of Dynamic Flows in Communication Networks," Springer-Verlag, Berlin, Heidelberg, and New York.

Franken, P., König, D., Arndt, U., and Schmidt, V. (1982). "Queues and Point Processes," Akademie-Verlag/Wiley, Berlin and New York.

Kleinrock, L., and Gail, R. (1996). "Queueing Systems: Problems and Solutions," John Wiley and Sons, New York, Chichester, Brisbane, Toronto, and Singapore.

Lazowska, E. D., Zahorjan, J., Graham, G. S., and Sevcik, K. C. (1984). "Quantitative System Performance: Computer System Analysis Using Queueing Network Models," Prentice-Hall, Englewood Cliffs, NJ.

Neuts, M. F. (1981). "Matrix–Geometric Solutions in Stochastic Models: An Algorithmic Approach," Johns Hopkins University Press, Baltimore, MD.

Newell, G. F. (1982). "Applications of Queueing Theory," 2nd ed., Chapman and Hall, New York.

Robertazzi, T. G. (1994). "Computer Networks and Systems: Queueing Theory and Performance Evaluation," 2nd ed., Springer-Verlag, New York, Berlin, and Heidelberg.

Srivastava, H. M., and Kashyap, B. R. K. (1982). "Special Functions in Queuing Theory and Related Stochastic Processes," Academic Press, New York.

Takagi, H. (1986). "Analysis of Polling Systems," MIT Press, Cambridge, MA.

Viswanadham, N., and Narahari, Y. (1992). "Performance Modeling of Automated Manufacturing Systems," Prentice-Hall, Englewood Cliffs, NJ.

Wolff, R. W. (1989). "Stochastic Modeling and the Theory of Queues," Prentice-Hall, Englewood Cliffs, NJ.

Woodgate, H. S. (1977). "Planning by Network: Project Planning and Control Using Network Techniques," 3rd ed., Business Books Limited, London.



Reduction of Dimensionality

Zhidong Bai
National University of Singapore

P. R. Krishnaiah
University of Pittsburgh

I. Introduction
II. Preliminaries
III. Tests for the Rank of the Regression Matrix
IV. Rank of Regression Matrix by Model Selection Methods
V. Reduction of Dimensionality under the Fanova Model
VI. Rank of the Covariance Matrix of Random Effects
VII. Variable Selection in Multivariate Regression
VIII. Variable Selection in Discriminant Analysis
IX. Tests for Rank of Canonical Correlation Matrix
X. Model Selection for Rank of Canonical Correlation Matrix
XI. Variable Selection in Canonical Correlation Analysis
XII. Dimensionality and Dependence in Two-way Contingency Table
XIII. Nonexact Test for the Two-Sample Problem
XIV. Multivariate Discrimination Analysis

GLOSSARY

Canonical correlation analysis Methodology dealing with the study of the dependence between two sets of variables.

Canonical variables Pairs of linear combinations of two sets of variables that best explain the dependence between the original sets of variables.

Discriminant analysis Methodology dealing with the classification of heterogeneous populations.

Finite intersection tests Procedures involving simultaneous tests of a finite number of hypotheses that are of interest.

Large dimensional random matrix Random matrix whose dimension increases proportionally to the degrees of freedom.

Linear discriminant functions Linear functions of the original variables that maximize the discrimination between populations.

Principal component analysis Technique of finding certain linear combinations of original variables that are important in explaining variation among experimental units.

Spectral analysis of large dimensional random matrices Theory of limiting distributions of eigenvalues of large dimensional random matrices.

Wishart distribution Distribution of the sample sums of squares and cross-products matrix when the underlying distribution of the observations is multivariate normal.

IN A NUMBER of situations, the data analysis is confronted with numerous variables or parameters. One way of finding more efficient statistical inferences is to select a smaller number of important variables or parameters, i.e., to reduce the dimension of data. The important variables selected may be the original variables or linear

combinations of the original variables. Some methods of reduction of dimensionality are principal component analysis, canonical correlation analysis, and finding the rank of the regression matrix. Sometimes the number of parameters to be estimated needs to be reduced, such as the nonexact tests by omitting the estimation of a large dimensional covariance matrix.

I. INTRODUCTION

Multivariate statistical analysis is meant to deal with high-dimensional data. On one hand, measurements on more variables must provide more information about the statistical problems. On the other hand, however, the model with more variables needs more parameters to describe and thus more efficiency would be lost due to estimation of the parameters. Therefore, suitably reducing the number of variables (i.e., excluding those of less importance) or proposing new statistical models involved with a smaller number of parameters has been become an interesting topic in both theoretical and applied statistics.

The techniques of multivariate regression analysis and canonical correlation analysis play important roles in the analysis of multivariate data in many disciplines. In the area of multivariate regression analysis, it is of interest to select a smaller number of variables that are adequate for prediction. Similarly, in canonical correlation analysis, it is of interest to select important variables that are adequate to explain the relationship between two sets of variables.

The performance of some traditional statistical approaches becomes poorer and poorer as the dimension of data increases, even breaking down when the dimension of data is larger than the degrees of freedom. Some remedies for the classical statistical methods have more or less purposely been proposed in the literature.

The main emphasis of this review is on techniques for determination of the ranks of the regression matrix and canonical correlation matrix, on methods for selection of important original variables in the areas of multivariate regression analysis and canonical correlation analysis. We also review some remedies to classical statistical tests when the latter perform poorly in dealing with high dimensional data. This review is by no means exhaustive. For lack of space, we restrict our attention to the reduction-of-dimensionality problems in the above-mentioned areas. For discussions of other important topics in multivariate analysis, the reader is referred to Anderson (1984a), Bai (1999), Gnanadesikan (1977), Krishnaiah (1980a), Krishnaiah and Kanal (1982), Kshirsagar (1972), and Rao (1973).

II. PRELIMINARIES

The following notation is used throughout this article. The transpose of a matrix A is denoted by A', whereas the inverse of a square matrix B is denoted by B^{-1}. The transpose of a conjugate of a complex matrix C is denoted by C^*. We now define elliptically symmetric distribution, complex multivariate normal distribution, and complex elliptically symmetric distribution.

A real random vector $\mathbf{x}$: $p \times 1$ is said to have elliptically symmetric distribution if its density is of the form

$$f(\mathbf{x}) = |\Sigma|^{-1/2} h[(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})] \tag{1}$$

Multivariate normal, multivariate t, and multivariate Cauchy distributions are special cases of the elliptically symmetric distribution. For more details on the elliptically symmetric distributions, the reader is referred to Kelker (1970).

A $p \times 1$ random vector $\mathbf{z} = \mathbf{x}_1 + i\mathbf{x}_2$ is said to be distributed as complex multivariate normal if the real vector $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$ is distributed as multivariate normal with mean vector $\boldsymbol{\mu}_0' = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')$ and covariance matrix Σ_0, where

$$\Sigma_0 = \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ -\Sigma_2 & \Sigma_1 \end{pmatrix} \tag{2}$$

$\Sigma_1' = \Sigma_1$ and $\Sigma_2' = -\Sigma_2$ are of order $p \times p$. The complex multivariate normal distribution was considered by Wooding (1956), Goodman (1963), and others. The density function of the complex multivariate normal distribution is of the form

$$f(\mathbf{z}) = \pi^{-p} |\Sigma|^{-1} \exp[-(\mathbf{z} - \boldsymbol{\mu})^* \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu})], \tag{3}$$

where $\Sigma = 2(\Sigma_1 - i\Sigma_2)$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_1 + i\boldsymbol{\mu}_2$. Note that a complex multivariate normal distribution is uniquely determine by its mean vector $\boldsymbol{\mu} = E\mathbf{z}$ and covariance matrix $\Sigma = E(\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^*$. For a review of the literature on complex multivariate distributions, the reader is referred to Krishnaiah (1976).

We now define the complex elliptically symmetric distribution introduced by Krishnaiah and Lin (1986). The $p \times 1$ random vector $\mathbf{z} = \mathbf{x}_1 + i\mathbf{x}_2$ is said to have a complex elliptically symmetric distribution if the real vector $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$ has an elliptically symmetric distribution with density

$$f(\mathbf{x}) = |\Sigma_0|^{-1/2} h\left[\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_0)' \Sigma_0^{-1} (\mathbf{x} - \boldsymbol{\mu}_0)\right], \tag{4}$$

where $\boldsymbol{\mu}_0' = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')$ and Σ_0 has the same expression as (2). Write $\Sigma = 2(\Sigma_1 - i\Sigma_2)$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_1 + i\boldsymbol{\mu}_2$. By the facts that

$$(\mathbf{z}-\boldsymbol{\mu})^*\Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu}) = \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)'\Sigma_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)$$

and

$$|\Sigma| = 2^p|\Sigma_0|^{1/2},$$

the density of $\mathbf{z}$ can be written in the form that

$$g(\mathbf{z}) = |\Sigma|^{-1}h[(\mathbf{z}-\boldsymbol{\mu})^*\Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu})]. \tag{5}$$

Complex multivariate normal and complex multivariate t distributions are special cases of the complex elliptically symmetric distribution.

III. TESTS FOR THE RANK OF THE REGRESSION MATRIX

We first discuss procedures for testing the hypothesis on the number of significant discriminant functions. Note that this is a special case of the testing problem for the rank of regression matrix.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_k$ (real) be distributed independently as multivariate normal (imnd.) with mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ and a common covariance matrix Σ. Furthermore, let $\mathbf{x}_{ij}$ $(j = 1, 2, \ldots, n_i)$ denote the jth independent observation on $\mathbf{x}_i$. Then, the between-group sums of squares and cross-product (SP) matrix is given by

$$S_b = \sum_{i=1}^{k} n_i(\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot\cdot})(\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot\cdot})' \tag{6}$$

whereas the within-group SP matrix is given by

$$S_w = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i\cdot})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i\cdot})' \tag{7}$$

where

$$n_i\bar{\mathbf{x}}_{i\cdot} = \sum_{j=1}^{n_i}\mathbf{x}_{ij}, \quad n\bar{\mathbf{x}}_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$$

and $n = n_1 + \cdots + n_k$.

Now, let $l_1 \geq \cdots \geq l_p$ denote the eigenvalues of $S_bS_w^{-1}$. Also, write

$$\Omega = \sum_{i=1}^{k} n_i(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})' \tag{8}$$

where $n\boldsymbol{\mu} = (n_1\boldsymbol{\mu}_1 + \cdots + n_k\boldsymbol{\mu}_k)$. Then the rank of Ω is equal to the number of significant discriminant functions. Fisher (1939) proposed to use $T_1 = (l_{r+1} + \cdots + l_s)$ as a test statistic for testing the hypothesis that the rank of Ω is equal to r where $s = \min(p, k-1)$. In general, we can use suitable functions $\psi(l_{r+1}, \ldots, l_s)$ of $l_{r+1}, \ldots, l_s$, to test for the rank of Ω. For example, $\psi(l_{r+1}, \ldots, l_s)$ may be l_{r+1}. But the exact distributions of these statistics involve $\lambda_1, \ldots, \lambda_r$ as nuisance parameters where $\lambda_1 \geq \cdots \geq \lambda_p$ are the eigenvalues of $\Omega\Sigma^{-1}/n$.

We now discuss the asymptotic joint distribution of the eigenvalues of $S_bS_w^{-1}$ without the assumption of normality, which was derived in Bai, Krishnaiah and Liang (1986), since it is useful in implementation of some test procedures for determination of the rank of Ω under general moment conditions. For each i, let $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}$ be independently, identically and elliptically symmetrically distributed (iiesd.) and have a common density of the form given in (1) with a mean vector $\boldsymbol{\mu}_i$ and a common covariance matrix Σ.

Let $l_1 \geq \cdots \geq l_p$ denote the eigenvalues of $S_bS_w^{-1}$. Also, denote by $\theta_1 \geq \cdots \geq \theta_p$ the eigenvalues of $\Omega\Sigma^{-1}/n$, whose multiplicities are given by

$$\begin{aligned}
\theta_1 = \cdots = \theta_{p_1^*} &= \delta_1 > \\
\theta_{p_1^*+1} = \cdots = \theta_{p_2^*} &= \delta_2 > \\
&\vdots \\
\theta_{p_{t-1}^*+1} = \cdots = \theta_{p_t^*} &= \delta_t > \\
\theta_{p_t^*+1} = \cdots = \theta_p &= 0
\end{aligned} \tag{9}$$

where $p_j^* = p_1 + \cdots + p_j$ $(j = 1, 2, \ldots, t+1)$, $r = p_t^*$, $p = p_1 + \cdots + p_{t+1}$, and $p_0^* = 0$.

Define

$$\begin{aligned}
u_{i_h} &= \sqrt{n}\left(2\delta_h^2 + 4\delta_h\right)^{-1/2}\left(l_{i_h} - \delta_h\right) \\
u_{r+j} &= nl_{r+j}
\end{aligned} \tag{10}$$

where $h = 1, 2, \ldots, t$, $i_h = p_{h-1}^* + 1, \ldots, p_h^*$, $j = 1, 2, \ldots, s-r$, and r is the number of nonzero eigenvalues of $\Omega\Sigma^{-1}/n$.

Now, let $n_i/n \to q_i$ $(q_i > 0)$, for $i = 1, 2, \ldots, k$. Then, Bai, Krishnaiah, and Liang (1986) proved that the random vectors $\mathbf{u}_j = (u_{p_{j-1}^*+1}, \ldots, u_{p_j^*})$, $j = 1, \ldots, t+1$, are asymptotically independent and that the limiting distribution of $\mathbf{u}_j$ as $n \to \infty$ is the same as the joint distribution of eigenvalues of the random matrix A_j, where for $j = 1, 2, \ldots, t$, the elements above or on the diagonal of A_j are independently and normally distributed (ind.) with mean zero, variance 1 for diagonal elements and 1/2 for off-diagonal elements. In other words, the random matrices $A_1, \ldots, A_t$ are distributed as standard Gaussian matrices. Also, A_{t+1}: $(p-r)\times(p-r)$ is distributed as a central Wishart matrix with $(k-1-r)$ degrees of freedom. That is, the joint limit distribution of $u_1, \ldots, u_s$ can be expressed as

$$f(u_1 \cdots u_s) = \prod_{j=1}^{t+1}\eta_j\left(u_{p_{j-1^*}+1}, \ldots, u_{p_j^*}\right). \tag{11}$$

Here $\eta_j(\cdot)$ $(j = 1, 2, \ldots, t+1)$ denotes the joint density of eigenvalues of A_j. Computational aspects of the percentage points of the individual eigenvalues of the standard Gaussian matrix and Wishart matrix are discussed

by Krishnaiah (1980b). When the underlying distribution is multivariate normal, the expression (11) was derived by Hsu (1941a), in which W. Q. Liang (personal communication) found an error in the proof. However, Bai (1985) confirmed that the final result of Hsu is still correct. Bai, Krishnaiah, and Liang (1986) extended the preceding result to the case where for each i, the observations are iid. with finite 4th moment. From the result of Bai, Krishnaiah and Liang (1986), it is obvious that when r is the rank of Ω, $n(l_{r+1}+\cdots+l_p)$ is asymptotically distributed as χ^2 with $(p-r)(k-1-r)$ degrees of freedom, provided the kurtosis of the underlying distribution is zero.

Krishnaiah (1982a) proposed the following sequential test procedure for the rank of Ω when $n_1,\ldots,n_k$ tend to infinity in such a way that $(n_i/n),\ldots,(n_k/n)$ tend to (say) $q_1,\ldots,q_k$, respectively. The hypothesis $\Omega=0$ is accepted or rejected according to

$$l_1 \underset{>}{\overset{<}{}} c_{\alpha 1} \tag{12}$$

where

$$P[l_1 \le c_{\alpha 1}|\Omega=0]=(1-\alpha_1). \tag{13}$$

If $\Omega=0$ is accepted, we do not proceed further. Otherwise, we accept or reject H_1 according to

$$l_2 \underset{>}{\overset{<}{}} c_{\alpha 2} \tag{14}$$

where

$$P[l_2 \le c_{\alpha 2}|H_1\colon l_1 > c_{\alpha 1}]=(1-\alpha_2) \tag{15}$$

and H_t denotes the hypothesis that the rank of Ω is equal to t. When H_1 is true, the asymptotic distribution of l_2 is independent of l_1. If H_1 is accepted, we do not proceed further. Otherwise, we accept or reject H_2 according as

$$l_3 \underset{>}{\overset{<}{}} c_{\alpha 3} \tag{16}$$

where

$$P[l_3 \le c_{\alpha 3}|H_2\colon l_2 > c_{\alpha 2}]=(1-\alpha_3) \tag{17}$$

We continue this method until a decision about the rank of Ω is reached.

We now discuss the problem of testing for the rank of the regression matrix. Consider the model

$$Y=XB+E \tag{18}$$

where the rows of $E\colon n\times p$ are distributed as a multivariate normal with mean vector 0 and covariance matrix Σ. Also, let $X\colon n\times q$ denote the design matrix and $B\colon q\times p$, the regression matrix. We assume that $q\ge p$. Tintner (1945) derived the likelihood ratio test (LRT) statistic for the rank of B when Σ is known. Anderson (1951) derived the following expression for the LRT statistic to test the hypothesis H_r, which states that if the rank of B is r, then the LRT is given by

$$L_r=\prod_{j=r+1}^{p}(1+l_j)^{n/2} \tag{19}$$

where $l_1\ge\cdots\ge l_p$ denote the eigenvalues of S_1S^{-1} and

$$S_1=Y'X(X'X)^{-1}X'Y \tag{20}$$

$$S=Y'[I-X(X'X)^{-1}X']Y \tag{21}$$

Fujikoshi (1977) derived expressions for the asymptotic distributions of the test statistics m_1T_1, m_2T_2, and m_3T_3 where

$$\begin{aligned} T_1&=\textstyle\sum_{j=r+1}^{p}\log(1+l_j)\\ T_2&=\textstyle\sum_{j=r+1}^{p}l_j\\ T_3&=\textstyle\sum_{j=r+1}^{p}\left\{\frac{l_j}{1+l_j}\right\} \end{aligned} \tag{22}$$

Here m_1, m_2, and m_3 are certain correction factors. In discussion of limiting properties of these tests, we may choose m_i as n. In addition, we usually assume that $\lim(\Omega/n)=0(1)$ where $\Omega=B'(X'X)B\Sigma^{-1}$. The asymptotic distributions of nT_1, nT_2, and nT_3 under H_0 are χ^2-distribution with $(p-r)(q-r)$ degrees of freedom. Fujikoshi also derived the nonnull distributions of the above test statistics in terms of normal density and its derivatives when the eigenvalues of Ω have multiplicities.

Krishnaiah, Lin, and Wang (1985a) derived the LRT statistics for testing the hypothesis on the rank of B when the underlying distribution is elliptically symmetric. They also investigated the asymptotic distributions of the above statistics, a review of which is given below.

Let E be distributed as an elliptically symmetric distribution with density

$$f(E)=\frac{1}{|\Sigma|^{n/2}}h(tr\,\Sigma^{-1}E'E) \tag{23}$$

where $h(x)$ is a strictly decreasing and differentiable function in x. Also, let

$$\Delta=CB \tag{24}$$

where $C\colon u\times k$ is known and of rank u. Let H_{1r} denote the hypothesis that the rank of Δ is r, whereas H_{2r} denotes the hypothesis that the rows of Δ lie in an r-dimensional plane in p-dimensional space.

Now, let

$$\begin{aligned}\Pi_r(\mathbf{a})=\{&(GF'+\mathbf{ab}')D_{p\times p};\quad GG'=FF'\\ &=I_r,\ D_{p\times p}\text{ positive definite and }\mathbf{b}_{p\times 1}\}.\end{aligned}$$

Then H_{1r} is equivalent to the hypothesis that $\Delta\in\Pi_r(0)$, and H_{2r} is equivalent to $\Delta\in\Pi_r(\mathbf{1})$ where $\mathbf{1}'=(1,\ldots,1)$. Now, let

$$M = C(X'X)^{-1}C'$$
$$\widehat{B} = (X'X)^{-1}X'Y$$
$$S_h(\widehat{B}) = (C\widehat{B})'M^{-1}(C\widehat{B})$$
$$S_f(\widehat{B}) = (C\widehat{B})'\{M^{-1} - M^{-1}\mathbf{1}(\mathbf{1}'M^{-1}\mathbf{1})^{-1}\}C\widehat{B}$$
$$S = Y'(I - X(X'X)^{-1}X')Y.$$

Then, the LRT statistics for testing the hypothesis H_{1r} against $H_{1r'}$ for some $r' > r$, when Σ is known and unknown, are given respectively by

$$T_4 = \frac{h(\phi_{r+1} + \cdots + \phi_s + tr\Sigma^{-1}S)}{h(tr\Sigma^{-1}S)} \tag{25}$$

$$T_5 = \prod_{j=r+1}^{s} (1 + d_j)^{-n/2} \tag{26}$$

where $s = \min(u, p)$, $\phi_1 \geq \cdots \geq \phi_s$ are the first s largest nonzero eigenvalues of $S_h(\widehat{B})\Sigma^{-1}$, and $d_1 \geq \cdots \geq d_s$ the first s largest eigenvalues of $nS_h(\widehat{B})S^{-1}$.

Next, the LRT statistics for testing the hypothesis H_{2r} against $H_{2r'}$ for some $r' > r$ when Σ is known and unknown are given respectively by

$$T_6 = \frac{h(\psi_{r+1} + \cdots + \psi_{\bar{s}} + tr\Sigma^{-1}S)}{h(tr\Sigma^{-1}S)} \tag{27}$$

$$T_7 = \prod_{j=r+1}^{\bar{s}} (1 + l_j)^{-n/2} \tag{28}$$

where $\bar{s} = \min(u - 1, p)$, $\psi_1 \geq \cdots \geq \psi_{\bar{s}}$ are the first $\bar{s}$ largest eigenvalues of $S_f(\widehat{B})\Sigma^{-1}$, and $l_1 \geq \cdots \geq l_{\bar{s}}$ the first $\bar{s}$ largest eigenvalues of $nS_f(\widehat{B})S^{-1}$.

Krishnaiah, Lin, and Wang (1985a) also derived the LRT statistics analogous to T_4, T_5, T_6, and T_7 when the underlying distribution is complex elliptically symmetric. These authors also derived asymptotic joint distributions of $(d_1, \ldots, d_s)$ and $(l_1, \ldots, l_{\bar{s}})$. On the basis of the preceding results, they pointed out that $-2\log T_5$ and $-2\log T_7$ are distributed asymptotically as χ^2. When the underlying distribution is multivariate normal, the asymptotics for T_6 was derived in Rao (1973).

In a number of situations, it may be more realistic to assume that the rows of E are iiesd than to assume the joint distribution of the observations Y is elliptically symmetric. The two situations just described become identical when the underlying distribution is multivariate normal. Krishnaiah, Lin, and Wang (1985a) have derived asymptotic joint distributions of $(d_1, \ldots, d_s)$ and $(l_1, \ldots, l_{\bar{s}})$ when the rows of E are independently and elliptically symmetrically distributed (iesd.) with mean vector 0 and the same dispersion matrix.

IV. RANK OF REGRESSION MATRIX BY MODEL SELECTION METHODS

In the model of Eq. (18), we assume that the rows of E are iid. with mean vector 0 and covariance matrix Σ. Also, let $\Delta = CB$ be as defined in the preceding section, where we discussed the problem of testing the hypothesis that the rank of Δ is r when r is specified. But situations often arise in which the experimenter does not know which of the hypotheses $H_{10}, H_{11}, \ldots, H_{1u}$ to test. In these situations, it is of interest to select one of the models $M_0, M_1, \ldots, M_u$ where M_j denotes the model that the rank of Δ is j. We now give a review of Bai *et al.* (1989) for the determination of the rank of Δ using model selection methods.

Let

$$L(r) = \sum_{j=r+1}^{s} \phi_j + rC_n \tag{29}$$

where $-\frac{1}{2}(\phi_{r+1} + \cdots + \phi_s)$ is the logarithm of the LRT statistic for testing the hypothesis that the rank of Δ is r when Σ is known and the underlying distribution is multivariate normal. The statistics are as defined in the preceding section. Also, C_n satisfies the following conditions:

$$\text{(a)} \quad \lim_{n\to\infty}\left\{\frac{C_n}{\log n}\right\} = \infty, \quad \text{(b)} \quad \lim_{n\to\infty}\left\{\frac{C_n}{\lambda_*}\right\} = 0,$$
$$\text{(c)} \quad \lim_{n\to\infty}\left\{\frac{\lambda_*}{\log n}\right\} = \infty \tag{30}$$

where λ_* denotes the smallest eigenvalue of $X'X$. Then, according to the procedure of Bai *et al.* (1989), the rank of Δ when Σ is known is estimated with $\hat{q}$, where

$$L(\hat{q}) = \min\{L(0), L(l), \ldots, L(s)\}. \tag{31}$$

It is also proved in the work that $\hat{q}$ is a consistent estimate of the rank of Δ.

When Σ is unknown, let

$$L^*(r) = n\sum_{j=r+1}^{s} \log(1 + d_j) + rC_n, \tag{32}$$

where $d_1 \geq \cdots \geq d_s$ are the first s largest eigenvalues of $S_h(\hat{B})S^{-1}$ in the preceding section, and C_n satisfies the following conditions:

$$\lim_{n\to\infty}\left\{\frac{C_n}{\log n}\right\} = \infty \tag{33a}$$

$$\limsup_{n\to\infty}(C_n/n) < \frac{1}{3}\log 2 \tag{33b}$$

$$\lim_{n\to\infty}\left\{\frac{C_n}{\lambda_*}\right\} = 0 \tag{33c}$$

We also make the following assumptions on λ^* (the largest eigenvalue of $X'X$) and λ_*:

$$\lim_{n\to\infty}\left\{\frac{\lambda_*}{\log n}\right\}=\infty \tag{34}$$

$$\lambda^*=O\left(\frac{n\log n}{\log\log n}\right) \tag{35}$$

Then, Bai *et al*. (1989) proposed using $\hat{q}$ as an estimate of the rank of Δ where

$$L^*(\hat{q})=\min\{L^*(0),L^*(l),\ldots,L^*(s)\}$$

They also proved that $\hat{q}$ is a consistent estimate of the rank of Δ.

We may consider alternative model selection criteria similar to those considered by Akaike (1972), Rissanen (1973), and Schwartz (1978) in some other problems.

Next, consider the case when X is also stochastic and the rows of (Y, X) are imnd. with mean vector 0 and unknown covariance matrix. When B is not of full rank, Izenman (1975) considered the problem of estimation of B and the asymptotic distribution of the estimate of B. We can propose model selection procedures, similar to those discussed in present section, to determine the rank of B.

V. REDUCTION OF DIMENSIONALITY UNDER THE FANOVA MODEL

Consider the following two-way classification model with one observation per cell:

$$x_{ij}=\mu+\alpha_i+\beta_j+\eta_{ij}+\varepsilon_{ij} \tag{36}$$

for $i=1,2,\ldots,r$ and $j=1,2,\cdots,s$ where

$$\sum_{i=1}^{r}\alpha_i=\sum_{j=1}^{s}\beta_j=\sum_{i=1}^{r}\eta_{ij}=\sum_{j=1}^{s}\eta_{ij}=0 \tag{37}$$

Here μ,α_i,β_j and η_{ij}, respectively, denote the general mean, the effect due to the ith row, the effect due to the jth column, and interaction in the ith row and jth column. Without loss of generality, we assume that $r\le s$. The problem of finding the rank of the interaction matrix $\boldsymbol{\eta}=(\eta_{ij})$ is of interest and has received attention in the literature. The usual F test statistics to test the hypotheses of no row effect and no column effect were proposed in the literature under the assumption of no interactions. If there is interaction, then the F statistics are distributed as doubly noncentral F distributions instead of central F distributions (see Fang and Krishnaiah, 1984), even when the null hypotheses are true; so the usual tests are no longer valid. Thus, it is of interest to test the hypothesis that the rank of $\boldsymbol{\eta}$ is zero; this problem is known in the literature as testing for additivity. Fisher and MacKenzie (1923), Tukey (1949), and Williams (1952) are the early workers on the problem of testing for additivity when $\boldsymbol{\eta}$ has special structures. When $\boldsymbol{\eta}\neq 0$, knowledge of the rank of $\boldsymbol{\eta}$ helps to estimate the parameters more efficiently. So it is of interest to test for the rank of $\boldsymbol{\eta}$. We now discuss this problem.

Suppose $\boldsymbol{\eta}$ is of rank c. Then it is known, by singular value decomposition of the matrix, that

$$\boldsymbol{\eta}=\theta_1\boldsymbol{\mu}_1\boldsymbol{\nu}_1'+\cdots+\theta_c\boldsymbol{\mu}_c\boldsymbol{\nu}_c' \tag{38}$$

where $\theta_1^2\ge\cdots\ge\theta_c^2$ are the eigenvalues of $\boldsymbol{\eta}\boldsymbol{\eta}'$ and $\boldsymbol{\mu}_j$ and $\boldsymbol{\nu}_j$ the eigenvectors of $\boldsymbol{\eta}\boldsymbol{\eta}'$ and $\boldsymbol{\eta}'\boldsymbol{\eta}$ corresponding to θ_j^2. Now, let $l_1\ge\cdots\ge l_{r-1}$ denote the nonzero eigenvalues of DD' where $D=(d_{ij})$ and $d_{ij}=x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..}$. Gollob (1968) considered the problem of testing the hypotheses $\theta_j=0$, and his tests are based on the assumption that all l_j are distributed independently as χ^2 variables. But the above assumption is not correct. Mandel (1969) proposed heuristically to examine the magnitude of $l_j/\gamma_j\hat{\sigma}^2$ to test for $\theta_j=0$ where $\gamma_j=E(l_j)$ and $\hat{\sigma}^2=(l_{c+1}+\cdots+l_{r-1})/(\gamma_{c+1}+\cdots+\gamma_{r-1})$. But the distributions of the above test statistics not only are complicated but also involve nuisance parameters. Corsten and van Eijnsbergen (1972) derived the following likelihood ratio test. Accept or reject H_0: $\theta_1=\cdots=\theta_c=0$ according to

$$L_1\overset{<}{>}c_{1\alpha} \tag{39}$$

where

$$P[L_1\le c_{1\alpha}|H_0]=(1-\alpha) \tag{40}$$

and $L_1=(l_1+\cdots+l_c)/(l_1+\cdots+l_{r-1})$. For $c=1$, the LRT was derived independently by Johnson and Graybill (1972). Yochmowitz and Cornell (1978) discussed the LRT for testing the hypothesis H_j: $\theta_j=\cdots=\theta_c=0$ against the alternative H_{j+1}: $\theta_j\neq 0$ and $\theta_{j+1}=\cdots=\theta_c=0$. When H_0 is true, it is known (e.g., see Johnson and Graybill, 1972) that $l_1,\ldots,l_{r-1}$ are jointly distributed as the joint distribution of the eigenvalues of the central $(r-1)\times(r-1)$ Wishart matrix W with $(s-1)$ degrees of freedom and $E(W)=(s-l)I_{r-1}$. Schuurmann, Krishnaiah, and Chattopadhyay (1973) derived the exact distribution of $l_1/(l_1+\cdots+l_{r-1})$ and $l_{r-1}/(l_1+\cdots+l_{r-1})$ when H_0 is true and computed some of the percentage points of the above statistic. Krishnaiah and Schuurmann (1974) derived the exact distributions of $l_j/(l_1+\cdots+l_{c-1})$ for $j=2,3,\ldots,c-1$ when H_0 is true. The problem of testing for the rank of $\boldsymbol{\eta}$ can be tackled by using the techniques of principal component analysis. Schuurmann, Krishnaiah, and Chattopadhyay (1973) proposed the following simultaneous test procedure in the spirit of the simultaneous test procedures

of Krishnaiah and Waikar (1971, 1972) in the area of principal component analysis. Accept or reject $\theta_i = 0$ according

$$\frac{l_i}{l_i + \cdots + l_{r-1}} \underset{>}{\overset{<}{}} c_{2\alpha} \tag{41}$$

where

$$P\left\{\frac{l_i}{l_i + \cdots + l_{r-1}} \leq c_{2\alpha} | H_i\right\} = (1 - \alpha) \tag{42}$$

For details of other simultaneous test procedures, the reader is referred to Krishnaiah and Yochmowitz (1980).

We now review a work of Rao (1985) on a more general problem of reduction of dimensionality.

Let Y: $n \times p$ be a random matrix that is distributed as multivariate normal with $E(Y) = M$ and the covariance matrix of $\mathbf{y}$ is $C \otimes \Sigma$, where $\mathbf{y}$ is the vector obtained by writing the rows of Y vertically one below the other, starting from the first, and C is a known positive definite matrix. Also, let S: $p \times p$ be distributed independent of Y as a central Wishart matrix with s degrees of freedom and $E(S) = s\Sigma$. Under the above model, Rao (1985) derived the LRT for testing the hypothesis H_0 that

$$H\colon M = X\psi + \Phi W' + \Gamma \tag{43}$$

where Σ has general structure of the form

$$\Sigma = \sigma_1^2 V_1 V_1' + \cdots + \sigma_f^2 V_f V_f' \tag{44}$$

where $\sigma_1^2, \ldots, \sigma_f^2$ are unknown and V_i: $p \times g_i$ $(i = 1, 2, \ldots, f)$ is a known matrix of rank g_i such that $p = g_1 + \cdots + g_f$.

In Eq. (43), $Xn \times b$ is a known matrix of rank b, W: $p \times c$ a given matrix of rank c, Ψ and Φ are matrices of unknown parameters and Γ is a matrix of specified rank $r \leq \min(p - b, p - c)$. If X is an $n \times 1$ vector of unities and W is the null matrix, the above problem reduces to the problem of specifying the dimensionality of row mean vectors in M considered by Fisher (1939), Fujikoshi (1974), Krishnaiah, Lin, and Wang (1985a), and others. If X is an $n \times 1$ vector of unities and W a $p \times 1$ vector of unities, then H is the hypothesis specifying the rank of interaction in a two-way classification with one observation per cell; this problem was considered when $\Sigma = \sigma^2 I$ and $C = I$.

VI. RANK OF THE COVARIANCE MATRIX OF RANDOM EFFECTS

Consider the one-way components of the covariance model,

$$\mathbf{x}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij} \tag{45}$$

for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, m_i$ where $\boldsymbol{\mu}$ is the general mean vector, $\boldsymbol{\alpha}_i$: $p \times 1$ the vector of random effects, $\boldsymbol{\varepsilon}_{ij}$ the vector of errors and $\mathbf{x}_{ij}$ denotes the jth observation in the ith group. Also, $\boldsymbol{\alpha}_i$ and $\boldsymbol{\varepsilon}_{ij}$ are distributed independently of each other as multivariate normal with $E(\boldsymbol{\alpha}_i) = E(\boldsymbol{\varepsilon}_{ij}) = 0$, and covariance matrices are given by $\Psi = E(\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i')$ and $\Sigma_1 = E(\boldsymbol{\varepsilon}_{ij} \boldsymbol{\varepsilon}_{ij}')$. We also assume that $E(\boldsymbol{\alpha}_i \boldsymbol{\alpha}_j') = 0$ for $i \neq j$, and $E(\boldsymbol{\varepsilon}_{ij} \boldsymbol{\varepsilon}_{i',j'}) = 0$ for $i \neq i'$ and/or $j \neq j'$. The covariance matrix of $\mathbf{x}_{ij}$ is given by Σ_2 where

$$\Sigma_2 = \Psi + \Sigma_1. \tag{46}$$

We assume that Ψ is not of full rank, and we are interested in finding its rank. If the rank of Ψ is r, then there exists a full rank matrix B: $(p - r) \times p$ such that $B\Phi = 0$. If the rank of Φ is zero, then we conclude that there is no difference between the effects of the groups. Knowledge about the rank of Ψ helps us estimate Ψ more efficiently.

When $m_1 = m_2 = \cdots = m_k = m$, the between-group SP matrix and within-group SP matrix are given by S_b and S_w, respectively, where

$$\begin{aligned} S_b &= m \sum_{i=1}^{k} (\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot\cdot})(\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot\cdot})' \\ S_w &= \sum_{i=1}^{k} \sum_{j=1}^{m} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i\cdot})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i\cdot})' \\ m\bar{\mathbf{x}}_{i\cdot} &= \sum_{j=1}^{m} \mathbf{x}_{ij}, \quad mk\bar{\mathbf{x}}\colon = \sum_{i=1}^{k} \sum_{j=1}^{m} \mathbf{x}_{ij} \end{aligned} \tag{47}$$

Then, S_b and S_w are distributed independently as central Wishart matrices with $(k - 1)$ and $k(m - 1)$ degrees of freedom, respectively, $E[S_b/(k-1)] = \Sigma_2$, $E[S_w/k(m-1)] = \Sigma_1$, and $\Sigma_2 = \Sigma_1 + m\Psi$. When the sample sizes are unequal, S_b is not distributed as a Wishart matrix. When all m_i are equal, Anderson (1984b, 1985) has derived the likelihood ratio test statistic for testing the hypothesis that the rank of Ψ is not greater than r. Schott and Saw (1984) derived the likelihood ratio test for $\text{rank}(\Psi) \leq r$ against the alternative $\text{rank}(\Psi) = r + 1$.

We now discuss a more general problem considered by Rao (1983) and Zhao, Krishnaiah, and Bai (1986b). Let S_1 and S_2 be distributed independently as central Wishart matrices with n_1 and n_2 degrees of freedom, respectively, and let $E(S_i/n_i) = \Sigma_i$, for $i = 1, 2$. Also, let $\Sigma_2 = \Gamma + \Sigma_1$ where Γ is a nonnegative definite matrix. Then, we are interested in finding the rank of Γ. Rao (1983) proposed a modified LRT statistic for testing the hypothesis that the rank of Γ is a specified value. We now discuss the model selection method proposed by Zhao, Krishnaiah, and Bai

(1986b) for estimating the rank of Γ. Let $\delta_1 \geq \cdots \geq \delta_p$ denote the eigenvalues of $S_1 S_2^{-1} n_2/n_1$. Also, let

$$L_q = \prod_{i=1+\min(q,\tau)}^{p} \{(\alpha_n + \beta_n \delta_i)^{-n/2} \delta_i^{n\beta_n/2}\} \tag{48}$$

where τ denotes the number of δ_i that are greater than 1, $\alpha_n = n_1/n$, $\beta_n = n_2/n$, and $n = n_1 + n_2$. In addition, let

$$L_{qt} = \prod_{i=1+\min(q,\tau)}^{1+\min(t,\tau)} \{(\alpha_n + \beta_n \delta_i)^{-n/2} \delta_i^{n\beta_n/2}\} \tag{49}$$

Zhao, Krishnaiah, and Bai (1986b) showed that L_q is the likelihood ratio test statistic for testing H_q against the alternative that Γ is arbitrary, and L_{qt} the likelihood ratio test statistic for testing H_q against H_t $(q < t)$ where H_j denotes the hypothesis that the rank of Γ is equal to j. Now let

$$EDC(a, C_n) = -\log L_a + v(a, p) C_n \tag{50}$$

where $v(a, p) = \frac{1}{2} a(2p - a + 1)$ and C_n satisfies the following conditions:

$$\lim_{n\to\infty} \left(\frac{C_n}{n}\right) = 0 \tag{51a}$$

$$\lim_{n\to\infty} \left(\frac{C_n}{\log\log n}\right) = \infty \tag{51b}$$

Zhao, Krishnaiah, and Bai (1986b) estimated the unknown rank of Γ with $\hat{q}$ where

$$EDC(\hat{q}, C_n) = \min\{EDC(0, C_n), \ldots, EDC(p-1, C_n)\} \tag{52}$$

They have also proved that $\hat{q}$ is strongly consistent. The foregoing procedure can be used to draw inference on the rank of the covariance matrix of the vector of random effects in one-way components of the covariance model. When $\Sigma_1 = \sigma^2 I_p$, we can use the techniques of Zhao, Krishnaiah, and Bai (1986a) to find the rank of Γ.

VII. VARIABLE SELECTION IN MULTIVARIATE REGRESSION

In the area of univariate regression analysis, it is of interest to select variables that are important for prediction. Reviews of the literature on some methods of selection of variables are given in Krishnaiah (1982a) and Thompson (1978a, 1978b). In this section, we review procedures for selection of independent variables that are important for the prediction of a set of dependent variables under the classical multivariate regression model.

Consider the multivariate regression model of Eq. (18) where $X = [\mathbf{x}_1, \ldots, \mathbf{x}_q]$ and $\mathbf{x}_i$: $n \times 1$ is a vector of n independent observations on the ith independent variable $\mathbf{x}_i'$. Also, let $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_p]$ where $\mathbf{y}_i$: $n \times 1$ denotes the vector of n independent observations on the ith dependent variable. Then, it is of interest to find which of the variables $\mathbf{x}_1, \ldots, \mathbf{x}_q$ are important. We can use Roy's largest root test, the $T^2_{\max}$ test, or Krishnaiah's finite intersection test for the selection of important variables. Now, let $B' = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$ where $\boldsymbol{\beta}_i$ is of order $p \times 1$. Also, let H_i: $\boldsymbol{\beta}_i = 0$ and

$$T_i^2 = \frac{(n-q)\hat{\boldsymbol{\beta}}_i' S^{-1} \hat{\boldsymbol{\beta}}_i}{e_{ii}} \tag{53}$$

where $\hat{B} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_q)' = (X'X)^{-1}(X'Y)$, $S = (S_{ij}) = Y'(I - X(X'X)^{-1}X')Y$, and $e_{ii}\Sigma$ is the covariance matrix of $\hat{\boldsymbol{\beta}}_i$. According to Roy's largest root test, we accept or reject H_i according to

$$T_i^2 \overset{<}{>} c_\alpha \tag{54}$$

where c_α is chosen such that

$$P\left[(n-q) C_L\left(S_1 S^{-1}\right) \leq q c_\alpha | H\right] = (1 - \alpha) \tag{55}$$

and $H = \cap_{i=1}^{q} H_i$, $C_L(A)$ denotes the largest eigenvalue of A, and $S_1 = Y'X(X'X)^{-1}X'Y = \hat{B}'(X'X)\hat{B}$. Percentage points of c_α are given in Krishnaiah (1980b). If we use $T^2_{\max}$ test (e.g., see Krishnaiah, 1969; and Siotani, 1959), we accept or reject H_0: rank$(B) = 0$ according as

$$T^2_{\max} \overset{<}{>} c_{\alpha 1} \tag{56}$$

where

$$P[T_i^2 \leq c_{\alpha 1}; \quad i = 1, 2, \ldots, q | H] = (1 - \alpha) \tag{57}$$

Approximate values of $c_{\alpha 1}$ can be obtained from the results of Siotani (1959, 1960, 1961) for some cases. We conclude that the independent variable x_i is important or unimportant for the prediction of $(y_1, \ldots, y_p)$ according to whether H_i is rejected or accepted. We now discuss Krishnaiah's finite intersection tests (Krishnaiah, 1965) for the selection of variables. For an illustration of the application of the finite intersection test, the reader is referred to Schmidhammer (1982).

Let Σ_k denote the top-left $k \times k$ corner of $\Sigma = (\sigma_{ij})$ and $\sigma_{k+1}^2 = |\Sigma_{k+1}|/|\Sigma_k|$ for $k = 0, 1, \ldots, p-1$ with $|\Sigma_0| = 1$. Also, let $Y_j = [\mathbf{y}_1, \ldots, \mathbf{y}_j]$, $X_j = [\mathbf{x}_1, \ldots, \mathbf{x}_j]$ and $B_j = [\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_j^*]$ for $j = 1, 2, \ldots, p$, where $\boldsymbol{\beta}_j^*$ is the jth column of B. In addition, let

$$\boldsymbol{\zeta}_j = \Sigma_j^{-1} \begin{pmatrix} \sigma_{1,j+1} \\ \vdots \\ \sigma_{j,j+1} \end{pmatrix} \tag{58}$$

for $j = 1, 2, \ldots, p-1$ and $\boldsymbol{\zeta}_0 = 0$. We know the conditional distribution of $\mathbf{y}_{j+1}$, given that Y_j is multivariate

normal with covariance matrix $\sigma_{j+1}^2 I_n$ and the mean vector

$$E_c(\mathbf{y}_{j+1}) = X\boldsymbol{\eta}_{j+1} + Y_j\boldsymbol{\zeta}_j = [X, Y_j]\begin{pmatrix}\boldsymbol{\eta}_{j+1}\\ \boldsymbol{\zeta}_j\end{pmatrix} \quad (59)$$

where $\boldsymbol{\eta}_{j+1} = \boldsymbol{\beta}_{j+1}^* - B_j\boldsymbol{\zeta}_j$ with the understanding that $\boldsymbol{\eta}_1 = \boldsymbol{\beta}_1$. Now, let H_{ij}: $\mathbf{c}_i'\boldsymbol{\eta}_j = 0$, where $\mathbf{c}_i' = (c_{i1}, \ldots, c_{iq})$ for $i = 1, 2, \ldots, q$ with

$$c_{ih} = \begin{cases} 0, & h \neq i \\ 1, & h = i \end{cases}$$

Then, the hypothesis H_i can be expressed as $H_i = \cap_{j=1}^{p} H_{ij}$. So the problem of testing the hypotheses $H_1, \ldots, H_q$ simultaneously is equivalent to testing the hypotheses H_{ij} simultaneously. Now let

$$F_{ij} = \frac{(\mathbf{c}_i'\hat{\boldsymbol{\eta}}_j)^2(n-j-q+1)}{d_{ij}s_j^2} \quad (60)$$

where $d_{ij}\sigma_j^2$ is the variance of $\mathbf{c}_i'\hat{\boldsymbol{\eta}}_j$, $\hat{\boldsymbol{\eta}}_j$ is the least squares estimate of $\boldsymbol{\eta}_j$ under the model of Eq. (59), and $s_{j+1}^2 = |S_{j+1}|/|S_j|$, where S_j is the top $j \times j$ left-hand corner of S. Then we accept or reject H_{ij} according as

$$F_{ij} \underset{>}{\overset{<}{}} F_\alpha \quad (61)$$

where

$$\begin{aligned} &P[F_{ij} \leq F_\alpha; \quad i = 1, 2, \ldots, q, \quad j = 1, 2, \ldots, p|H] \\ &= \prod_{j=1}^{p} P[F_{ij} \leq F_\alpha; \quad i = 1, 2, \ldots, q|H] \\ &= (1-\alpha) \end{aligned} \quad (62)$$

When H is true, the joint distribution of $F_{1j}, \ldots, F_{qj}$ is a multivariate F distribution with $(1, n-q-j+1)$ degrees of freedom. Evaluation of the probability integrals of the multivariate F distribution was discussed in Krishnaiah and Armitage (1970). The hypothesis H_i is accepted if $H_{i1}, \ldots, H_{ip}$ are accepted, and it is rejected otherwise. If H_i is rejected, then we conclude that the independent variable x_i is important for prediction of the set $(y_1, \ldots, y_p)$ of dependent variables. One may use the step-down procedure proposed by J. Roy (1958) also, but the lengths of the confidence intervals associated with the finite intersection tests are shorter than the lengths of the corresponding confidence intervals associated with the step-down procedure. Fujikoshi (1985) proposed a procedure, based on an information theoretic criterion, to select a subset of variables that are important for discrimination. Rao (1948) proposed a procedure to find whether the addition of some independent variables make a significant contribution in the prediction of dependent variables.

VIII. VARIABLE SELECTION IN DISCRIMINANT ANALYSIS

We now discuss the stepwise procedures for the selection of variables in the area of discriminant analysis for several groups. These procedures are used widely since computer programs for the implementation of these procedures are available in the BMDP and SPSS packages. Stepwise procedures for the selection of variables in discriminant analysis were proposed in the literature in a way similar to that for the corresponding procedures in regression analysis (Krishnaiah, 1982a).

Consider the following model:

$$E(\mathbf{y}_j) = A\boldsymbol{\theta}_j \quad (63)$$

where

$$A' = (A_1' \cdots A_k') \quad (64)$$

In the matrix A_i: $n_i \times k$, the elements in the ith column are equal to one, and other elements in the matrix are zero. Also, $\boldsymbol{\theta}_j' = (\boldsymbol{\mu}_{1j}, \ldots, \boldsymbol{\mu}_{kj})$, $\mathbf{y}_j' = (x_{1j1}, \ldots, x_{1jn_1}, \cdots, x_{kj1}, \cdots, x_{kjn_k})$, and x_{ijt} denotes observation on the jth variable, tth individual, and ith group. Let H_j: $C\boldsymbol{\theta}_j = 0$ where

$$C = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix} : (k-1) \times k \quad (65)$$

Let F_j denote the usual F statistic used for testing the hypothesis H_j. Then,

$$F_j = \frac{b_{jj}(n-k)}{w_{jj}(k-1)} \quad (66)$$

where $W = (w_{ij})$ and $B = (b_{ij})$ are the within-group SP matrix and between-group SP matrix, respectively. The likelihood ratio statistic for testing H_j is given by $\Lambda(x_j)$, where

$$\Lambda(x_j) = \frac{w_{jj}}{t_{jj}} \quad (67)$$

and $t_{jj} = b_{jj} + w_{jj}$. Obviously,

$$F_j = \frac{[1-\Lambda(x_j)](n-k)}{(k-1)} \quad (68)$$

If $\max(F_1, \ldots, F_p) \leq F_{1\alpha}$, we declare that none of the variables are important for discrimination and do not proceed further. Otherwise, we select the variable corresponding to the maximum of $F_1, \ldots, F_p$ as the most important. For example, let the variable be x_1.

At the second stage we test to find out as to whether any of the remaining variables $x_2, x_3, \ldots, x_p$ give additional information for discrimination between the

populations. A measure of the degree of additional information is provided by

$$\Lambda_{j\cdot 1} = \frac{\Lambda(x_1, x_j)}{\Lambda(x_1)} \tag{69}$$

where

$$\Lambda(x_1, x_j) = \frac{\begin{vmatrix} w_{11} & w_{1j} \\ w_{j1} & w_{jj} \end{vmatrix}}{\begin{vmatrix} t_{11} & t_{1j} \\ t_{j1} & t_{jj} \end{vmatrix}} \tag{70}$$

In Eq. (69), $\Lambda(x_1, x_j)$ is the likelihood ratio test statistic for testing the hypothesis that the mean vectors of (x_1, x_j) are the same in all populations. It can be viewed as a measure of the discriminating ability of x_1 and x_j, whereas $\Lambda(x_1)$ is a measure of the degree of discrimination of the variable x_1. As the value of $\Lambda(x_1, x_j)$ decreases, the discriminating ability of x_1 and x_j increase. We can write $\Lambda_{j\cdot 1}$ as

$$\Lambda_{j\cdot 1} = \frac{w_{j\cdot 1}}{t_{j\cdot 1}} \tag{71}$$

where $w_{j\cdot 1} = w_{jj} - w_{j1}w_{11}^{-1}w_{1j}$ and $t_{j\cdot 1} = t_{jj} - t_{j1}t_{11}^{-1}t_{1j}$. Now let

$$F_{j\cdot 1} = \frac{b_{j\cdot 1}(n-k-1)}{w_{j\cdot 1}(k-1)} \tag{72}$$

where $b_{j\cdot 1} = t_{j\cdot 1} - w_{j\cdot 1}$ is the adjusted between group sum of squares. We can write Eq. (72) as

$$F_{j\cdot 1} = \frac{(n-k-1)}{(k-1)} \frac{1-\Lambda_{j\cdot 1}}{\Lambda_{j\cdot 1}} \tag{73}$$

The preceding statistic proposed by Rao (1973) is nothing but the statistic used to test the hypothesis

$$H_{j\cdot 1}\colon\ \mu_{1j} - \beta_{j\cdot 1}\mu_{11} = \cdots = \mu_{kj} - \beta_{j\cdot 1}\mu_{k1} \tag{74}$$

where $\beta_{j\cdot 1} = \alpha_{j1}\sigma_{11}^{-1}$. If $\max(F_{2\cdot 1}, \ldots, F_{p\cdot 1} \le F_{2\alpha}$, we declare that none of the variables $x_2, x_3, \ldots, x_p$ are important; here $F_{2\alpha}$ is the upper $100\alpha\%$ point of the central F distribution with $(k-1, n-k-1)$ degrees of freedom. If $\max(F_{2\cdot 1}, \ldots, F_{p\cdot 1}) > F_{2\alpha}$, the variable corresponding to the maximum of $F_{2\cdot 1}, \ldots, F_{p\cdot 1}$ is declared to be important. For simplicity of notation, let us assume that this variable is, say, x_2.

After having selected x_2, we test whether the variable (in this case x_1) selected at the first stage is also good for discrimination in the presence of the variable x_2, which is the third step. This can be tested by using the following test statistic:

$$F_{1\cdot 2} = \frac{b_{1\cdot 2}(n-k-1)}{w_{1\cdot 2}(k-1)} \tag{75}$$

where $t_{1\cdot 2} = t_{11} - t_{12}t_{11}^{-1}t_{21}$, $w_{1\cdot 2} = w_{11} - w_{12}w_{11}^{-1}w_{21}$, and $b_{1\cdot 2} = t_{1\cdot 2} - w_{1\cdot 2}$. We decide to retain or exclude x_1 from the selected subset according to

$$F_{1\cdot 2} \lessgtr F_{2\alpha}. \tag{76}$$

Here we note that

$$F_{1\cdot 2} = \frac{(n-k-1)}{(k-1)} \frac{1-\Lambda_{1\cdot 2}}{\Lambda_{1\cdot 2}} \tag{77}$$

where

$$\Lambda_{1\cdot 2} = \frac{w_{1\cdot 2}}{t_{1\cdot 2}} \tag{78}$$

$w_{1\cdot 2} = w_{11} - w_{12}w_{11}^{-1}w_{21}$ and $t_{1\cdot 2} = t_{11} - t_{12}t_{11}^{-1}t_{21}$. If $\Lambda_{1\cdot 2}^{*} = 1/\Lambda_{1\cdot 2}$, then

$$F_{1\cdot 2} = \frac{(n-k-1)}{(k-1)}(\Lambda_{1\cdot 2}^{*} - 1) \tag{79}$$

In the fourth step, we either select one of the variables $x_3, \ldots, x_p$ or decide not to select any more on the basis of the discriminating ability of these variables individually in the presence of x_1 and x_2. If we discard x_1 at the third step, then we consider the discriminating ability of the variables $x_3, \ldots, x_p$ in the presence of x_2 only. This procedure is continued until a decision is made not to select any more variables or until all the variables are selected. Suppose that after a few stages we selected $x_3, \ldots, x_j$ and that x_j is the latest addition to the selected subset. Then we test whether $x_3, \ldots, x_j$ are individually important in the presence of the remaining variables. For example, we test whether x_4 is important in the presence of the variables $x_3, x_5, x_6, \ldots, x_j$. The statistic used to test whether x_i $(i = 3, 4, \ldots, j-1)$ is important is given by

$$F_{i\cdot(3,4,\ldots,j)} = \frac{b_{i\cdot(3,4,\ldots,j)}}{w_{i\cdot(3,4,\ldots,j)}} \frac{(n-k-j+3)}{(k-1)} \tag{80}$$

with the understanding that the suffix i does not occur in the set $(3, 4, \ldots, j)$. Let

$$\Lambda_{i\cdot(3,4,\ldots,j)} = \frac{\Lambda(x_3, \ldots, x_i, x_{i+1}, \ldots, x_j)}{\Lambda(x_3, \ldots, x_{i-1}, x_{i+1}, \ldots, x_j)} \tag{81}$$

where $\Lambda(x_3, \ldots, x_i, x_{i+1}, \ldots, x_j)$ is the ratio of the determinant of the within-group SP matrix based on the variables $(x_3, \ldots, x_i, x_{i+1}, \ldots, x_j)$ and the determinant of the total SP matrix based on the same variables. Similarly, $\Lambda(x_3, \ldots, x_{i-1}, x_{i+1}, \ldots, x_j)$ can be defined. So,

$$\Lambda_{i\cdot(3,4,\ldots,j)} = \frac{w_{i\cdot(3,4,\ldots,j)}}{t_{i\cdot(3,4,\ldots,j)}} \tag{82}$$

Hence,

$$F_{i\cdot(3,4,\ldots,j)} = \frac{(n-k-f+3)}{(k-1)} \frac{(1-\Lambda_{i\cdot(3,4,\ldots,j)})}{\Lambda_{i\cdot(3,4,\ldots,j)}} \tag{83}$$

The variable x_i is retained or excluded according to whether $F_{i\cdot(3,4,\ldots,j)}$ is greater than or less than the upper $100\alpha\%$ point of the central F distribution with $(k-1, n-k-f+3)$ degrees of freedom. At any stage, we can test whether all the variables selected together will

discriminate between the groups by using many standard procedures. For example, suppose $x_1, x_2, \ldots, x_q$ are selected. Then we compute B_{11} and W_{11}, which are, respectively, the between-group SP matrix and the within-group SP matrix based on $x_1, \ldots, x_q$. They are given by

$$B_{11} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1q} \\ b_{21} & b_{22} & \cdots & b_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ b_{q1} & b_{q2} & \cdots & b_{qq} \end{bmatrix}$$

$$W_{11} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1q} \\ w_{21} & w_{22} & \cdots & w_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ w_{q1} & w_{q2} & \cdots & w_{qq} \end{bmatrix} \tag{84}$$

We can test whether the variables $x_1, \ldots, x_q$ together will discriminate between the populations by computing various functions of the eigenvalues of $B_{11}W_{11}^{-1}$. Some of these functions are $C_L(B_{11}W_{11}^{-1})$, $tr(B_{11}W_{11}^{-1})$, $tr(B_{11}(B_{11}+W_{11})^{-1})$, and $|B^{11}(B_{11}+W_{11})^{-1}|$. One can also use finite intersection tests.

We now have a critical look at the stepwise procedure for the selection of variables. At the first stage of the procedure, we choose the critical value $F_{1\alpha}$ such that

$$P[F_j \le F_{1\alpha} | H_j] = (1-\alpha) \tag{85}$$

Here the hypotheses $H_1, \ldots, H_q$ are tested individually. Since the decision whether or not to select any variable at the first stage is based on whether or not all the hypotheses are accepted simultaneously, it would be a natural thing to test them simultaneously and choose the critical value $F_{1\alpha}$ such that

$$P\left[F_j \le F_{1\alpha}; j = l, 2, \ldots, p \Big| \bigcap_{j=1}^{p} H_j\right] = (1-\alpha) \tag{86}$$

The joint distribution of $F_1, \ldots, F_p$ not only is complicated but also involves nuisance parameters. But we can use Bonferroni's inequality to compute an upper bound on $F_{1\alpha}$. At the first stage, we select one variable only as the most important, and no decision is made about other variables. But this most important variable may be discarded at a later stage. So there is some inconsistency in this method that will be discussed later.

At the second stage, the critical value $F_{2\alpha}$ is chosen such that

$$P[F_{j\cdot 1} \le F_{2\alpha} | H_{j\cdot 1}] = (1-\alpha) \tag{87}$$

We go to the second stage if and only if $\max(F_1, \ldots, F_p) \ge F_{1\alpha}$. So, at the second stage, we have to compute the following conditional probabilities instead of Eq. (85) even if we are testing the hypotheses $H_{j\cdot 1}$ individually:

$$P[F_{j\cdot 1} \le F_{2\alpha} | \max(F_1, \ldots, F_p) \ge F_{1\alpha}] \tag{88}$$

It is quite complicated to compute the foregoing probabilities. Apart from it, we have to test $H_{2\cdot 1}, \ldots, H_{p\cdot 1}$ simultaneously, instead of testing them individually. At the second stage, we select the variable (say, x_2) corresponding to $\max(F_{2\cdot 1} \cdots, F_{p\cdot 1})$. The statistic $F_{j\cdot 1}$ for any given j is useful for testing whether the variable x_j gives additional information for discrimination between the groups in presence of the important variable x_1. But the variable x_1, which is declared to be the most important at the first stage, may be discarded as being unimportant at a later stage; so the procedure may not be meaningful. Apart from it, the choice of the critical values is very arbitrary, and we cannot say what the Type I error of this procedure is. In view of the points raised earlier, we do not recommend the use of the foregoing stepwise procedures. Krishnaiah (1982a) discussed the disadvantages of using forward selection and backward selection procedures for selection of variables under univariate regression models. Similar criticism applies for forward selection and backward selection procedures for selection of variables in discriminant analysis.

IX. TESTS FOR RANK OF CANONICAL CORRELATION MATRIX

It is known that the multiple correlation coefficient is the maximum correlation between a variable and linear combinations of a set of variables. Hotelling (1935, 1936) generalized this concept to two sets of variables $\mathbf{x}_1'$: $1 \times p_1$ and $\mathbf{x}_2'$: $1 \times p_2$ and introduced canonical correlation analysis. Canonical correlation analysis is useful in studying the relationship between the two sets of variables. Let the covariance matrix of $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$ be Σ, where

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{89}$$

and Σ_{ii}: $p_i \times p_i$ is the covariance matrix of $\mathbf{x}_i$. Then $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ is known to be the canonical correlation matrix. Without loss of generality, we assume that $p_1 < p_2$ and $\rho_1^2 \ge \cdots \ge \rho_{p_1}^2$ denote the eigenvalues of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Here $\rho_1, \ldots, \rho_{p_1}$ are known as canonical correlations where ρ_i is the positive square root of ρ_i^2. Now let α_i and β_i denote the eigenvectors of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ and $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, respectively, corresponding to ρ_i^2. Then $\alpha_i'\mathbf{x}_1, \ldots, \alpha_{p_1}'\mathbf{x}_1$ and $\beta_i'\mathbf{x}_2, \ldots, \beta_{p_1}'\mathbf{x}_2$ are known as canonical variables. One of the important problems in the area of canonical correlation analysis is to find the number of canonical correlations that are significantly different from zero. In this section, we discuss some procedures

for testing the hypothesis on the rank of the canonical correlation matrix when the underlying distribution is multivariate normal.

Let X: $n \times p$ be a random matrix such that $E(X) = 0$ and $E(X'X) = n\Sigma$. Also let

$$S = X'X = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \tag{90}$$

where S_{ij} is of order $p_i \times p_j$. In addition, let $r_1^2 \geq \cdots \geq r_{p_1}^2$ denote the eigenvalues of $S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}$. Then $r_1, \ldots, r_{p_1}$ are known as the sample canonical correlations where r_i is the positive square root of r_i^2. Various functions of $r_1^2, \ldots, r_{p_1}^2$ were proposed in the literature as test statistics for determination of the rank of the canonical correlation matrix. We review these procedures in this section.

We first assume that the rows of X are imnd. In this case, Bartlett (1948) proposed a procedure for testing the hypothesis H_t, where H_t denotes $\rho_{t+1}^2 = \cdots = \rho_{p_1}^2 = 0$; he also derived the asymptotic distribution of the preceding statistic. Fujikoshi (1974) showed that the foregoing test statistic is the LRT statistic. Hsu (1941b) derived the asymptotic joint distribution of the sample canonical correlations when H_t is true. When the population canonical correlations $\rho_1, \ldots, \rho_{p_1}$ have multiplicities, and none of them is equal to zero, Fujikoshi (1978) derived the nonnull distribution of a single function of the sample canonical correlations, whereas Krishnaiah and Lee (1979) derived the asymptotic joint distribution of functions of the sample canonical correlations. The expressions derived by Krishnaiah and Lee involve multivariate normal density and multivariate Hermite polynomials. When the underlying distribution is not multivariate normal, Fang and Krishnaiah (1981, 1982) obtained results analogous to those obtained in the paper of Krishnaiah and Lee.

Now, let the joint distribution of the elements of X be elliptically symmetric, with density given by

$$f(X) = |\Sigma|^{-n/2} h(tr\,\Sigma^{-1} X'X) \tag{91}$$

Then, Krishnaiah, Lin, and Wang (1985b) showed that the LRT statistic for testing the hypothesis $\rho_{t+1} = \cdots = \rho_{p_1} = 0$ is given by

$$L(k) = \prod_{j=t+1}^{p_1} \left(1 - r_j^2\right)^{n/2}. \tag{92}$$

So the LRT statistic is the same as when the underlying distribution is multivariate normal. They also noted that the distribution of any function of $r_1^2, \ldots, r_{p_1}^2$ is independent of the form of the underlying distribution as long as the underlying distribution belongs to the family of elliptical distributions. We now review some of the work reported in the literature on canonical correlation analysis when it is assumed that the observations are iiesd. with the common density

$$f(\mathbf{x}) = |\Sigma|^{-1/2} h(\mathbf{x}'\Sigma^{-1}\mathbf{x}). \tag{93}$$

Now, let

$$c_i = \frac{\sqrt{n}\left(r_i^2 - \rho_i^2\right)}{2\rho_i^2\left(1 - \rho_i^2\right)}. \tag{94}$$

Then Muirhead and Waternaux (1980) showed that $c_1, \ldots, c_{p_1}$ are asymptotically distributed independently as normal, with mean 0 and variance $(\kappa + 1)$ when $\rho_1^2, \ldots, \rho_{p_1}^2$ are distinct. This is a special case of a result of Fang and Krishnaiah (1981, 1982). Krishnaiah, Lin, and Wang (1985b) derived the asymptotic joint distribution of the sample canonical correlations when the population canonical correlations have multiplicities and the last few population canonical correlations are zero. In particular, they showed that the joint asymptotic distribution of $(nr_{s+1}^2/\kappa + 1), \ldots, (nr_{p_1}^2/\kappa + 1)$, when H_s: $\rho_{s+1}^2 = \cdots = \rho_{p_1}^2 = 0$, is the same as the joint distribution of the eigenvalues of the central Wishart matrix $W_{p_1 - s}$ with $(p_2 - s)$ degrees of freedom and $E(W_{p_1 - s}) = (p_2 - s) I_{p_1 - s}$. This result is useful in the implementation of certain test procedures for H_s when the sample size is large. For example, we can use r_{s+1}^2 or $(r_{s+1}^2 + \cdots + r_{p_1}^2)$ as a test statistic for H_s.

We now discuss the problem of testing for the rank of the canonical correlation matrix under the correlated multivariate regression equations (CMRE) model considered by Kariya, Fujikoshi, and Krishnaiah (1984). Consider the CMRE model,

$$Y_i = X_i \boldsymbol{\theta}_i + E_i, \tag{95}$$

for $i = 1, 2$. In this model, the rows of (E_1, E_2) are imnd. with mean vector 0 and covariance matrix Σ, where

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{96}$$

and Σ_{ij} is of order $p_i \times p_j$. Also, X_i: $n \times r_i$ is the design matrix and $\boldsymbol{\theta}_{ii}$: $r_i \times p_i$ is the matrix of unknown parameters for $i = 1, 2$. Without loss of generality, we assume that $p_1 \leq p_2$. Now, let

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \tag{97}$$

where $S_{ij} = Y_i' Q_i Q_j Y_j$ and $Q_i = I - X_i (X_i' X_i)^{-1} X_i'$. Also, let $R = S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}$. Kariya, Fujikoshi, and Krishnaiah (1984) investigated the problem of testing the hypothesis that $\rho_t^2 = \cdots = \rho_{p_1}^2 = 0$. They also derived the asymptotic distributions of three statistics in the null case and under local alternatives. We can test the hypothesis

that $\rho_t^2 = \cdots = \rho_{p_1}^2 = 0$ by considering suitable functions of $r_t^2, \ldots, r_{p_1}^2$, such as $r_t^2, r_t^2 + \cdots + r_{p_1}^2$, and so on, where $r_1^2 \geq \cdots \geq r_{p_1}^2$ are the eigenvalues of the sample canonical correlation matrix $S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}$.

For an application of the canonical correlation analysis in econometrics, the reader is referred to Hannan (1967) and Chow and Ray-Chowdhuri (1967).

X. MODEL SELECTION FOR RANK OF CANONICAL CORRELATION MATRIX

Let $X' = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$: $p \times n$ be a random matrix whose columns are iimnd. with common mean vector 0 and covariance matrix Σ. Let $\mathbf{x}_i$ and Σ be partitioned as $\mathbf{x}_i' = (\mathbf{x}_{i1}', \mathbf{x}_{i2}')$ and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{98}$$

where Σ_{ij} is of order $p_i \times p_j$ and $\mathbf{x}_{ij}$ is of order $p_j \times 1$. Let $\rho_1^2 \geq \cdots \geq \rho_s^2$ denote the first largest s eigenvalues of the population canonical correlation matrix $\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$, where $s = \min(p_1, p_2)$. Also, let $r_1^2 \geq \cdots \geq r_s^2$ denote the first largest s eigenvalues of $S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}$, where

$$S = \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j' = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \tag{99}$$

and S_{ij} is of order $p_i \times p_j$. Let M_k $(k = 0, 1, \ldots, s)$ denote the model for which $\text{rank}(\Sigma_{12}) = k$, that is, the number of nonzero canonical correlations is equal to k. Also, let H_k denote the hypothesis that $\text{rank}(\Sigma_{12}) = k$. Let $L(k)$ denote the likelihood ratio test statistic for H_k. Then,

$$\log L(k) = \frac{n}{2} \sum_{i=k+1}^{s} \log\left(1 - r_i^2\right) \tag{100}$$

Now, let

$$G(k) = -\log L(k) + k C_n \tag{101}$$

where C_n satisfies the following conditions:

$$\lim_{n \to \infty} \left\{ \frac{C_n}{n} \right\} = 0 \tag{102a}$$

$$\liminf_{n \to \infty} \left\{ \frac{C_n}{\log \log n} \right\} = \infty \tag{102b}$$

Let q denote the true rank of Σ_{12}. Then Bai *et al.* (1989) proposed to use $\hat{q}$ as an estimate of q, where $\hat{q}$ is given by

$$G(\hat{q}) = \min\{G(0), \ldots, G(s)\} \tag{103}$$

These authors also showed that $\hat{q}$ is a strongly consistent estimate of q. Now, assume that the assumption of normality is violated, but $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are iid. vectors with $E(\mathbf{x}_1) = 0$, $E(\mathbf{x}_1 \mathbf{x}_1') = \Sigma$, and $E(\mathbf{x}_1' \mathbf{x}_1)^2 < \infty$. In this case, $L(k)$ need not be the LRT statistic for H_k, but we can still use it as in Eq. (101). Then Bai *et al.* (1989) showed that $\hat{q}$ in Eq. (103) is still a strongly consistent estimate of q under certain moment conditions.

XI. VARIABLE SELECTION IN CANONICAL CORRELATION ANALYSIS

Let us consider the set $(\mathbf{x}_1', \mathbf{x}_2')$ of $p_1 + p_2$ variables. We wish to select a set of r_2 important variables from the $\mathbf{x}_2$ set on the basis of the degree of dependence with the $\mathbf{x}_1$ set. There are $\binom{p_2}{r_2}$ sets. Let these sets be denoted by $\mathbf{x}_{f_i}$ and let the sample canonical correlation matrix between $\mathbf{x}_1$ set and $\mathbf{x}_{f_i}$ set be denoted by $S_{11}^{-1} S_{1f_i} S_{f_i f_i}^{-1} S_{f_i 1}$. We use the largest root of the canonical correlation matrix as a criterion to select the variables. We declare that none of these sets are important if

$$\max_i c_L\left(S_{11}^{-1} S_{1f_i} S_{f_i f_i}^{-1} S_{f_i 1}\right) \leq c_\alpha \tag{104}$$

where $c_L(A)$ denotes the largest eigenvalue of A. If

$$\max_i c_L\left(S_{11}^{-1} S_{1f_i} S_{f_i f_i}^{-1} S_{f_i 1}\right) > c_\alpha \tag{105}$$

then the set corresponding to $\max_i c_L(S_{11}^{-1} S_{1f_i} S_{f_i f_i}^{-1} S_{f_i 1})$ is declared to be the most important. The critical value c_α is chosen such that

$$P\left[\max_i c_L\left(S_{11}^{-1} S_{1f_i} S_{f_i f_i}^{-1} S_{f_i 1}\right) \leq c_\alpha | H\right] = (1 - \alpha) \tag{106}$$

and H: $\Sigma_{12} = 0$. In other words, the critical value c_α is chosen such that the probability of declaring that none of the $\binom{p_2}{r_2}$ sets is important when, in fact, none of the variables in the $\mathbf{x}_2$ set is correlated with the $\mathbf{x}_1$ set, which is $(1 - \alpha)$. But the distribution of $\max_i c_L(S_{11}^{-1} S_{1f_i} S_{f_i f_i}^{-1} S_{f_i 1})$ is very complicated to derive. So we use the following bound to get an approximate value of c_α:

$$\begin{aligned} &P\left[c_L\left(S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}\right) \leq c_\alpha | H\right] \\ &\leq P\left[\max_i c_L\left(S_{11}^{-1} S_{1f_i} S_{f_i f_i}^{-1} S_{f_i 1}\right) \leq c_\alpha | H\right) = (1 - \alpha) \end{aligned} \tag{107}$$

We now discuss an alternative procedure for the selection of the best subset of r_2 variables from the $\mathbf{x}_2$ set and let $\mathbf{x}_{f_i}$ $[i = 1, 2, \ldots, \binom{p_2}{r_2}]$ denote a subset of r_2 variables from the p_2 variables $\mathbf{x}_2$. As before, let $\Sigma_{11}^{-1} \Sigma_{1f_i} \Sigma_{f_i f_i}^{-1} \Sigma_{f_i 1}$ denote the canonical correlation matrix connected with the $\mathbf{x}_1$ set and the $\mathbf{x}_{f_i}$ set. Let ψ_i denote a suitable function of the eigenvalues of the above matrix. Also, let $\hat{\psi}_i$ denote the corresponding function of the eigenvalues of $S_{11}^{-1} S_{1f_i} S_{f_i f_i}^{-1} S_{f_i 1}$. In addition, let $\psi_1, \ldots, \psi_{p_0}$ be

ordered as $\psi_{[1]}, \ldots, \psi_{[p_0]}$, where $p_0 = \binom{p_2}{r_2}$. Then the subset associated with the maximum of $\hat{\psi}_1, \ldots, \hat{\psi}_{p_0}$ is declared to be the best subset. Suppose $\hat{\psi}_i$ is the largest of $\hat{\psi}_j$. In this case, the probability of correct decision is given by the probability of $\hat{\psi}_i$ being greater than $\hat{\psi}_j$ $(j = 1, \ldots, i-1, i+1, \ldots, p_0)$ when ψ_i is greater than ψ_j for $j \neq i$. This probability involves nuisance parameters. One may use bounds that are free from nuisance parameters.

We now discuss the problem of studying the effect of additional variables on the canonical correlations. Consider two sets of variables, $\mathbf{x}_1$: $p_1 \times 1$ and $\mathbf{y}_1$: $q_1 \times 1$. Without loss of generality, we assume that $p_1 \leq q_1$. Suppose the sets of variables $\mathbf{x}_1$ and $\mathbf{y}_1$ are augmented to $\mathbf{x}$: $p \times 1$ and $\mathbf{y}$: $q \times 1$ by adding extra sets of variables $\mathbf{x}_2$: $p_2 \times 1$ and $\mathbf{y}_2$: $q_2 \times 1$, respectively. Also, we assume that $(\mathbf{x}', \mathbf{y}')$ is distributed as multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, where

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \tag{108}$$

and Σ_{xx} is the covariance matrix of $\mathbf{x}$. Let $\rho_1 \geq \cdots \geq \rho_{p_1}$ denote the canonical correlations between the sets $\mathbf{x}_1$ and $\mathbf{y}_1$ and let $\tilde{\rho}_1 \geq \cdots \geq \tilde{\rho}_{p_1}$ denote the canonical correlations between $\mathbf{x}$ and $\mathbf{y}$. Also, let $\delta_\alpha = \tilde{\rho}_\alpha - \rho_\alpha$ for $\alpha = l, 2, \ldots, p_1$. Then $\delta_\alpha > 0$. Next, let

$$S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} \tag{109}$$

denote the sample covariance matrix based on $(n+1)$ observations on $(\mathbf{x}', \mathbf{y}')$, and the sample canonical correlations $\tilde{r}_1 \geq \cdots \geq \tilde{r}_p$ are the positive square roots of the eigenvalues of $S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx}$. Similarly, let $r_1 \geq \cdots r_{p_1}$ denote the sample canonical correlations based on $(n+1)$ observations on $(\mathbf{x}', \mathbf{y}')$.

Now, let $f(d_1, \ldots, d_{p_1})$ be a continuously differentiable function in a neighborhood of $\mathbf{d} = \boldsymbol{\delta}$, where $\mathbf{d}' = (d_1, \ldots, d_{p_1})$ and $\boldsymbol{\delta}' = (\delta_1, \ldots, \delta_{p_1})$. Then Fujikoshi, Krishnaiah, and Schmidhammer (1987) showed that $\sqrt{n}\{f(d_1, \ldots, d_{p_1}) - f(\delta_1, \ldots, \delta_{p_1})\}$ is distributed asymptotically as normal with mean zero and certain variance σ^2. When $f(d_1, \ldots, d_{p_1}) = d_1$ and $p_1 = p$ or $q_1 = q$, the preceding result was derived by Wijsman (1986). The result of Fujikoshi, Krishnaiah, and Schmidhammer (1987) can be used to find whether the addition of new variables to one or both of the sets $\mathbf{x}_1$ and $\mathbf{y}_1$ will have an effect on functions of the canonical correlations. For example, we can draw an inference as to whether the addition of variables will increase the values of the largest canonical correlation, the sum of the canonical correlations, and so on. If there is no significant increase, we conclude that the new variables are not important in explaining the association between the two sets of variables. Fujikoshi (1985) proposed a procedure based on an information theoretic criterion to select best variables in canonical correlation analysis.

XII. DIMENSIONALITY AND DEPENDENCE IN TWO-WAY CONTINGENCY TABLE

Consider a two-way contingency table, and let p_{ij} $(i = 1, 2, \ldots, r+1;\ j = 1, 2, \ldots, s+1)$ denote the probability of an observation falling in the ith row and jth column. Without loss of generality, we assume that $r \leq s$. We consider the model

$$p_{ij} = p_{i\cdot} p_{\cdot j} \zeta_{ij} \tag{110}$$

where $p_{i\cdot} = p_{i1} + \cdots + p_{i,s+1}$, $p_{\cdot j} = p_{1j} + \cdots + p_{r+1,j}$, and $\boldsymbol{\zeta} = (\zeta_{ij})$ is an unknown matrix. In a number of situations, we are interested in studying the structure of dependence between rows and columns if $p_{ij} \neq p_{i\cdot} p_{\cdot j}$. If we know the structure of dependence, we can estimate the unknown parameters more efficiently. Now, let $F = (f_{ij})$, where $f_{ij} = p_{ij} / \sqrt{p_{i\cdot} p_{\cdot j}}$. From the singular value decomposition of the matrix, it is known (e.g., see Lancaster, 1969) that

$$F = \boldsymbol{\xi}_0^* \boldsymbol{\eta}_0^{*\prime} \rho_0 + \sum_{u=1}^{r} \rho_u \boldsymbol{\xi}_u^* \boldsymbol{\eta}_u^{*\prime} \tag{111}$$

where $\rho_0 \geq \cdots \geq \rho_r$ are the positive square roots of the eigenvalues of FF', $\boldsymbol{\xi}_u^*$ is the eigenvector of FF' corresponding to ρ_u^2 and $\boldsymbol{\eta}_u^*$ is the eigenvector of $F'F$ corresponding to ρ_u^2. Here $\rho_0 = 1$, $\boldsymbol{\xi}_0^* = (\sqrt{p_{1\cdot}}, \ldots, \sqrt{p_{r+1\cdot}})'$, and $\boldsymbol{\eta}_0^* = (\sqrt{p_{\cdot 1}}, \ldots, \sqrt{p_{\cdot s+1}})'$. We now review the work of O'Neill (1978a, 1978b, 1981) and Bhaskara Rao, Krishnaiah, and Subramanyam (1985) for testing for the rank of the matrix $\boldsymbol{\zeta}$. We also review the work of Bai *et al.* (1989) for determination of the rank of $\boldsymbol{\zeta}$ by using model selection methods.

Let n_{ij} denote the frequency in the ith row and jth column, $n_{i\cdot} = n_{i1} + \cdots + n_{i,s+1}$, $n_{\cdot j} = n_{1j} + \cdots + n_{r+l,j}$ and $n = \Sigma\Sigma n_{ij}$. Also, let $B = (b_{ij})$ where $b_{ij} = n_{ij} / \sqrt{n_{i\cdot} n_{\cdot j}}$. Now, let $\hat{\rho}_0^2 \geq \cdots \geq \hat{\rho}_r^2$ denote the eigenvalues of BB' where $\hat{\rho}_0 = 1$. We assume that n is fixed and the joint distribution of the cell frequencies is given by

$$n! \prod_{ij} \frac{1}{n_{ij}!} p_{ij}^{n_{ij}} \tag{112}$$

The classical test statistic for testing the hypothesis $p_{ij} = p_{i\cdot} p_{\cdot j}$ of independence is given by

$$\chi_0^2 = \sum_{i=1}^{r+1} \sum_{j=1}^{s+1} \frac{[n n_{ij} - n_{i\cdot} n_{\cdot j}]^2}{n^2 n_{i\cdot} n_{\cdot j}} \tag{113}$$

When the null hypothesis is true, χ_0^2 is distributed asymptotically as χ^2 with rs degrees of freedom. The preceding hypothesis is equivalent to the hypothesis that $\rho_1^2 = \cdots = \rho_r^2 = 0$ and one can choose $\hat{\rho}_1^2 + \cdots + \hat{\rho}_r^2$ as a test statistic. This test is equivalent to the χ^2 test for independence since $\chi_0^2 = n(\hat{\rho}_1^2 + \cdots + \hat{\rho}_r^2)$. Now, let H_t denote the hypothesis that $\rho_t^2 = 0$. This hypothesis is equivalent to the hypothesis that the rank of ζ is t. O'Neill (1978a, 1978b) showed that the joint asymptotic distribution of $n\hat{\rho}_1^2, \ldots, n\hat{\rho}_r^2$, when H is true, is the same as the joint distribution of the eigenvalues of the central Wishart matrix W with s degrees of freedom and $E(W) = sI_r$. Tables for percentage points of the largest eigenvalue of the central Wishart matrix are given in Krishnaiah (1980b).

We now review the work of Bhaskara Rao, Krishnaiah, and Subramanyam (1985) for determination of the rank of ζ. They suggested functions of $\hat{\rho}_1^2, \ldots, \hat{\rho}_r^2$ as test statistics for testing H_1. For example, one may use $\hat{\rho}_1^2, \hat{\rho}_1^2 + \cdots + \hat{\rho}_r^2$ as test statistics. These authors also suggested the following simultaneous test procedure. We accept or reject H_t, according to

$$\hat{\rho}_t^2 \gtrless c_\alpha \tag{114}$$

where

$$P[\hat{\rho}_t^2 \leq c_\alpha | H_t] = (1 - \alpha) \tag{115}$$

If H_t is accepted and H_{t-1} is rejected, then the rank of ζ is t. Bhaskara Rao, Krishnaiah, and Subramanyam (1985) derived asymptotic joint distribution of functions of $\hat{\rho}_1^2, \ldots, \hat{\rho}_r^2$ when $\rho_1^2, \ldots, \rho_r^2$ have multiplicities. O'Neill (1978a) suggested using $n(\hat{\rho}_t^2 + \cdots + \hat{\rho}_r^2)$ as a test statistic for testing the hypothesis that the rank of ζ is t. In general, we can use a suitable function of $\hat{\rho}_t^2, \ldots, \hat{\rho}_r^2$ such as the foregoing test statistic or $n\hat{\rho}_t^2$ to test the hypothesis that the rank of ζ is t. But, unfortunately, the distributions of these test statistics involve nuisance parameters even asymptotically. As an ad hoc procedure, one can replace the nuisance parameters with their consistent estimates.

Bai *et al.* (1992) proposed the following procedure for determination of the rank of $P = (p_{ij})$. Let

$$G(k) = n \sum_{j=k+1}^{r} \hat{\rho}_j^2 + kC_n \tag{116}$$

where C_n satisfies the following conditions:

$$\lim_{n\to\infty} \left(\frac{C_n}{n}\right) = 0 \tag{117a}$$

$$\lim_{n\to\infty} \left(\frac{C_n}{\log\log n}\right) = \infty \tag{117b}$$

Then the unknown rank q of P is estimated with $\hat{q}$ where $\hat{q}$ is given by

$$G(\hat{q}) = \min\{G(1), \ldots, G(r)\} \tag{118}$$

Bai, Krishnaiah, and Zhao (1992) showed that $\hat{q}$ is a consistent estimate of q.

XIII. NONEXACT TEST FOR THE TWO-SAMPLE PROBLEM

In this and the next sections, we consider the problem of reduction of dimensionality of parameters. As an example, we first consider the two-sample location problem. Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ are random samples from two p-variate normal populations with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and a common covariance matrix Σ. Our problem is to test the hypothesis H: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ against K: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. The classical approach uses the Hotelling test (or T^2-test), with

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{x}} - \bar{\mathbf{y}})' A^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}})$$

$$\bar{\mathbf{x}} = \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1} \mathbf{x}_i, \bar{\mathbf{y}} = \tfrac{1}{n_2}\sum_{i=1}^{n_2} \mathbf{y}_i, \quad \text{and}$$

$$A = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1} (\mathbf{x}_i - \bar{\mathbf{x}})(x_i - \bar{\mathbf{x}})' + \sum_{i=1}^{n_2} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'\right)$$

The T^2 test has lots of good properties, but it is not well defined when the degrees of freedom $(n_1 + n_2 - 2)$ are less than the dimension (p) of the data.

As a remedy, Dempster (1958, 1960) proposed the so-called *nonexact test* (NET) by using the chi-square approximation technique. Dempster's method is to construct an $(n_1 + n_2) \times (n_1 + n_2)$ orthogonal matrix H with the first column consisting of $1/\sqrt{n_1 + n_2}$ and the second column consisting of $n_1\ \sqrt{n_2/n_1(n_1 + n_2)}$ and $n_2\ -\sqrt{n_1/n_2(n_1 + n_2)}$. Write $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{n_1+n_2}) = (\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}, \mathbf{y}_1, \ldots, \mathbf{y}_{n_2})$ and make a transformation

$$\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_{n_1+n_2}) = \mathbf{ZH}$$

It is easy to verify that

$$\mathbf{w}_1 \sim N_p\left(\frac{n_1\boldsymbol{\mu}_1 + n_2\boldsymbol{\mu}_2}{\sqrt{n_1 + n_2}}, \Sigma\right),$$

$$\mathbf{w}_2 \sim N_p\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \Sigma\right)$$

and

$$\mathbf{w}_j \sim N_p(0, \Sigma) \quad \text{for} \quad j = 3, \ldots, n_1 + n_2$$

and all $\mathbf{w}_j$'s are independent of each other.

Then, Dempster used $F=(n_1+n_2-2)Q_2/(Q_3+\cdots+Q_{n_1+n_2})$ as the test statistic, where $Q_j=\mathbf{w}_j'\mathbf{w}_j$. Then, he used the χ^2 approximation for $j=3,\ldots,n_1+n_2$,

$$Q_j \simeq \sigma^2\chi_r^2 \tag{119}$$

Dempster proposed two complicated methods to estimate the parameter r, by the following equations:

$$t=\left(\frac{1}{\hat{r}_1}+\frac{1+(n_1+n_2-2)^{-1}}{3\hat{r}_1^2}\right)(n_1+n_2-3)$$

and

$$t+w=\left(\frac{1}{\hat{r}_2}+\frac{1+(n_1+n_2-2)^{-1}}{3\hat{r}_2^2}\right)(n_1+n_2-3)$$
$$+\left(\frac{1}{\hat{r}_2}+\frac{3}{2\hat{r}_2^2}\right)\binom{n_1+n_2-2}{2}$$

where $t=(n_1+n_2-2)[\ln\frac{1}{n}(\sum_{j=3}^{n_1+n_2}Q_j)]-\sum_{j=3}^{n_1+n_2}\ln Q_j$ and $w=-\sum_{3\le i<j\le n_1+n_2}\ln\sin^2\theta_{ij}$, with θ_{ij} the angle between random vectors $\mathbf{w}_i$ and $\mathbf{w}_j$.

In fact, the parameters σ^2 and r can be estimated by solving the equations

$$r\sigma^2=\mathrm{tr}(\Sigma) \quad \text{and} \quad (r^2+2r)\sigma^2=(\mathrm{tr}(\Sigma))^2+2\mathrm{tr}(\Sigma^2),$$

and Σ can be estimated by the pooled sample covariance matrix A.

The approximation (119) is also valid for $j=2$ under H. Therefore, under H, the distribution F is approximately $F_{r,r(n_1+n_2-2)}$.

Later, Bai and Saranadasa (1996) revisited Dempster's NET and found that the NET is not only a remedy for the T^2 test when $p>n_1+n_2-2$, but also much more powerful than the T^2 test in many general situations when T^2 is well defined. One difficulty in computing Dempster's test statistic is the construction of a high dimensional orthogonal matrix, and the other is the estimation of the degrees of freedom of the chi-square approximation. Bai and Saranadasa (1996) proposed a new test, the asymptotic normal test (ANT), in which the test statistic is $M=\|\overline{\mathbf{x}}-\overline{\mathbf{y}}\|^2-\tau\mathrm{tr}(A)$, where A is the pooled sample covariance matrix and $\tau=\frac{1}{n_1}+\frac{1}{n_2}$.

If we normalize M by an estimator of its variance, the ANT statistic is then given by

$$Z=\frac{\tau^{-1}M}{\sqrt{\frac{2(n_1+n_2-1)(n_1+n_2-2)}{(n_1+n_2)(n_1+n_2-3)}\left(\mathrm{tr}A^2-(n_1+n_2-2)^{-1}(\mathrm{tr}A)^2\right)}}$$

It has been shown that the ANT is asymptotically equivalent to NET, and simulations show that ANT is slightly more powerful than NET. It is easy to show that both the NET and ANT are asymptotically exact, that is, the type I error for both tests tends to the prechosen level. Simulation results show that NET and ANT gain a great amount of power with a slight loss of the exactness of the type I error. Note that *nonexact* does not mean that the error is always larger than the prechosen value.

It is worth to indicate that Bai and Saranadasa's ANT does not need the normality assumption. It requires only the existence of the 4th moment. For details of this method, the reader is referred to Bai and Saranadasa (1996).

Now, let us analyze why this happens. Under the normality assumption, if Σ is known, then the "most powerful test statistic" should be $(\overline{\mathbf{x}}-\overline{\mathbf{y}})'\Sigma^{-1}(\overline{\mathbf{x}}-\overline{\mathbf{y}})$. Since Σ is actually unknown, the matrix A plays the role of an estimator of Σ. Then there is the problem of how close A^{-1} is to the inverse of the population covariance matrix Σ^{-1}.

Let us rewrite the matrix A^{-1} in the form $\Sigma^{-1/2}S^{-1}\Sigma^{-1/2}$, where $\mathbf{S}$ is the sample covariance of $n=n_1+n_2-2$ iid random p-vectors with mean zero and identity covariance matrix. Obviously, the approximation is good if S^{-1} is close to I and not otherwise. Unfortunately, this is not really the case. By the theory of spectral analysis of large dimensional random matrix (see Bai and Yin, 1993; Yin, 1986; and Yin, Bai, and Krishnaiah, 1988), when $p/n=0.25$, the ratio of the largest eigenvalue of S^{-1} to the smallest can be as large as 9. Even when p/n is as small as 0.01, this ratio can be as large as 1.493. This shows that it is practically impossible to get a "good" estimate of the inverse covariance matrix. In other words, if the ratio $(\sqrt{n}+\sqrt{p})^2/(\sqrt{n}-\sqrt{p})^2$ is not in a tolerable range from 1 (e.g., 9 for $p/n=0.25$, 3.7 for $p/n=0.1$, and 1.493 for $p/n=0.01$), NET and ANT give better tests than T^2.

From the viewpoint of dimensionality of parameters to be estimated, we find that the T^2 test needs to estimate $\frac{1}{2}p^2+\frac{5}{2}p$ parameters (i.e., $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ) and the NET and ANT need only estimate $2p+2$ parameters [i.e., $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\mathrm{tr}(\Sigma)$, and $\mathrm{tr}(\Sigma^2)$].

A similar but simpler case is the one-sample problem. As in Bai and Saranadasa (1996), it can be shown that NET and ANT are better than the T^2 test. This phenomenon happens in many statistical inference problems, such as large contingency tables, MANOVA, discretized density estimation, linear models with large number of parameters, and the error-in-variable models. Once the dimension of the parameter is large, the performance of the classical estimators become poor and corrections may be needed.

XIV. MULTIVARIATE DISCRIMINATION ANALYSIS

Suppose that $\mathbf{x}$ is a sample drawn from one of two populations with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and a common

covariance matrix Σ. Our problem is to classify the present sample $\mathbf{x}$ into one of the two populations. If $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and Σ are known, then the best discriminant function is $d = (\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2))' \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, i.e., assign $\mathbf{x}$ to Population 1 if $d > 0$.

When both the mean vectors and the covariance matrix are unknown, assume that training samples $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ from the two populations are available. Then we can substitute the MLE $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$, and A of the mean vectors and covariance matrix into the discriminant function. Obviously, this is impossible if $n = n_1 + n_2 - 2 < p$. The problem is again whether this criterion has the smallest misclassification probability when p is large. If not, what discrimination criterion is better? Based on the same discussion in the last subsection, one may guess that the criterion $d = (\mathbf{x} - \frac{1}{2}(\mathbf{x} + \mathbf{y}))'(\mathbf{x} - \mathbf{y})$ should be better. Using the limiting theorems of spectral analysis of a large sample covariance matrix, this was theoretically proved in Saranadasa (1993). Simulation results presented in his paper strongly support the theoretical results, even for moderately large n and p.

SEE ALSO THE FOLLOWING ARTICLES

STATISTICS, FOUNDATIONS • STATISTICS, MULTIVARIATE

BIBLIOGRAPHY

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. *In* "Proceedings of the Second International Symposium on Information Theory, Suppl. to Problems of Control and Information Theory," pp. 267–281.

Amemiya, Y., and Fuller, W. (1984). Estimation for the multivariate errors-in-variables model with estimated error covariance matrix. *Ann. Statist.* **12,** 497–509.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* **22,** 327–351; correction (1980). *Ann. Statist*. **8,** 1400.

Anderson, T. W. (1984a). "An Introduction to Multivariate Statistical Analysis," Wiley, New York.

Anderson, T. W. (1984b). Estimating linear statistical relationships. *Ann. Statist*. **12,** 1–45.

Anderson, T. W. (1985). Components of variance in MANOVA. *In* "Multivariate Analysis—VI" (P. R. Krishnaiah, ed.), pp. 1–8, North-Holland, Amsterdam.

Anderson, B. M., Anderson, T. W., and Olkin, I. (1986). Maximum likelihood estimators and likelihood ratio criteria in multivariate components of variance. *Ann. Statist*. **12,** 1–45.

Bai, Z. D. (1985). A note on limiting distribution of the eigenvalues of a class of random matrices. *J. Math. Res. Exposition* **5,** 113–118.

Bai, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices, A review. *Statistica Sinica* **9,** 611–677.

Bai, Z. D., and Saranadasa, H. (1996). Effect of high dimension comparison of significance tests for a high dimensional two sample problem. *Statistica Sinica* **6,** 311–329.

Bai, Z. D., and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of large dimensional covariance matrix. *Ann. Probab.* **21,** 1275–1294.

Bai, Z. D., Krishnaiah, P. R., and Liang, W. Q. (1986). On asymptotic joint distribution of the eigenvalues of a noncentral manova matrix for nonnormal populations. *Sankhya, Ser. B* **48,** 153–162.

Bai, Z. D., Krishnaiah, P. R., Subramanyam, K., and Zhao, L. C. (1989). Inference on the ranks of regression matrix and canonical correlation matrix using model selection methods, Tech. Report # 89–20, Center for Multivariate Analysis, Penn State University.

Bai, Z. D., Krishnaiah, P. R., Sambamoorthi, N., and Zhao, L. C. (1992). Model selection for log-linear models. *Sankhya: Ind. J. Statist.* **54, Ser. (B), Pt. 2,** 200–219.

Bartlett, M. (1948). A note on the statistical estimation of demand and supply relations from time series. *Econometrica* **16,** 323–329.

Bhaskara Rao, M., Krishnaiah, P. R., and Subramanyam, K. (1985). A structure theorem on bivariate positive quadrant dependent distributions and tests for independence in two-way contingency tables, Technical Report No. 85-48, Center for Multivariate Analysis, University of Pittsburgh.

Chow, G. C., and Ray-Choudhuri, D. K. (1967). An alternative proof of Hannan's theorem on canonical conditions and multiple equation systems. *Econometrica* **35,** 139–142.

Corsten, L. C. A., and van Eijnsbergen, A. C. (1972). Multiplicative effects in two-way analysis of variance. *Statistica Neerlandica* **26,** 61–68.

Dempster, A. P. (1958). A high dimensional two sample significance test. *Ann. Math. Statist*. **29,** 995–1010.

Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics* **16,** 41–50.

Fang, C., and Krishnaiah, P. R. (1981). Asymptotic distributions of functions of the eigenvalues of real and complex noncentral Wishart matrices. *In* "Statistics and Related Topics" M. Csorgo *et al.*, eds.), pp. 89–108, North-Holland, Amsterdam.

Fang, C., and Krishnaiah, P. R. (1982). Asymptotic distributions of functions of the eigenvalues of some random matrices for non-normal populations. *J. Multivariate Anal*. **12,** 39–63.

Fang, C., and Krishnaiah, P. R. (1984). Asymptotic distributions of functions of the eigenvalues of the doubly noncentral multivariate F matrix. *In* "Statistics: Applications and New Directions," pp. 254–269, Indian Statistical Institute, Calcutta.

Fisher, R. A. (1939). The sampling distribution of some statistics obtained from nonlinear equations. *Ann. Eugen*. **9,** 283–349.

Fisher, R. A., and Mackenzie, W. A. (1923). Studies in crop variation II. The manurial response of different potato varieties. *J. Agric. Sci.* **13,** 311.

Fujikoshi, Y. (1974). The likelihood ratio tests for the dimensionality of regression coefficients. *J. Multivariate Anal*. **4,** 327–340.

Fujikoshi, Y. (1977). Asymptotic expansions for the distributions of some multivariate tests. *In* "Multivariate Analysis IV" (P. R. Krishnaiah, ed.), pp. 55–71. North-Holland, Amsterdam.

Fujikoshi, Y. (1978). Asymptotic expansions for the distributions of some functions of the latent roots of matrices in three situations. *J. Multivariate Anal*. **8,** 63–72.

Fujikoshi, Y. (1985). Selection of variables in discriminant analysis and canonical correlation analysis. *In* "Multivariate Analysis VI" (P. R. Krishnaiah, ed.), pp. 219–236, North-Holland, New York.

Fujikoshi, Y., Krishnaiah, P. R., and Schmidhammer, J. (1987) Effect of additional variables in principal component analysis, discriminant analysis and canonical correlation analysis. *In* "Advances in Multivariate Statistical Analysis," 45–61, Theory Decis. Lib. Ser. B: Math. Statist. Methods, Reidel, Boston.

Gnanadesikan, R. (1977). "Methods of Statistical Data Analysis of Multivariate Observations," Wiley, New York.

Gollob, H. F. (1968). A statistical model which combines features of factor analysis and analysis of variance techniques. *Psychometrika* **33,** 73–116.

Goodman, N. R. (1963). Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction). *Ann. Math. Stat*. **34,** 152–176.

Hannan, E. J. (1967). Canonical correlation and multiple equation systems in economics. *Econometrica* **12,** 124–138.

Hotelling, H. O. (1935). The most predictable criterion. *J. Educational Psychol.* **26,** 139–142.

Hotelling, H. O. (1936). Relations between two sets of variates. *Biometrika* **28,** 321–377.

Hsu, P. L. (1941a). On the limiting distribution of roots of a determinantal equation. *J. London Math. Soc*. **16,** 183–194.

Hsu, P. L. (1941b). On the limiting distribution of the canonical correlations. *Biometrika* **32,** 38–45.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal*. **5,** 248–264.

Johnson, D. E., and Graybill, F. A. (1972). An analysis of a two-way model with interaction and no replication. *J. Amer. Statist. Assoc*. **67,** 862–868.

Kariya, T., Fujikoshi, Y., and Krishnaiah, P. R. (1984). Tests for independence of two multivariate regression equations with different design matrices. *J. Multivariate Anal*. **15,** 383–407.

Kelker, D. (1970). Distribution theory of spherical distribution and a location scale parameter generalization. *Sankhya A* **43,** 419–430.

Krishnaiah, P. R. (1965). Multiple comparison tests in multi-response experiments. *Sankhya A* **27,** 65–72.

Krishnaiah, P. R. (1969). Simultaneous test procedures under general MANOVA models. *In* "Multivariate Analysis II" (P. R. Krishnaiah, ed.), pp. 121–143, Academic Press, New York.

Krishnaiah, P. R. (1976). Some recent developments on complex multivariate distributions. *J. Multivariate Anal*. **6,** 1–30.

Krishnaiah, P. R. (ed.) (1980a). "Analysis of Variance, Handbook of Statistics," Vol. 1, North-Holland, Amsterdam.

Krishnaiah, P. R. (1980b). Computations of some multi-variate distributions. *In* "Handbook of Statistics," Vol. 1 (P. R. Krishnaiah, ed.), pp. 745–971, North-Holland, Amsterdam.

Krishnaiah, P. R. (1982a). Selection of variables under univariate regression models. *In* "Handbook of Statistics," Vol. 2 (P. R. Krishnaiah, ed.), pp. 805–820, North-Holland, New York.

Krishnaiah, P. R. (1982b). Selection of variables in discriminant analysis. *In* "Handbook of Statistics," Vol. 2 (P. R. Krishnaiah and L. N. Kanal, eds.), pp. 883–892, North-Holland, New York.

Krishnaiah, P. R., and Armitage, T. V. (1970). On a multivariate *F* distribution. *In* "Essays in Probability and Statistics" (R. C. Bose *et al.*, eds.), pp. 439–468, University of North Carolina, Chapel Hill.

Krishnaiah, P. R., and Kanal, L. N. (eds.) (1982). "Classification, Pattern Recognition and Reduction of Dimensionality," *Handbook of Statistics,* Vol. 2, North-Holland, Amsterdam.

Krishnaiah, P. R., and Lee, J. C. (1979). On the asymptotic joint distributions of certain functions of the eigenvalues of some random matrices. *J. Multivariate Anal*. **9,** 248–258.

Krishnaiah, P. R., and Lin, J. (1986). Complex elliptical distributions. *Commun. Statist.* **15,** 3693–3718.

Krishnaiah, P. R., and Schuurmann, F. J. (1974). On the evaluation of some distributions that arise in simultaneous tests for the equality of the latent roots of the covariance matrix. *J. Multivariate Anal*. **4,** 265–282.

Krishnaiah, P. R., and Waikar, V. B. (1971). Simultaneous tests for equality of latent roots against certain alternatives—I. *Ann. Inst. Statist. Math*. **23,** 451–468.

Krishnaiah, P. R., and Waikar, V. B. (1972). Simultaneous tests for equality of latent roots against certain alternatives—II. *Ann. Inst. Statist. Math*. **24,** 81–85.

Krishnaiah, P. R., and Yochmowitz, M. G. (1980). Inference on the structure of interaction in two-way classification model. *In* "Handbook of Statistics," Vol. 1 (P. R. Krishnaiah, ed.), pp. 973–994, North-Holland, Amsterdam.

Krishnaiah and Zhoa (1992). p. 32.

Krishnaiah, P. R., Lin, J., and Wang, L. (1985a). Tests for the dimensionality of the regression matrices when the underlying distributions are elliptically symmetric, Technical Report No. 85-36, Center for Multivariate Analysis, University of Pittsburgh.

Krishnaiah, P. R., Lin, J., and Wang, L. (1985b). Inference on the ranks of the canonical correlation matrices for elliptically symmetric populations, Technical Report No. 85-14, Center for Multivariate Analysis, University of Pittsburgh.

Kshirsagar, A. M. (1972). "Multivariate Analysis," Marcel Dekker, New York.

Lancaster, H. O. (1969). "The Chi-Square Distribution," Wiley, New York.

Mandel, J. (1969). Partitioning the interaction in analysis of variance. *J. Res. Natl. Bur. Standards* **73B,** 309–328.

Muirhead, R. J., and Waternaux, C. (1980). Asymptotic distributions in canonical correlation analysis and other multivariate procedures for nonnormal populations. *Biometrika* **67,** 31–43.

O'Neill, M. E. (1978a). Asymptotic distributions of the canonical correlations from contingency tables. *Austral. J. Statist*. **20,** 75–82.

O'Neill, M. E. (1978b). Distributional expansions for canonical correlations from contingency tables. *J. Roy. Statist. Soc. Ser. B* **40,** 303–312.

O'Neill, M. E. (1981). A note on canonical correlations from contingency tables. *Austral. J. Statist*. **23,** 58–66.

Rao, C. R. (1948). Tests of significance in multivariate analysis. *Biometrika* **35,** 58–79.

Rao, C. R. (1973). "Linear Statistical Inference and Its Applications," Wiley, New York.

Rao, C. R. (1983). Likelihood ratio tests for relationships between two covariance matrices. *In* "Studies in Econometrics, Time Series and Multivariate Statistics" (T. Amemiya, S. Karlin, and L. Goodman, eds.), pp. 529–544, Academic Press, New York.

Rao, C. R. (1985). Tests for dimensionality and interactions of mean vectors under general and reducible covariance structures. *J. Multivariate Anal*. **16,** 173–184.

Rissanen, J. (1973). Modeling by shortest data description. *Automatica* **14,** 465–471.

Roy, J. (1958). Step-down procedure in multivariate analysis. *Ann. Math. Statist.* **29,** 1177–1187.

Saranadasa, H. (1993). Asymptotic expansion of the misclassification probabilities of *D*- and *A*-criteria for discrimination from two high dimensional populations using the theory of large dimensional random matrices *J. Multivariate Anal.* **46,** 154–174.

Schmidhammer, J. L. (1982). On the selection of variables under regression models using Krishnaiah's finite intersection tests. *In* "Handbook of Statistics," Vol. 2 (P. R. Krishnaiah and L. N. Kanal, eds.), pp. 821–833, North-Holland, Amsterdam.

Schott, J. R., and Saw, J. G. (1984). A multivariate one-way classification model with random effects. *J. Multivariate Anal*. **14,** 1–12.

Schuurmann, F. J., Krishnaiah, P. R., and Chattopadhyay, A. K. (1973). On the distributions of the ratios of the extreme roots to the trace of the Wishart matrix. *J. Multivariate Anal*. **3,** 445–453.

Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Statist*. **6,** 461–464.

Siotani, M. (1959). The extreme value of the generalized distances of the individual points in the multi-variate normal sample. *Ann. Inst. Statist. Math*. **10,** 183–203.

Siotani, M. (1960). Notes on multivariate confidence bounds. *Ann. Inst. Statist. Math*. **11,** 167–182.

Siotani, M. (1961). The extreme value of the generalized distances and its applications. *Bull. Int. Statist. Inst.* **38,** 591–599.

Thompson, M. L. (1978a). Selection of variables in multiple regression; Part I, A review and evaluation. *Int. Statist. Rev*. **46,** 1–19.

Thompson, M. L. (1978b). Selection of variables in multiple regression; Part II, Chosen procedures, computations and examples. *Int. Statist Rev.* **46,** 129–146.

Tintner, G. (1945). A note on rank, multicollinearity and multiple regression. *Ann. Math. Statist*. **16,** 304–308.

Tukey, J. W. (1949). One degree of freedom for nonadditivity. *Biometrics* **5,** 232–242.

Wijsman, R. A. (1986). Asymptotic distribution of the increase of the largest canonical correlation when one of the vectors is augmented. *J. Multivariate Anal*. **18,** 169–177.

Williams, E. J. (1952). The interpretation of interactions in factorial experiments. *Biometrika* **39,** 65–81.

Wooding, R. A. (1956). The multivariate distribution of complex normal variables. *Biometrika* **43,** 212–215.

Yin, Y. Q. (1986). LSD' for a class of random matrices. *J. Multivariate Anal.* **20,** 50–68.

Yin, Y. Q., Bai, Z. D., and Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probab. Th. Rel. Fields* **78,** 509–521.

Yochmowitz, M. G., and Cornell, R. G. (1978). Stepwise tests for multiplicative components of interaction. *Technometrics* **20,** 79–84.

Zhao, L. C., Krishnaiah, P. R., and Bai, Z. D. (1986a). On detection of number of signals in presence of white noise. *J. Multivariate Anal*. **19,** 1–25.

Zhao, L. C., Krishnaiah, P. R., and Bai, Z. D. (1986b). On detection of number of signals in presence of colored noise using information theoretic criteria. *J. Multivariate Anal*. **19,** 26–49.



Set Theory

Marion Scheepers
Boise State University

I. The Beginnings: Galilei and Cantor
II. Basic Concepts
III. Six Problems of Classical Set Theory
IV. Modern Set Theory's Verdict
V. Conclusion

GLOSSARY

Axiom of choice The statement that for each set of pairwise disjoint nonempty sets, there exists a set which contains exactly one member from each.
Continuum hypothesis The statement that the set of real numbers has the smallest uncountable cardinality.
Determined game A two-person game which never ends in a draw and for which one of the players has a winning strategy.
Measurable cardinal An initial ordinal κ for which there is a κ-complete two-valued measure on all its subsets.
Model of set theory A mathematical object in which all the Zermelo–Fraenkel axioms, as well as the choice axiom, are true.
Ordinal A measure for the length of a well-ordered list.
Uncountable set There is no one-to-one function from the set to the integers.

SET THEORY is the discipline of mathematics concerned with the abstract concepts of "set" and "element," and their applications in mathematics. Much of mathematics can be simulated by using the simple notions of "set" and "element." Whenever this occurs set theory is a foundation for the simulated area of mathematics.

Traditionally the main sources of inspiration for set theory have been problems having roots in the study of the real line. The body of set theoretic knowledge is voluminous and diverse. Here six carefully selected classical problems will serve as an introduction to recent activity in set theory.

I. THE BEGINNINGS: GALILEI AND CANTOR

Galileo Galilei spent a significant part of the dialog during the first day in his "**Discourses and Mathematical Demonstrations Concerning Two New Sciences Pertaining to Mechanics and Local Motions**" (published in 1638) on infinite sets. In this dialog Galilei points out that there is a one-to-one function from the set of positive integers onto the set of squares of the positive integers. He concludes that these two infinite sets have the same number of elements. Then he contrasts this with: For each positive integer N there are approximately $\sqrt{N}$ positive integers whose squares are below N. Thus the

proportion of squares below N to positive integers below N is $\frac{\sqrt{N}}{N}$. This converges to zero as N increases. Thus, though the two sets have the same number of elements, one set is insignificantly small in proportion to the other. These types of contrasts—a set being "big" in one sense and "small" in another—will be called *Galilean paradoxes*.

In 1873 Cantor took the existence of a one-to-one and onto function between sets as the definition for having the same number of elements. Two sets are said to be *equinumerous* if they have the same number of elements. A set A is said to be *less numerous* than a set B if there is a one-to-one function from A to B, but no function from A to B is onto. Cantor proved: The set of positive integers is equinumerous to the set of rational numbers; the real line is equinumerous to the Cartesian plane; the real line is equinumerous to the set of all subsets of the integers. In 1873 he also proved that the set of positive integers is less numerous than the unit interval. Later he proved that every set is less numerous than the set of all its subsets. Thus, two infinite sets need not have the same number of elements. Moreover, for each infinite set, there is one with a larger number of elements. These results can be taken as the official birth of set theory.

The two most used adjectives in set theory are *countable* and *uncountable*. A set is said to be countable if there is a one-to-one function from it to the integers. Otherwise, a set is said to be uncountable.

II. BASIC CONCEPTS

Cantor's results caused a vigorous discussion among mathematicians and philosophers. Careful justification for his proof techniques were called for. Of the several axiomatizations proposed for set theory the Zermelo-Fraenkel axiomatization (denoted ZF) is currently used most universally. With little exception the axiom of choice (AC) is also used as an axiom. ZFC denotes the resulting axiom system. To date no inconsistencies have been discovered in ZFC; it is believed that there are none.

A. The ZFC Axioms

Like the axioms of group theory declare when a mathematical object is a group, the ZFC axioms declare when a mathematical object is a *model of set theory*. The members of a model of set theory are said to be *sets*. ZFC addresses only the *membership* relation among members of a model of set theory. The symbol $\in$ denotes this membership relation. For sets A and B, if $A \in B$ then A is said to be an element of B. Here is a listing of the ZFC axioms, formulated somewhat informally:

Empty-set axiom: There exists a set which has no elements (It is denoted $\emptyset$).
Power set axiom: For each set A there is a set whose only elements are the subsets of A (It is denoted $\mathcal{P}(A)$, and called the *power set of A*).
Pairing axiom: For all sets A and B there is a set whose only members are A and B (It is denoted $\{A, B\}$).
Union axiom: For each set A there is a set whose only members are sets which are members of members of A (It is denoted $\cup A$).
Comprehension axiom: For each set A, and for each set-theoretic property P, there is a set whose only elements are those elements of A having the specified property (It is denoted $\{x \in A : P(x)\}$).
Extensionality axiom: Sets A and B are equal if, and only if, they have the same elements.
Foundation axiom: Each nonempty set A has a member which itself has no members in common with A.
Infinity axiom: There is an infinite set.
Replacement axiom: For $P(\cdot, \cdot)$ a set theoretic property and A a set: If there is for each element a of A a unique b such that $P(a, b)$ is true, then there is a set B such that: for each a in A there is a b in B for which $P(a, b)$ is true.
Choice axiom: For each set A, all of whose members are nonempty and pairwise have elements in common, there is a set which contains exactly one member from each set in A.

The Comprehension- and Replacement axioms are actually infinite lists of axioms, one axiom per set-theoretic property. As Bertrand Russell pointed out, for a proper treatment of these axioms one must make precise the notion of a "set-theoretic property." Doing this would take us a little too far afield.

One must be careful with the membership relation: A model of set theory is itself a collection of objects, named "sets," and the symbol $\in$ refers to the membership relation among those. Often $\in$ is abused and also used to indicate that a set is a member of a given model of set theory. But the axioms do not permit reference to the model of set theory.

B. Ordinal Numbers

Ordinal numbers (ordinals) were introduced by Cantor. Ordinals count position in a list: first, second, third, and so on. One can formalize this concept so that it applies to infinite lists. Here is one way of using sets in a model of set theory to simulate the notion of an ordinal: An ordinal is a set A with the following properties: (i) Each element

of A is a subset of A and (ii) for B and C elements of A, either $B \in C$, or else $C \in B$, or else $B = C$.

Thus $\in$ defines a linear ordering on an ordinal. Also, for any ordinals A and B, either $A \in B$, or else $B \in A$, or else $A = B$. We call an ordinal A *smaller than* (or less than) an ordinal B if $A \in B$. The Foundation Axiom implies that the linear ordering defined on an ordinal by $\in$ contains no infinite descending sequences—i.e., it is a well-ordering. Ordinals are unique in the following sense: If A and B are ordinals and if f is a function from A onto B such that $x \in y$ if, and only if, $f(x) \in f(y)$, then $A = B$.

The empty set $\emptyset$ is an ordinal and as ordinal is denoted by 0. For each ordinal α the set $\alpha \cup \{\alpha\}$ is an ordinal, and is denoted $\alpha + 1$. The union of a set of ordinals is an ordinal. By these two facts and the ZFC axioms the list of ordinals continues indefinitely. One obtains the following picture:

$$0, 1, 2, \ldots, \omega, \omega + 1, \ldots, \omega + \omega, \ldots, \omega_1,$$
$$\omega_1 + 1, \ldots, \omega_2, \ldots, \ldots, \omega_\omega, \ldots$$

Here we have $1 = \{\emptyset\}$, $2 = \{\emptyset, \{\emptyset\}\}$, and so on. ω denotes the least infinite ordinal and is the set $\{0, 1, 2, \ldots\}$. $\omega + 1$ denotes $\omega \cup \{\omega\}$ and ω_1 denotes the smallest ordinal beyond ω for which there is no function from ω onto it. Next are the ordinals $\omega_1 + 1$, $\omega_1 + 2$, and so on. ω_2 denotes the least ordinal beyond ω_1 for which there is no function from ω_1 onto it. One defines $\omega_3, \omega_4, \ldots, \omega_n, \ldots$ (for finite n) analogously. ω_ω is the least ordinal beyond each ω_n for which there is no function from any ω_n onto it. Then comes $\omega_\omega + 1, \omega_\omega + 2, \ldots, \omega_\omega + \omega_\omega, \ldots$ and so on. $\omega_{\omega+1}$ denotes the least ordinal beyond ω_ω for which there is no function from any of the earlier ones onto it. Indeed, for each ordinal α there is a corresponding later ordinal ω_α.

An ordinal is said to be an *initial ordinal* if there is no function from an earlier ordinal onto it. It is said to be a *successor ordinal* if it is of the form $\alpha + 1$. If an ordinal is not a successor ordinal, it is said tobe a *limit ordinal*. The *cofinality* of a limit ordinal A is defined to be the least ordinal B such that there is a function f from B to A with the property that for each $a \in A$ there is a $b \in B$ with $a \in f(b)$. For example, ω is the cofinality of ω_ω, and also of $\omega_2 + \omega$. If a limit ordinal has cofinality ω, it is said to have *countable cofinality*; otherwise, it has *uncountable cofinality*. The cofinality of a limit ordinal is always an initial ordinal.

C. Simulating Mathematics in Set Theory

Much of mathematics can be simulated in set theory. Here is a simulation of number theory: Recall the simulation of ordinals: $0 = \emptyset$, $1 = \{0\}$, $2 = \{0, 1\}$, ... Define the operation S by: $S(x) = x \cup \{x\}$. The ZFC axioms imply that the Peano axioms hold for the set $\omega = \{0, 1, 2, \ldots\}$, endowed with S as successor operation. Also the integers, rational-, real-, and complex numbers can be simulated in set theory.

First, simulate the notion of an ordered pair by sets as follows: Let sets x and y be given. Then $\{x, y\}$ is a set (by the pairing axiom), and is the same as the set $\{y, x\}$ (by the extensionality axiom). Since $\{x\}$ is a set, also $\{\{x\}, \{x, y\}\}$ is a set. The latter set, denoted (x, y), is the simulation of the ordered pair of x and y. One can show that if x is not equal to y, then (x, y) is not equal to (y, x) and if $(x, y) = (u, v)$, then $x = u$ and $y = v$.

For sets A and B let $A \times B$, the Cartesian product of A and B, be simulated by the set $\{(a, b): a \in A \text{ and } b \in B\}$. Subsets of $A \times B$ are used to simulate binary relations. For example, the "less than" relation between natural numbers is simulated by $\{(m, n): m \in \omega, n \in \omega, \text{ and } m \in n\}$.

To simulate integers form the Cartesian product $\omega \times \omega$. Define $\mathbb{Z}$, the set of integers, to be $\{(m, n): m \in \omega, n \in \omega, \text{ and } m = 0 \text{ or } n = 0\}$. Pairs of the form $(0, n)$ represent $-n$ and pairs of the form $(m, 0)$ represent m. Now addition, subtraction, and "less than" can be simulated.

Simulate the rational numbers from the simulation of the integers. Then simulate the real numbers from the simulation of the rational numbers by using Dedekind cuts. An ordered pair (A, B) where A and B are subsets of $\mathbb{Q}$ (the set of simulations of the rational numbers) is said to be a *Dedekind cut* if: (i) $A \cup B = \mathbb{Q}$, (ii) both A and B are nonempty, and (iii) each rational number in A is less than each rational number in B. The real numbers are simulated by such pairs, and $\mathbb{R}$ (the set of simulations of real numbers) is $\{(A, B): (A, B) \text{ it is a Dedekind cut}\}$. The order relation on the set of real numbers is simulated by declaring for the Dedekind cuts (A_1, B_1) and (A_2, B_2) that $(A_1, B_1) < (A_2, B_2)$ if A_1 is a proper subset of A_2. Then simulate the operations of addition and multiplication. The ZFC axioms imply that this simulation is an ordered field, that the completeness axiom holds, and that there are between any two real numbers a rational number.

For sets A and B a function f from A to B is simulated by the set $\{(x, f(x)): x \in A\}$. Next one simulates convergence, continuity, differentiability, integrability, and the other concepts of calculus. The ZFC axioms imply the usual theorems of calculus such as the Intermediate Value Theorem, the Fundamental Theorem of Calculus, and so on, for these simulations.

D. The Cumulative Hierarchy

Let $\mathbf{V}$ be a model of set theory which contains nothing but the essentials. What does it look like? $\mathbf{V}$ is recursively built up as follows: Put $V_0 = \emptyset$. Let α be an ordinal such that V_α has been defined. Define: $V_{\alpha+1} = \mathcal{P}(V_\alpha)$. Let β be

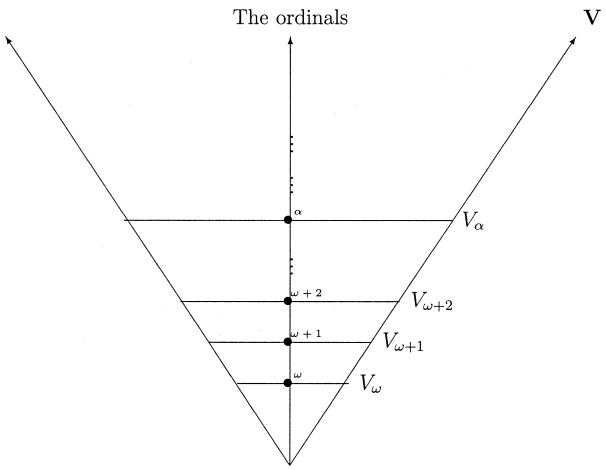


FIGURE 1 A model **V** of set theory.

a limit ordinal such that V_α is defined for each $\alpha \in \beta$. Define: $V_\beta = \cup\{V_\alpha : \alpha \in \beta\}$. Thus we have: $V_1 = \mathcal{P}(V_0)$, $V_2 = \mathcal{P}(V_1), \ldots, V_\omega = \cup\{V_n : n \in \omega\}$, $V_{\omega+1} = \mathcal{P}(V_\omega)$, and so on indefinitely.

For α less than β each element of V_α is an element of V_β. Each α is a member of $V_{\alpha+1}$. For each $\alpha > 0$ the set V_α is a model for all the ZFC axioms except perhaps the power set-, pairing-, infinity- and replacement axioms. For $\alpha > \omega$ the infinity axiom also holds in V_α. For limit ordinals $\alpha > \omega$ the pairing- and power-set axioms also hold in V_α; only the replacement axiom potentially fails. $V_{\omega+\omega}$ is a model for all ZFC axioms except replacement.

V is the collection of sets occuring as members of the V_α's. For each set X in **V** the least α with $X \in V_\alpha$ is said to be the rank of X. We write: $rank(X) = \alpha$. Here are some examples: $rank(\omega) = \omega + 1$, $rank(\mathcal{P}(\omega)) = \omega + 2$, $rank(\mathbb{R}) = \omega + 2$, $rank(\mathcal{P}(\mathbb{R})) = \omega + 3$. Figure 1 depicts **V**.

E. Cardinal Numbers

Cardinals count the number of items in a collection–like one, two, three, and so on. These are simulated as follows by ordinals: Zermelo proved that the axiom of choice is equivalent to: There is for each set an ordinal and a one-to-one function from the set onto that ordinal. For a set X the least ordinal for which there is a one-to-one function from X to that ordinal is said to be the cardinal number, or cardinality, of X and is denoted $|X|$.

The cardinals are the initial ordinals. When emphasizing cardinality characteristics, denote ω by $\aleph_0$, ω_1 by $\aleph_1$, ω_2 by $\aleph_2$, and so on. $\aleph_\omega$ is the least cardinal above each $\aleph_n$; the smallest cardinal above $\aleph_\omega$ is $\aleph_{\omega+1}$.

If a cardinal has an immediate predecessor in this list it is said to be a *successor* cardinal; otherwise, it is a *limit* cardinal. $\aleph_0$ and $\aleph_\omega$ are limit cardinals. $\aleph_1$, $\aleph_2$, and $\aleph_{\omega+1}$ are successor cardinals. The cardinal κ is said to be *singular* if its cofinality is smaller than κ. Else, κ is said to be *regular*. $\aleph_\omega$ and $\aleph_{\omega_1}$ are singular cardinals; $\aleph_1$ is regular, as is every successor cardinal.

Cantor introduced an arithmetic for ordinals and for cardinals. For both there is an addition, a multiplication and exponentiation. Ordinal arithmetic is easier than cardinal arithmetic, but both extend the usual arithmetic of the nonnegative integers.

F. Borel Sets and Projective Sets

Sets encountered in mathematical practice are usually describable in terms of simple set theoretic operations, starting from simply describable sets. Descriptive set theory is devoted to the study of such sets.

To develop a useful theory one must make precise the notions "simple set theoretic operations," and "simply describable sets." In a 1905 paper Borel defined a hierarchy of such "simple" sets as follows: Σ^0_0 is the collection of sets of real numbers which are unions of open intervals with rational boundary points; Π^0_0 is the collection of the complements of sets in Σ^0_0. For α a nonzero countable ordinal such that Π^0_β is defined for all $\beta \in \alpha$, define Σ^0_α as the set of all the unions of countably many sets from $\cup\{\Pi^0_\beta : \beta \in \alpha\}$; Π^0_α is the set of all the complements of sets in Σ^0_α.

For $\beta \in \alpha$ one has: each element of Σ^0_β or Π^0_β is in Σ^0_α and in Π^0_α. Each successively defined set in this list has members not in the previously defined sets. Sets which are elements of $\cup\{\Sigma^0_\alpha : \alpha \in \omega_1\}$ are said to be *Borel* sets.

Similar Borel hierarchies are defined for other spaces commonly used in mathematics—like the complex plane, Euclidean three-space, and the space of continuous real-valued functions on the closed unit interval. Borel sets are now the norm for what is meant by "simply describable" sets.

In mathematics sets are obtained not only as unions and complements, but also as the images of functions. Suslin discovered sets of real numbers which are not Borel, but are the continuous image of a Borel set. This discovery was the official birth of descriptive set theory.

A set of real numbers is said to be *analytic* if it is the continuous image of some Borel set. Σ^1_1 denotes the collection of analytic sets. If a set is the complement of an analytic set it is said to be co-analytic. Π^1_1 denotes the collection of co-analytic sets. For each n, Σ^1_{n+1} denotes the collection of sets which are continuous images of sets in Π^1_n, and Π^1_{n+1} denotes the collection of complements of sets in Σ^1_{n+1}.

For each n the sets in Σ^1_n or in Π^1_n are also in both Σ^1_{n+1} and Π^1_{n+1}. Each set in this list has an element which is not in any prior set in the list. If a set is in some Σ^1_n it is said to be *projective*. Suslin discovered that a projective set is a Borel set if, and only if, it is both analytic and co-analytic.

G. Gale–Stewart Games, and Determined Sets

The following game defined by Gale and Stewart in the 1950s subsumed the games considered in the 1930s by Banach, Mazur, and Ulam: Let a set S of sequences of positive integers be given. Two players, ONE and TWO, play infinitely many innings as follows: In the first inning ONE chooses a positive integer o_1, and TWO responds with t_1. In inning two ONE chooses a positive integer o_2, and TWO responds with t_2, and so on. They play an inning per positive integer. At the conclusion of these innings they have constructed a sequence $s = (o_1, t_1, \ldots, o_n, t_n, \ldots)$ of positive integers. ONE is declared the winner of this *play* s of the game if s is a member of A; otherwise, TWO wins. Let $\mathsf{G}(S)$ denote this game.

For each play of $\mathsf{G}(S)$ there is a winner, but who wins may differ from play to play. Depending on the specific properties of S the game may be more favorable to one of the players. To make the notion of "favorable" precise, we define the notions of a strategy and of a winning strategy. A strategy for player ONE is a function, F, with input the current history of the game and as output the positive integer player ONE is to choose next. Let F be a strategy for ONE. The play s is said to be an F-play if $o_1 = F(\emptyset)$, $o_2 = F(o_1, t_1), \ldots, o_{n+1} = F(o_1, t_1, \ldots, o_n, t_n), \ldots$. F is said to be a winning strategy for ONE if *every* F-play is won by ONE. The notions of a strategy and a winning strategy for TWO are defined similarly.

A set S of sequences of positive integers is said to be *determined* if one of the players has a winning strategy in the game $\mathsf{G}(S)$. If neither player has a winnings strategy then S is said to be *undetermined*. In any model of (ZFC) set theory there are undetermined sets. The choice axiom plays a critical role in this demonstration and the examples obtained from it have pathological properties. Do undetermined sets have to be "pathological"?

Using continued fractions one can show that the set of irrational numbers is essentially the set of sequences of nonnegative integers. Thus, Gale–Stewart games can be viewed as played on the real line. Are projective sets of real numbers determined?

H. The Gödel Incompleteness Theorems

Is it true that all mathematical statements about sets in a model of set theory can be proven or disproven from the ZFC axioms? According to Gödel's First Incompleteness Theorem there is for any list of axioms which is a "reasonable" extension of Peano arithmetic a statement about the objects being axiomatized such that the axioms neither prove, nor disprove the statement. Gödel's examples of such statements are not the type of statement encountered in mathematical practice. The question arose whether any of the notorious unsolved mathematical problems is such an undecidable statement. The first examples of undecidable statements arising from mathematical practice appeared in set theory, as will be seen later, but many have since been discovered in virtually all areas of infinitary mathematics.

Recall that for any limit ordinal $\alpha > \omega$ the set V_α satisfies all the ZFC axioms except perhaps the replacement axiom. If we could prove from ZFC that there is a limit ordinal $\alpha > \omega$ such that V_α also satisfies the replacement axiom, then we would have proved from ZFC that there is a set which is a model of set theory, and thus that ZFC is consistent. But Gödel's Second Incompleteness Theorem states exactly that if ZFC is consistent, then the statement "ZFC is consistent" is not decidable from ZFC. Thus, if ZFC is consistent then it cannot prove that there is a limit ordinal $\alpha > \omega$ such that the replacement axiom holds in V_α.

III. SIX PROBLEMS OF CLASSICAL SET THEORY

A. Cardinal Exponentiation

Cardinal exponentiation led to the first significant set theoretic problem. For κ a cardinal number, 2^κ is defined to be the cardinality of the power set of κ. Thus, $2^{\aleph_0}$ denotes the cardinality of $\mathcal{P}(\omega)$.

1. Cantor's Continuum Problem

$2^{\aleph_0}$ is also the cardinality of the real line. The equation $2^{\aleph_0} = \aleph_1$ is known as Cantor's Continuum Hypothesis (CH). Cantor's Continuum Problem is the question whether CH holds in every model of (ZFC) set theory.

In 1938 Gödel showed that any model of (ZF) set theory contains a "smallest inner model," now called the Constructible Universe, and denoted by **L**. (This is analogous to: Every characteristic zero field has a smallest subfield, "the" field of rational numbers.) He showed that **L** is a model of set theory and CH provably holds in **L**. Thus ZFC cannot disprove CH.

In 1963 Cohen discovered *forcing*, a method to extend a given model of set theory to a larger one. This method is analogous to extending a field to one which contains a root for a given polynomial. Cohen showed by forcing how to extend any model of set theory in which CH holds to one in which CH is false. Thus ZFC cannot prove CH. Conclusion: CH is undecidable from ZFC!

2. Exponentiation of Singular Cardinals

Cohen's breakthrough raised the question: Does ZFC impose any restrictions on the value of 2^κ? It was proved that $2^{\aleph_0}$ can be any cardinal of uncountable cofinality. Easton

showed that for κ any regular cardinal ZFC imposes no restrictions on the value of 2^κ besides (i) the cofinality of 2^κ is larger than κ, and (ii) $\kappa < \lambda$ implies $2^\kappa \leq 2^\lambda$.

For κ a singular cardinal of uncountable cofinality ZFC imposes restrictions. Silver proved: If for each $\lambda < \kappa$ the equation $2^\lambda = \lambda^+$ holds, then $2^\kappa = \kappa^+$ holds. Shortly afterwards Galvin and Hajnal proved: If $\aleph_\alpha$ has uncountable cofinality, and if for each $\kappa < \aleph_\alpha$ also $2^\kappa < \aleph_\alpha$, then $2^{\aleph_\alpha} < \aleph_{(2^{|\alpha|})^+}$.

Let $\mathrm{cof}(\kappa)$ denote the cofinality of κ. For κ singular a careful analysis indicated that the key question is: What is the value of $\kappa^{\mathrm{cof}(\kappa)}$ when $2^{\mathrm{cof}(\kappa)} < \kappa$ holds? Here $\kappa^{\mathrm{cof}(\kappa)}$ is the cardinality of the set of functions from $\mathrm{cof}(\kappa)$ to κ. Classical results imply that $2^{\mathrm{cof}(\kappa)} \neq \kappa$, and that $2^{\mathrm{cof}(\kappa)} > \kappa$ implies $\kappa^{\mathrm{cof}}(\kappa) = 2^{\mathrm{cof}(\kappa)}$. The Singular Cardinal Hypothesis (SCH) is the statement: $2^{\mathrm{cof}(\kappa)} < \kappa$ implies that $\kappa^{\mathrm{cof}(\kappa)}$ is κ^+, the smallest cardinal larger than κ.

Could SCH fail in a model of set theory? Is there in each model of set theory a cardinal κ such that $2^\kappa = \kappa^+$? Do Silver's theorem and the Galvin–Hajnal theorem hold for uncountable cardinals of countable cofinality?

3. The Perfect Subset Property

For Γ a set of subsets of the real line let $\mathrm{CH}(\Gamma)$ denote the Continuum Hypothesis for Γ, which is: Each infinite set of real numbers in Γ has cardinality either $\aleph_0$ or else $2^{\aleph_0}$. Thus, CH is $CH(\mathcal{P}(\mathbb{R}))$.

Cantor proved $CH(\Pi^0_0)$. Later, Young extended this by proving $CH(\Pi^0_1)$. By 1916 Hausdorff and independently Aleksandrov proved $CH(Borel)$. In all these results the proof was based on the notion of a perfect set: A set A of real numbers is said to be *perfect* if: (i) for each point in A there is a sequence in A with all terms different from the point, but it converges to the point, and (ii) for every convergent sequence with all terms from A, the limit of the sequence is also in A. The interval [0,1] is an example of a perfect set. Cantor's famous "middle-thirds" set is a more exotic example. Cantor proved that each perfect set has cardinality $2^{\aleph_0}$.

These results motivate the *Perfect Subset Property*: A family Γ of subsets of $\mathbb{R}$ has the perfect subset property if: Every infinite set in Γ is either countable, or else has a perfect subset. If a family of sets has the perfect subset property, then the Continuum Hypothesis for that family is witnessed by these perfect sets. Bernstein showed that not every set of real numbers of cardinality $2^{\aleph_0}$ has a perfect subset. Thus, having a perfect subset is a stronger property than having cardinality $2^{\aleph_0}$.

Suslin generalized Hausdorff and Aleksandrov's theorem by proving that the family of analytic sets has the perfect subset property. Thus $CH(\Sigma^1_1)$ holds. No proof was forthcoming that Π^1_1 has the perfect subset property. Sierpiński proved that each Π^1_1 set is a union of $\aleph_1$ Borel sets. Thus the cardinality of infinite Π^1_1 sets could be $\aleph_0$, $\aleph_1$ or $2^{\aleph_0}$.

Gödel showed that in $\mathbf{L}$ there is an uncountable Π^1_1 set with no perfect subset. Thus for Π^1_1 the perfect subset property is not provable from ZFC. Is there in every model of set theory an uncountable Π^1_1 set with no perfect subset?

B. Lebesgue Measurability

A set X of real numbers is said to have (Lebesgue) *measure zero* if there is for each positive real ϵ a sequence $(I_n : n < \infty)$ of intervals such that X is covered by their union, and the sum of their lengths is less than ϵ. Measure zero is a notion of smallness but does not mean small cardinality. One has the following Galilean paradox: The Cantor middle-thirds set of real numbers has cardinality equal to that of the real line, but has Lebesgue measure zero.

A set S of real numbers is *Lebesgue measurable* if there is a Borel set B and a measure zero set N such that $S = (B \setminus N) \cup (N \setminus B)$. Thus, a set is Lebesgue measurable if it is only "slightly" different from some Borel set: The set of points where it is different is of Lebesgue measure zero.

There is a function, μ, the Lebesgue measure, defined on the set of all Lebesgue-measurable sets such that:

1. $\mu(X) = 0$ if, and only if, X has Lebesgue measure zero.
2. If $(X_n : n < \infty)$ is a sequence of pairwise disjoint Lebesgue measurable sets, then $\mu(\cup_{n<\infty} X_n) = \sum_{n<\infty} \mu(X_n)$.

A set is Lebesgue measurable if, and only if, it is in the domain of μ. The second property in the list is called "countable additivity."

In each model of (ZFC) set theory there are Lebesgue nonmeasurable sets. The earliest examples were considered pathological. Thus: Can the Lebesgue measure be extended to a countably additive measure ν such that each set of real numbers is measurable with respect to ν? This is *the measure problem*. Are projective sets of real numbers Lebesgue measurable?

1. The Measure Problem

Banach and Kuratowski proved: In any model of set theory in which CH holds there are countably many sets of real numbers such that no countably additive extension of the Lebesgue measure measures all these sets. Thus, ZFC + CH implies a negative answer to the measure problem.

What are the properties of a model of set theory in which the Lebesgue measure can be extended to a countably additive measure ν which measures all sets of reals? Restricting such a ν to subsets of [0,1] gives a real-valued measure on [0,1]. Using a one-to-one function this measure can be transferred to any set of cardinality $2^{\aleph_0}$. Thus, the measure problem may be considered for cardinals. For the least cardinal κ which permits a real-valued measure we have $\kappa \leq 2^{\aleph_0}$ and (by the Banach-Kuratowski result) $\kappa > \aleph_1$.

Ulam proved that if there is any set at all carrying a real-valued measure then the least κ such that there is a real-valued measure on κ is either no larger than $2^{\aleph_0}$, or else carries a two-valued measure. Moreover, if κ is the least cardinal carrying a real-valued measure, then for any family of fewer than κ subsets of κ, each of measure zero, the union of the family has measure zero: This property of the measure is called κ-completeness. Ulam showed that no successor cardinal could be real-valued measurable and that a real-valued measurable cardinal could not be singular. Thus in any model of set theory where the measure problem has a positive solution there would be an uncountable cardinal number $\kappa \leq 2^{\aleph_0}$ which is not a successor cardinal, and not a singular cardinal. Such a cardinal is said to be *weakly inaccessible*.

An uncountable cardinal κ is said to be *measurable* if it carries a two-valued κ-complete measure. Ulam showed that if κ is a measurable cardinal then for each cardinal $\lambda < \kappa$ one has $2^\lambda < \kappa$. Cardinal numbers having the latter property are called *strongly inaccessible*. The concepts of weakly and strongly inaccessible cardinals made their debut in Hausdorff's works, predating Ulam's paper by more than 20 years.

a. Inaccessibility properties of cardinals, and ZFC + I. If a model of set theory contains a weakly inaccessible cardinal κ, then in $\mathbf{L}$ of that model, κ is strongly inaccessible. Moreover, if some model of set theory contains a strongly inaccessible cardinal κ then the *set* V_κ in that model is itself a model of set theory. Thus, if ZFC proved that there is a strongly inaccessible cardinal, then it would prove that there is a set which is a model of set theory. But then ZFC would prove its own consistency, violating Gödel's Second Incompleteness Theorem.

Let I denote the statement "there is a strongly inaccessible cardinal" and let ZFC + I be the extension of ZFC which includes I as an axiom. Since ZFC + I proves the consistency of ZFC it transcends ZFC in axiomatic strength. Ironically, ZFC + I also proves the consistency of ZFC + ¬I: If κ is the least strongly inaccessible cardinal in a model of set theory, then V_κ is a model of set theory which contains no strongly inaccessible cardinals.

b. Measurability properties of cardinals, and ZFC + M. Let M denote the statement "there is a measurable cardinal," and let ZFC + M be the extension of ZFC which includes M as an axiom. By Ulam's results, ZFC + M proves the consistency of ZFC + I. But ZFC + M is much stronger than ZFC + I. In fact, if in a model of set theory κ is a measurable cardinal, then in that model there are κ many strongly inaccessible cardinals below κ. Also, ZFC + M proves the consistency of ZFC + ¬M: If κ is the least measurable cardinal in a model of set theory, then V_κ is a model of set theory which contains no measurable cardinals.

Solovay showed: If there is a model of set theory in which ZFC + M holds, then there is a model of set theory in which the measure problem has a positive solution. The latter model is obtained by extending the former by means of forcing. He also showed: If there is a model of set theory in which the measure problem has a positive solution, then there is a model of set theory in which ZFC + M holds. The latter is obtained from the former by considering an appropriate inner model. In technical language, "a positive solution of the measure problem is equiconsistent with the existence of measurable cardinals."

Let $\mathbf{V}$ be a model of set theory and let $\mathbf{L}$ be the constructible universe of $\mathbf{V}$. Let κ be an uncountable measurable cardinal in $\mathbf{V}$, and let ν in $\mathbf{V}$ be a two-valued κ-complete measure. Scott showed by 1961 that ν cannot be in $\mathbf{L}$. In 1964 Rowbottom gave another dramatic example of how different $\mathbf{V}$ and $\mathbf{L}$ are: the real numbers which are in $\mathbf{L}$ is a *countable* set of $\mathbf{V}$ (most real numbers of $\mathbf{V}$ are not in $\mathbf{L}$). Kunen showed that $\mathbf{V}$ contains a canonical inner model which satsifies ZFC, contains ν, and in which ν witnesses that κ is a measurable cardinal. It is customary that $\mathbf{L}^\nu$ denote this canonical inner model. (This is analogous to the situation where for a polynomial with rational coefficients the field of complex numbers contains a canonical subfield which contains roots of that polynomial—the corresponding Galois extension of the rationals.)

It is surprising that properties of cardinal numbers much larger than $2^{\aleph_0}$ should have any influence on the real numbers. Which properties of the real line are affected by "large" cardinals? Could large cardinals in conjunction with ZFC for example decide CH?

2. Lebesgue Measurability of Projective Sets of Reals

Lusin showed that Σ^1_1 sets are Lebesgue measurable, and thus so are Π^1_1 sets. But the Σ^1_2- and Π^1_2- sets posed problems. Gödel showed that in $\mathbf{L}$ there is a set of real numbers which is both Σ^1_2 and Π^1_2, and yet not Lebesgue measurable. Is there in every model of set theory a set of real

numbers which is both Σ^1_2 and Π^1_2, and yet not Lebesgue measurable?

C. The Baire Property

Another notion of smallness emerged with the study of continuity: Nowhere denseness and *first category*. A set N of reals is *nowhere dense* if there is for each nonempty open interval U a nonempty open subinterval V such that N is disjoint from V. A set is a first category set if it is a union of countably many nowhere dense sets.

Though first category is a notion of smallness it does not imply small cardinality. One has the following Galilean paradox: The Cantor middle-thirds set has cardinality equal to that of the real line, but it is nowhere dense. Also, first category does not imply measure zero, and *vice versa*. In fact one has the following Galilean paradox: The real line is a union of two sets A and B where A has Lebesgue measure zero and B has first category.

A set S of real numbers is said to have the *Baire property* if there area Borel set B and a first category set F such that $S = (B \setminus F) \cup (F \setminus B)$. Thus, sets having the Baire property differ only "slightly" from Borel sets in that they differ from a Borel set by a first category set of points.

There are sets of real numbers which do not have the Baire property. The first examples were considered pathological. Moreover, there are many analogies between the notions of Lebesgue measure zero and first category. Do projective sets have the Baire property? Are theorems mentioning Lebesgue measurable/measure zero or Baire property/first category in their hypotheses or conclusions true when Lebesgue measurable and Baire property are interchanged, and Lebesgue measure zero and first category are interchanged?

1. Projective Sets and the Baire Property

Each Borel set has the Baire property. Lusin and Sierpiński proved that every Σ^1_1 set has the Baire property, and thus every Π^1_1 set has the Baire property. Gödel showed that in $\mathbf{L}$ there is a set of real numbers which is both Σ^1_2 and Π^1_2, but does not have the Baire property.

2. Sierpiński's Duality Program

Sierpiński observed that often a theorem remains true when all occurrences of "Lebesgue measure zero" are replaced with "first category" and vice versa. This observation is summarized by saying that "measure and category are dual notions." Erdös and Sierpiński proved that in every model of set theory in which CH holds the following is true: *Let σ be a statement involving only the notions of Lebesgue measure zero, first category, and set theoretic properties, and let σ^* be the statement obtained by interchanging all occurrences of "measure zero" and "first category" in σ. Then σ implies σ^*.* The statement in italics is known as the Erdös–Sierpiński duality principle.

Does the duality principle hold in all models of set theory? Rothberger introduced the following combinatorial approach to the problem: Let $\mathcal{N}$ denote the set of all Lebesgue measure zero sets of real numbers, and let $\mathcal{M}$ denote the set of all first category sets of reals. Define the cardinal numbers:

$\mathsf{add}(\mathcal{N})$ The minimal cardinality of a family of measure zero sets whose union is not measure zero;
$\mathsf{cov}(\mathcal{N})$ The minimal cardinality of a family of measurezero sets whose union is the real line;
$\mathsf{unif}(\mathcal{N})$ The minimal cardinality of a set of real numbers which is not of measure zero;
$\mathsf{cof}(\mathcal{N})$ The minimal cardinality of a family of measure zero sets such that each measure zero set is a subset of some member of this family.

The cardinals $\mathsf{add}(\mathcal{M})$, $\mathsf{cov}(\mathcal{M})$, $\mathsf{unif}(\mathcal{M})$, and $\mathsf{cof}(\mathcal{M})$ are defined similarly.

Rothberger proved that in any model of set theory both $\mathsf{cov}(\mathcal{M}) \leq \mathsf{unif}(\mathcal{N})$ and $\mathsf{cov}(\mathcal{N}) \leq \mathsf{unif}(\mathcal{M})$ hold. This is an example of duality: Either statement is obtainable from the other by switching $\mathcal{M}$ and $\mathcal{N}$.

D. Choice and Uniformization

Any nonempty subset S of $A \times A$ can be viewed as a collection of pairwise disjoint sets as follows: For $x \in A$ define $S_x := \{(x, y) : (x, y) \in S\}$. Then the family $\{S_x : x \in A \text{ and } S_x \neq \emptyset\}$ is a collection of pairwise disjoint sets. By the axiom of choice there is a subset B of S which meets each nonempty S_x in exactly one point. It is traditional to say that B *uniformizes* S or that B is a uniformization of S.

If a collection of pairwise disjoint nonempty sets is in some sense simply definable, can a simply definable witness of the choice axiom for this collection be found? In particular, do projective subsets of $\mathbb{R} \times \mathbb{R}$ have projective uniformizations? By 1939 Kondô proved that Π^1_1 subsets of $\mathbb{R} \times \mathbb{R}$ have Π^1_1 uniformizations. This implies that Σ^1_2-sets have Σ^1_2 uniformizations. However, it was not clear if Π^1_2-sets have (any) projective uniformizations.

E. Well-Ordering the Real Line

By 1883 Cantor promoted the idea that every well-defined set can be well-ordered. Zermelo proved from ZF that this statement is equivalent to the axiom of choice. Gödel showed that the axiom of choice holds in $\mathbf{L}$ and thus cannot be disproved by ZF. Cohen showed that any model of set

theory can be extended to a model of ZF in which the axiom of choice is false. Thus, the axiom of choice also is not provable from the other axioms.

Do simply definable sets have simply definable well-orderings? One way to approach this is as follows: A well-ordering $\prec$ of a set A can be simulated as a subset of $A \times A$:

$$\prec = \{(a, b) \in A \times A : a \prec b\}.$$

Thus when a well-order of a set of reals is viewed as a subset of the plane it makes sense to ask if it is a projective set. It has long been known that only countable sets could have Σ^1_1 well-orderings.

The unit interval is a very simple Borel set. One can use Fubini's theorem and elementary properties of the Lebesgue measure to show that if $\prec$ is any well-ordering of the unit interval, then as a subset of the unit square it is nonmeasurable. This shows that a projective well-ordering of the real line is incompatible with all projective sets being Lebesgue measurable. Similarly one can use the Kuratowski-Ulam theorem (a category analog of Fubini's Theorem) to show that $\prec$ does not have the Baire property.

Gödel showed that in $\mathbf{L}$ the real line has a Σ^1_2 well-ordering. It follows that there is a Σ^1_2 subset of the plane which is not Lebesgue measurable. D.A. Martin showed that if a set of reals has a Σ^1_2 well-ordering, then it has cardinality at most $\aleph_1$. R.B. Mansfield proved the following very satisfying converse to Gödel's result: If in a model of set theory the real line has a Σ^1_2 well-ordering, then each real number of that model is present in $\mathbf{L}$.

F. Determined Sets of Reals

A result of Banach from the 1930s showed that if a set S is determined then it has the Baire property. In the early 1960s Mycielski and Swierczkowski showed that if the set S is determined then it is Lebesgue measurable and Davis showed that the collection of determined sets have the perfect subset property. By the early 1970s it was also known that determined subsets of $\mathbb{R} \times \mathbb{R}$ have certain uniformization properties. Thus, determined sets have all the regularity properties discussed earlier. For a set Γ of subsets of the real line, let $Det(\Gamma)$ denote that each element of Γ is determined.

As was mentioned earlier, all determined sets of reals have the basic regularity properties: They have the Baire property, the Perfect Subset property, they are Lebesgue measurable, and they have uniformization properties. All the proofs follow the same pattern: Associate an appropriate game with a property, code it as a Gale–Stewart game, then show that if either player has a winning strategy, the original set has the appropriate property.

This indicated a way to unite all the classical results about Borel sets and analytic sets: Show that these sets are determined. Work on this started in the 1950s. At first progress was slow. In 1975 D.A. Martin proved that every Borel set is determined. This made the theory of Borel sets part of the theory of determined sets.

Moreover, these properties of determined sets indicated a strategy for showing that the classical problems about projective sets are not decided by ZFC: Find a model of set theory where the appropriate collection of projective sets is determined. Then in this model of set theory projective sets have regularity properties which projective sets fail to have in $\mathbf{L}$.

IV. MODERN SET THEORY'S VERDICT

The roots of modern set theory lie in the techniques of constructing inner models (as initiated by Gödel) and extensions of models of set theory (as initiated by Cohen) and of associating certain "large" cardinals with regularity properties of sets of real numbers (as initiated by Ulam) and "small" uncountable cardinals with combinatorial properties of families of sets of reals (as initiated by Rothberger).

A. Descriptive Set Theory, Large Cardinals and Determinacy

Solovay showed that if there is a model of set theory in which ZFC + I holds, then there is a model of set theory in which ZF + "every set of reals is Lebesgue measurable" holds. In the latter model of set theory, the full choice axiom fails but a weak version of it, sufficient to conduct classical analysis survives. He also showed that if there is a model of set theory in which ZFC + I holds, then there is a model of set theory in which every projective set of reals is Lebesgue measurable. Shelah later showed that conversely, if there is a model of set theory in which ZF + "every set of reals is Lebesgue measurable" holds, then there is a model of set theory in which ZFC + I holds. This connected the problem of Lebesgue measurability of all projective sets of real numbers with the notion of a strongly inaccessible cardinal. (We already mentioned Solovay's equiconsistency results regarding existence of measurable cardinals and extendibility of Lebesgue measure.)

These results are not direct implications of the form: If a model of set theory has a measurable cardinal, then such and such is true for sets of real numbers of that model. Instead, one finds that if some model has a targeted property then some potentially different model has some other property.

Some of the beautiful discoveries of modern set theory are to the effect that the presence of certain large cardinals in a model of set theory directly influences the properties of certain sets of real numbers in that same model of set theory. Here is an example: Suppose a model of set theory contains a measurable cardinal. By Rowbottom's theorem,

the set of real numbers of **L** of that model is countable. By Mansfield's theorem, if the set of real numbers of a model has a Σ^1_2 well-ordering, then all the reals of that model are in **L** for that model. Thus: If a model of set theory has a measurable cardinal, then the set of reals of that model does not have a Σ^1_2 well-ordering.

Here is another example: Martin showed that if a model of set theory has a measurable cardinal, then every Π^1_1 set of reals of that model is determined. This implies that every Π^1_1-set has the Perfect Subset property, and more since Kechris and Martin proved that if each Π^1_n-set of reals is determined, then each Σ^1_{n+1} set of reals is Lebesgue measurable, has the Baire property, and has the perfect subset property. Thus: If a model of set theory has a measurable cardinal, then every Σ^1_2 set of real numbers is Lebesgue measurable, has the Baire property, and Σ^1_2 has the perfect subset property.

New techniques developed by Foreman, Magidor and Shelah pointed the way to new breakthroughs regarding projective sets of real numbers. Woodin isolated the essential large cardinal property which underlies proofs of determinacy of certain projective sets. The cardinals having the (very technical) property isolated by Woodin are now called Woodin cardinals. Shelah and Woodin showed that for each n: In a model of set theory where there are n Woodin cardinals as well as a measurable cardinal above them all, each Σ^1_{n+2} set of reals is Lebesgue measurable. Martin and Steele then proved the following strengthening: For each n: In a model of set theory which has n Woodin cardinals as well as a measurable cardinal above them all each Π^1_{n+1} set of reals is determined. Moreover, they showed that it is necessary to have the measurable cardinal above the n Woodin cardinals. For they showed that if there is a model of set theory with n Woodin cardinals in it, then there is amodel of set theory with n Woodin cardinals in which the real line has a Σ^1_{n+2} well-ordering. This well-ordering is not Lebesgue measurable, and so by the Kechris-Martin theorem there is an undetermined Π^1_{n+1} set. Woodin finally proved the beautiful theorem that there is a model of set theory in which there are infinitely many Woodin cardinals if, and only if, there is a model of (ZF) set theory in which *every* set of real numbers is determined.

In view of these direct influences of large cardinals on properties of sets of real numbers one may suspect that large cardinals may directly imply something about the cardinality of the set of reals. Thus far no large cardinal ever considered had this effect. For measurable cardinals the canonical inner model for a measurable cardinal satisfies CH. Levy and Solovay also showed that given a model of set theory in which there is a measurable cardinal, it can be extended to a model in which that cardinal is still measurable, but CH fails. Thus, ZFC + M does not decide the cardinality of the real line.

B. Combinatorial Set Theory, Singular Cardinals and pcf Theory

The question whether the Singular Cardinal Hypothesis holds in every model of set theory is also intimately connected with measurable cardinals. It was namely proved that there is a model of set theory in which the Singular Cardinal Hypothesis fails if, and only if, there is a model of set theory which contains a measurable cardinal of a specific kind. The method of proof was typical: For one implication find an appropriate inner model, for the other find an appropriate extension. Moreover, Foreman and Woodin showed that if there is a model of set theory in which there are appropriate large cardinals, then there is a model of set theory where for each infinite cardinal κ, $2^\kappa > \kappa^+$.

Could Silver's theorem fail at some singular cardinal of countable cofinality? By forcing Magidor obtained from a model of set theory containing certain large cardinals a model of set theory in which for each n $2^{\aleph_n} = \aleph_{n+1}$, but $2^{\aleph_\omega} = \aleph_{\omega+2}$. Extending Magidor's work, Shelah obtained from a model of set theory containing certain large cardinals for each infinite $\alpha < \omega_1$ a model of set theory in which for each n $2^{\aleph_n} = \aleph_{n+1}$, but $2^{\aleph_\omega} = \aleph_{\alpha+1}$.

Shelah showed that for singular κ of countable cofinality ZFC imposes restrictions similar to those for singular cardinals of uncountable cofinality (as in the Galvin-Hajnal theorem) on 2^κ. For example: If for each n $2^{\aleph_n} < \aleph_\omega$ then $2^{\aleph_\omega} < \aleph_{(2^{\aleph_0})^+}$. Recently Shelah proved one of the most surprising results in cardinal arithmetic: $2^{\aleph_0} < \aleph_\omega$ implies that $\aleph_\omega^{\aleph_0} < \aleph_{\omega_4}$. It is not clear whether ω_4 is optimal. These ZFC results are a small illustration of the power of pcf theory, a new branch of combinatorial set theory.

It is unknown if $2^{\aleph_\omega}$ could exceed $\aleph_{\omega_1}$ if for each n $2^{\aleph_n} < \aleph_\omega$.

C. Combinatorial Set Theory and the Duality Program

The Sierpiński–Rothberger duality program also saw incredible development since the early 1980s. To give a flavor of this large area of research in set theory, we introduce two more combinatorially defined cardinal numbers. First, let ${}^\omega\omega$ denote the set of functions from ω to ω. For f and g such functions, $f \prec g$ denotes that $\lim_{n\to\infty}(g(n) - f(n)) = \infty$.

The symbol $\mathfrak{b}$ denotes the minimal cardinality attainable by families $\mathcal{F}$ of elements of ${}^\omega\omega$ which have the following property: There is no g in ${}^\omega\omega$ such that for each $f \in \mathcal{F}$, $f \prec g$ holds. The symbol $\mathfrak{d}$ denotes the minimal cardinality attainable by families $\mathcal{G}$ of elements of ${}^\omega\omega$ which have the following property: For each f in ${}^\omega\omega$, there is a $g \in \mathcal{G}$ such that $f \prec g$.

Rothberger proved that $\mathfrak{b}$ and $\mathfrak{d}$ are related to Lebesgue measure and first category: $\mathfrak{b} \leq \mathsf{unif}(\mathcal{M})$ and

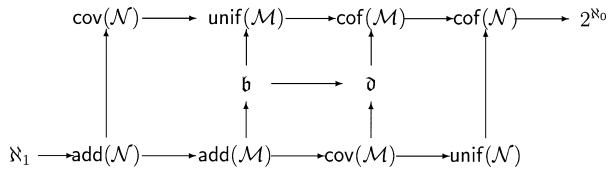


FIGURE 2 Cichoń's diagram.

$\mathsf{cov}(\mathcal{M}) \leq \mathfrak{d}$. It follows from their definitions that also $\mathfrak{b} \leq \mathfrak{d}$.

Early in the 1980s Miller proved that $\mathsf{add}(\mathcal{M}) = \min\{\mathfrak{b}, \mathsf{cov}(\mathcal{M})\}$ and that $\mathsf{cof}(\mathcal{M}) = \max\{\mathfrak{d}\}, \mathsf{unif}(\mathcal{M})\}$. Soon after Bartoszyński an independently Raisonnier and Stern proved that in every model of set theory $\mathsf{add}(\mathcal{N}) \leq \mathsf{add}(\mathcal{M})$ and $\mathsf{cof}(\mathcal{M}) \leq \mathsf{cof}(\mathcal{N})$. The surveyed results are summarized if Fig. 2. In this diagram an arrow indicates that in any model of set theory the cardinal at the initial point of the arrow is no larger than the cardinal at its terminal point.

These results indicate how to proceed to find models of set theory in which Sierpiński's duality principle fails. By the Bartoszyński–Raisonnier-Stern theorem, $\mathsf{add}(\mathcal{N}) \leq \mathsf{cov}(\mathcal{M})$ is true in any model of set theory. Thus, if the Sierpiński duality principle were true in all models of set theory, then also $\mathsf{add}(\mathcal{M}) \leq \mathsf{cov}(\mathcal{N})$ should be true in all models of set theory. Similarly, since $\mathsf{unif}(\mathcal{M}) \leq \mathsf{cof}(\mathcal{N})$ is true in all models of set theory, also $\mathsf{unif}(\mathcal{N}) \leq \mathsf{cof}(\mathcal{M})$ should be true in all models of set theory. Cohen's forcing method was used to show that ZFC does not decide any other relationships than those depicted in Fig. 2. For example, ZFC does not decide either of the inequalities $\mathsf{add}(\mathcal{M}) \leq \mathsf{cov}(\mathcal{N})$ or $\mathsf{cov}(\mathcal{N}) \leq \mathsf{add}(\mathcal{M})$. These were the first examples of "asymmetry" between measure and category given by the combinatorial agenda.

Does ZFC determine anything about the cofinalities of the cardinals in Fig. 2? For all but $\mathsf{cov}(\mathcal{N})$ it was known that these cardinals must have countable cofinality. Only late in the 1990's Shelah finally showed by forcing how to obtain from a model of set theory one in which $\mathsf{cov}(\mathcal{N}) = \aleph_\omega$.

During the 1980s and 1990s it was discovered that some of the cardinal numbers in Fig. 2 are intimately related with open covering properties of topological spaces. In the 1920s the dimension theorists Menger and Hurewicz each introduced a covering property that emerged from their research. Hurewicz showed that $\mathfrak{b}$ is the minimal cardinality of a set of reals not having his covering property, and that $\mathfrak{d}$ is the minimal cardinality of a set of real numbers not having Menger's covering property. In 1938 Rothberger introduced covering properties motivated by his study of a 1919 conjecture by Borel. In the 1980's Fremlin and Miller showed that $\mathsf{cov}(\mathcal{M})$ is the minimal cardinality of a set of real numbers which does not have the Rothberger property. In 1981 Gerlitz and Nagy introduced a covering property in connection with their study of pointwise convergence properties of continuous functions. In the 1990s Nowik, Weiss, and Scheepers showed that a set of real numbers has the Gerlitz–Nagy covering property if, and only if, it has both Hurewicz's and Rothberger's covering properties. It follows that $\mathsf{add}(\mathcal{M})$ is the minimal cardinality for a set of real numbers not having the Gerlitz–Nagy covering property.

V. CONCLUSION

It is curious that the central concept which gave rise to set theory, the notion of a one-to-one function from one set onto another, laid dormant for over two centuries after Galilei's published observations. Carl Boyer says in his well-known history book that "Galileo, like Moses, came within site of the promised land, but he could not enter it."

Though this survey attempted to be comprehensive, time- and space-constraints necessitated neglecting mention of very important work of many able set theorists. Also, many applications of set theory to other areas of mathematics had to be overlooked. Some of this material can be gleaned from the textbooks listed in the bibliography.

SEE ALSO THE FOLLOWING ARTICLES

CALCULUS • CONVEX SETS • GROUP THEORY • RELATIONS AND FUNCTIONS

BIBLIOGRAPHY

Bartoszyński, T., and Judah, H. (1995). *Set Theory*: *On the structure of the real line*, A. K. Peters, Wellesley.

Dales, H. G., and Woodin, W. H. (1987). *An introduction to independence for analysts*, London Mathematical Society Lecture Note Series 115, Cambridge University Press, Cambridge.

Erdös, P., Hajnal, A., Máté, A., and Rado, R. (1984). *Combinatorial Set Theory: Partition relations for cardinals*, Studies in Logic and the Foundations of Mathematics 106, North-Holland Publishing Company, New York.

Farah, I., and Todorčević, S. (1995). *Some applications of the method of forcing*, Yenisei, Moscow.

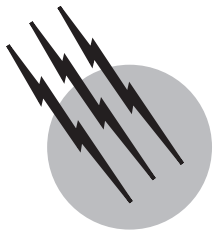
Jech, T. (1995). *Singular cardinals and the pcf theory*, The Bulletin of Symbolic Logic **1**(4), 408–424.

Jensen, R. (1995). *Inner models and large cardinals*, The Bulletin of Symbolic Logic **1**(4), 393–407.

Kanamori, A. (1994). *The Higher Infinite*, Springer-Verlag, Berlin.

Kechris, A. S. (1995). *Classical Descriptive Set Theory*, Graduate Texts in Mathematics 156, Springer-Verlag, New-York.

Woodin, W. H. (1999). *The Axiom of Determinacy, Forcing Axioms and the Nonstationary Ideal*, de Gruyter Series in Logic and its Applications Vol. 1, de Gruyter, Berlin-New York.



Solitons

A. S. Fokas
University of Cambridge

I. Introduction
II. Important Developments in Soliton Theory
III. Coherent Structures
IV. The Inverse Spectral Method
V. A Unification

GLOSSARY

Dromion A particular solution of a nonlinear evoluation equation in two space variables which is localized in space and which moves in certain predetermined tracks (Greek "dromos").

Integrable evolution equations A distinctive class of nonlinear evolution equations which can be analyzed exactly using the inverse scattering method.

Inverse scattering transform method A method which can be used to solve the initial value problem on the infinite line for a nonlinear integrable evolution equation. It is a nonlinear generalization of the Fourier transform method.

Riemann–Hilbert problem A mathematical problem in the theory of complex-valued analytic functions.

Soliton A particular solution of a nonlinear equation which is localized in space and which retains its shape upon interaction with any other localized disturbance.

THE SOLITON was discovered in 1965 by Zabusky and Kruskal, who also introduced this name in order to emphasize the analogy with particles ("soli" for solitary and "tons" for particles). Solitons appear in a surprisingly large number of physical circumstances, including fluid mechanics, nonlinear optics, plasma physics, quantum field theory, relativity, elasticity, biological models, and nonlinear networks (Crighton, 1995). This is a consequence of the fact that a soliton is the realization of a certain physical coherence which is natural to a variety of nonlinear phenomena. These phenomena are modeled by ordinary differential equations (ODEs), partial differential equations (PDEs), singular integrodifferential equations, difference equations, cellular automata, etc. In this review we will concentrate on the occurrence of solitons in nonlinear evolution equations in one and two space dimensions.

I. INTRODUCTION

A soliton was first observed in 1834 by a British experimentalist, J. Scott Russell. He first observed a soliton while riding on horseback beside a narrow barge channel. When the boat he was observing stopped, Russell (1844, p. 311) noted that it set forth "a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change

of form or diminution of speed ... Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon." Russell, impressed by this phenomenon challenged the theoreticians of the day to explain this discovery: "It now remained to the mathematicians to predict the discovery after it happened, that is to give an a priori demonstration a posterori." This work created a controversy which, in fact, lasted almost 50 years. It was resolved by Korteweg and deVries (1895), who derived the KdV equation as an approximation to water waves,

$$\frac{\partial q}{\partial t} + 6q\frac{\partial q}{\partial x} + \frac{\partial^3 q}{\partial x^3} = 0. \tag{1}$$

This equation is a nonlinear PDE of the evolution type, where t and x are related to time and space respectively, and $q(x, t)$ is related to the height of the wave above the mean water level (see Fig. 1). The controversy was resolved because it was shown by Korteweg and deVries (1895) that Eq. (1) supports a particular solution exhibiting the behavior described by Russell. This solution, which was later called the one-soliton solution, is given by

$$q_1(x - p^2 t) = \frac{p^2/2}{\cosh^2\left[\frac{1}{2}p(x - p^2 t) + c\right]}, \tag{2}$$

where p and c are constants. The location of this soliton at time t, i.e., its maximum position, is given by $p^2 t - c/p$, its velocity is given by p^2, and its amplitude by $p^2/2$. Thus, faster solitons are higher and narrower. It should be noted that q_1 is a *traveling wave* solution, i.e., q_1 depends only on the variable $X = x - p^2 t$, thus in this case the PDE (1) reduces (after integration) to the second-order ODE

$$-p^2 q_1(X) + 3q_1^2(X) + \frac{d^2 q_1}{dX^2}(X) = 0.$$

Under the assumption that q and dq/dX tend to zero as $|X| \to \infty$, this ODE yields the solution given by Eq. (2). The problem of finding a solution describing the interaction of two one-soliton solutions is much more difficult and was not addressed by Korteweg and deVries.

The KdV equation is a particularly natural equation combining a simple dispersion with a typical quadratic convective nonlinearity. Actually, this equation is *generic*, in a sense to be explained later, and hence appears in a wide variety of applications. It is therefore surprising that the first new application, other than water waves, was as late as 1960 in plasma physics. Shortly thereafter Kruskal and Zabusky (1965), while trying to understand some puzzling experimental results of Fermi, Pasta, and Ulam, derived the KdV equation as a continuous approximation of a certain anharmonic lattice. Studying numerically the

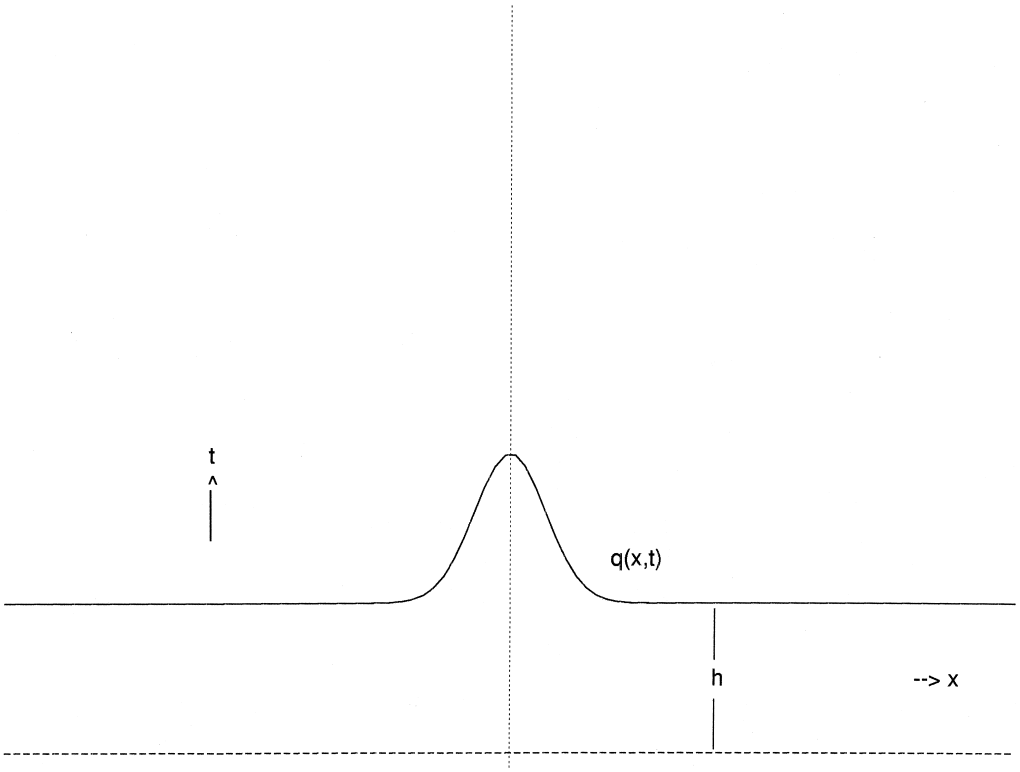


FIGURE 1 The modeling of water waves by the KdV.

interaction of two solutions of the form (2) (i.e., two solutions corresponding to two different p_1 and p_2), they discovered the defining property of solitons: After interaction these waves regained exactly the shapes they had before. This posed a new challenge to mathematicians, namely to explain analytically the interaction properties of such coherent waves. In order to resolve this challenge one needs to develop a larger class of solutions than the one-soliton solution. We note that Eq. (1) is nonlinear and no effective method to solve such nonlinear equations existed at that time.

Gardner *et al*. (1967) not only derived an explicit solution describing the interaction of an arbitrary number of solitons, but also discovered what was to evolve into a new method of mathematical physics. Regarding the explicit solution, we note that the two-soliton solution is given by

$$q_2(x,t) = \frac{2\left(p_1^2 e^{\eta_1} + p_2^2 e^{\eta_2}\right) + 4e^{\eta_1+\eta_2}(p_1-p_2)^2 + 2A_{12}\left(p_2^2 e^{2\eta_1+\eta_2} + p_1^2 e^{\eta_1+2\eta_2}\right)}{\left(1 + e^{\eta_1} + e^{\eta_2} + A_{12}e^{\eta_1+\eta_2}\right)^2}, \tag{3}$$

where

$$\eta_j = p_j x - p_j^3 t + \eta_j^0, \quad j = 1,2; \quad A_{12} = \frac{(p_1-p_2)^2}{(p_1+p_2)^2},$$

with p_j and η_j^0 constants. A snapshot of this solution with $p_1 = 1$, $p_2 = 2$ is given in Fig. 2. After some time the taller soliton will overtake the shorter one and the only effect of the interaction will be a "phase shift," i.e., a change in the position the two solitons would have reached without interaction. These statements can be verified analytically:

a. Comoving with soliton 1. Replace x by the new variable $\xi = x - p_1^2 t$ and let $t \to \pm\infty$; in this frame η_1 is constant and $\eta_2 \to \mp\infty$,

$$t \to +\infty: \quad q \to \frac{p_1^2/2}{\cosh^2\left(\frac{1}{2}\eta_1\right)}.$$

$$t \to -\infty: \quad q \to \frac{p_1^2/2}{\cosh^2\left(\frac{1}{2}\eta_1 + \frac{1}{2}\varphi_{12}\right)}, \quad \varphi_{12} = \ln(A_{12}).$$

b. Comoving with soliton 2. Replace x by the new variable $\xi = x - p_2^2 t$ and let $t \to \pm\infty$; in this frame η_2 is constant and $\eta_1 \to \pm\infty$,

$$t \to -\infty: \qquad q \to \frac{p_2^2/2}{\cosh^2\left(\frac{1}{2}\eta_2\right)}.$$

$$t \to +\infty: \qquad q \to \frac{p_2^2/2}{\cosh^2\left(\frac{1}{2}\eta_2 + \frac{1}{2}\varphi_{12}\right)}.$$

The positions of soliton 1 at $+\infty$ and at $-\infty$ are given by $\frac{1}{2}(p_1 x - p_1^3)t$ and $\frac{1}{2}[p_1(x + \varphi_{12}/p_1) - p_1^3 t]$, respectively; thus there exists a phase shift of φ_{12}/p_1. Similarly for soliton 2.

Regarding the general method introduced by Gardner *et al*. (1967), we note that if Eq. (1) is formulated on the infinite line, then the most interesting problem is the

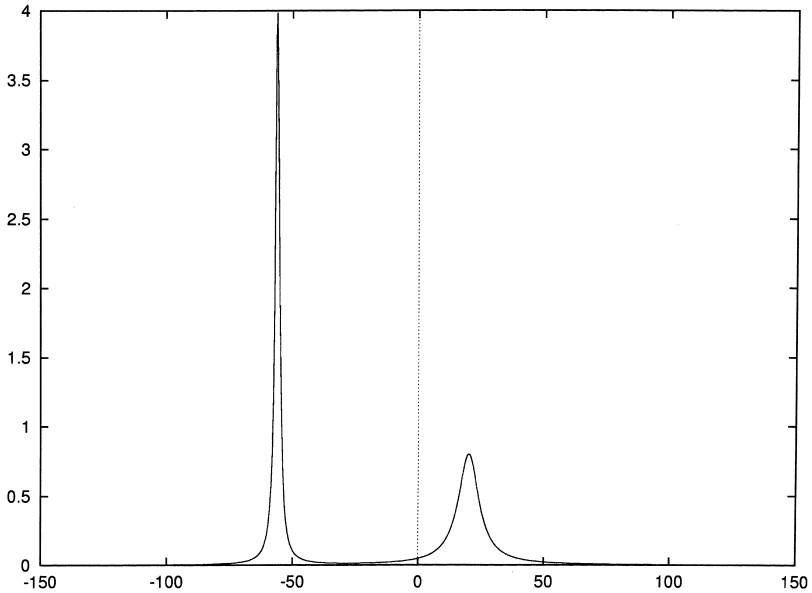


FIGURE 2 A two-soliton solution for the KdV.

solution of the initial-value problem: Given initial data $q(x,0)=q_0(x)$ which decay as $|x|\to\infty$, find $q(x,t)$, i.e., find how these data evolve under the nonlinear PDE (1). If q_0 is small and qq_x can be neglected, then Eq. (1) becomes linear and $q(x,t)$ can be found using the Fourier transform,

$$q(x,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{ikx+ik^3t}\hat{q}_0(k)\,dk, \tag{4a}$$

where

$$\hat{q}_0(k)=\int_{-\infty}^{\infty}e^{-ikx}q_0(x)\,dx. \tag{4b}$$

The remarkable discovery of Gardner *et al.* (1967) is that for Eq. (1) there exists a nonlinear analogue of the Fourier transform capable of solving the initial value problem even if q_0 is not small. Unfortunately, this nonlinear Fourier transform cannot in general be written in closed form, thus $q(x,t)$ in general can be expressed through the solution of a linear integral equation, or more precisely through the solution of a 2×2 matrix Riemann–Hilbert problem as will be explained in Section IV. This linear integral equation is uniquely specified in terms of $q_0(x)$. For particular initial data, $q(x,t)$ can be written explicitly. For example if $q_0(x)=q_1(x)$, where $q_1(x)$ is obtained by evaluating Eq. (2) at $t=0$, then $q(x,t)=q_1(x-p^2t)$. Similarly, if $q_0(x)=q_2(x,0)$, where $q_2(x,0)$ is obtained by evaluating Eq. (3) at $t=0$, then $q(x,t)=q_2(x,t)$.

A most important question, both physically and mathematically, is the description of the long-time behavior of the solution of the above initial value problem. If the nonlinear term of Eq. (1) can be neglected, one finds a linear dispersive equation. In this case different waves travel with different wave speeds, these waves cancel each other out, and the solution decays to zero as $t\to\infty$. Indeed, using the stationary phase method to compute the large-t behavior of the integral appearing in Eq. (4a), we can show that $q(x,t)$ decays like $0(1/\sqrt{t})$ as $t\to\infty$, $x/t=O(1)$. The situation with the KdV equation is more interesting: dispersion is balanced by nonlinearity and $q(x,t)$ has a nontrivial asymptotic behavior as $t\to\infty$. Indeed, using a nonlinear analogue of the steepest descent method discovered by Deift and Zhou (see Fokas and Zakharov, 1993) to analyze the Riemann–Hilbert problem mentioned earlier, it can be shown that $q(x,t)$ asymptotes to $q_N(x,t)$, where $q_N(x,t)$ is the exact N-soliton solution. This underlines the physical and mathematical significance of solitons: *They are the coherent structures emerging from any initial data as* $t\to\infty$. This implies that if a nonlinear phenomenon is modeled by the KdV equation on the infinite line, then one can immediately predict the structure of the solution as $t\to\infty$, $x/t=O(1)$: It will consist of N ordered single solitons, where the highest soliton occurs to the right; the number N and the parameters p_j and η_j^0 depend on the particular initial data $q_0(x)$.

So far we have concentrated on the KdV equation. However, there exist numerous other equations which exhibit similar behavior. Such equations are called *integrable* and the method of solving their initial value problem is called the *inverse scattering* or *inverse spectral* method. Some of these equations together with some typical soliton solutions will be given in Section III. The inverse spectral method will be introduced in Section IV. The extension of this method to boundary value problems will be briefly discussed in Section V. In the next section we present a brief historical review of some of the important developments of soliton theory.

II. IMPORTANT DEVELOPMENTS IN SOLITON THEORY

Were the results of Gardner *et al.* (1967) a fluke, or were there other equations which could be solved in a similar way? Before this question could be answered it was necessary to extract the essence of this ingenious discovery. This was done by Lax (1968), who introduced the so-called Lax pair formulation of the KdV, namely Eq. (1) can be written as the compatibility condition of the following pair of linear eigenvalue equations for the eigenfunction $\psi(x,t,k)$:

$$\psi_{xx}+(q+k^2)\psi=0, \tag{5a}$$

$$\psi_t+(2q-4k^2)\psi_x-q_x\psi=0. \tag{5b}$$

The nonlinear Fourier transform mentioned in Section I can be obtained by performing the spectral analysis of Eq. (5a), while Eq. (5b) is used only for the determination of the time evolution of the nonlinear Fourier data, which are now called spectral data (see Section IV). Following Lax's formulation, Zakharov and Shabat (1972) solved the nonlinear Schrödinger (NLS) equation

$$iq_t+q_{xx}-2\lambda|q|^2q=0,\qquad \lambda=\pm 1 \tag{6}$$

which has ubiquitous physical applications including nonlinear optics. Soon thereafter the sine-Gordon equation

$$q_{xx}-q_{tt}=\sin q \tag{7}$$

and the modified KdV equation

$$q_t+6q^2q_x+q_{xxx}=0 \tag{8}$$

were solved (Ablowitz *et al.*, 1973; Wadati, 1972). Since then numerous nonlinear equations have been solved. Thus the mathematical technique introduced by the authors of Gardner *et al.* (1967) for the solution of a particular physical equation was applicable to a wide range

of other problems. In this way a new method in mathematical physics was born, the so-called inverse scattering (spectral) method. Among the most important equations solved by this method are a particular two-dimensional reduction of Einstein's equation and the self-dual Yang–Mills equations (Ablowitz and Clarkson, 1991).

The next important development in the theory of integrable equations was the study of the KdV with space-periodic initial data. This occurred in the mid 1970s (Belokolos *et al.*, 1994). This method involves algebraic–geometric techniques; in particular, there exists a periodic analogue of the N-soliton solution which can be expressed in terms of a certain Riemann-theta function of genus N.

In the mid 1970s it was also realized that there exist *integrable ODEs*. For example, a stationary reduction of some of the equations introduced in connection with the space-periodic problem mentioned above led to the integration of some classical tops. Furthermore, the similarity reduction of some of the integrable PDEs led to the classical Painlevé equations. For example, letting $q = t^{-1/3}u(\xi)$, $\xi = xt^{-1/3}$ in the modified KdV equation (8) and integrating, we find

$$\frac{d^2u}{d\xi^2} + \frac{1}{3}\xi u + 3u^2 + \alpha = 0, \tag{9}$$

where α is a constant. This is Painlevé II, i.e., the second equation in the list of six classical ODEs introduced by Painlevé and his school around 1900. These equations are nonlinear analogues of the linear special functions such as Airy, Bessel, etc. The connection between integrable PDEs and ODEs of the Painlevé type was established by Ablowitz and Segur (1997). Their work marked a new era in the theory of these equations. Indeed, soon thereafter, Flaschka and Newell (1980) introduced an extension of the inverse spectral method, the so-called isomonodromy method, capable of integrating these equations. This method can also be used to construct nonlinear analogues of the classical connection formulas that exist for the linear special functions. For example, the Airy equation $u_{\xi\xi} = \xi u$ has the following asymptotic behavior:

$$x \to -\infty: \quad u(x) \sim \alpha(-x)^{-1/4} \sin\left[\frac{2}{3}(-x)^{3/2} + \varphi\right], \qquad \alpha < 0, \quad 0 \le \varphi < 2\pi.$$

$$x \to \infty: \quad u(x) \sim \alpha \sin\left(\varphi - \frac{\pi}{4}\right) x^{-1/4} \exp\left(\frac{2}{3}x^{3/2}\right), \qquad \varphi \ne \frac{\pi}{4}, \quad \varphi \ne \frac{5\pi}{4}.$$

Using the isomonodromy method, one can obtain the following nonlinear analogue of these formulas for Painlevé II:

$$x \to -\infty: \quad u(x) \sim \alpha(-x)^{-1/4} \times \sin\left[\frac{2}{3}(-x)^{3/2} + \frac{3}{4}\alpha^2 \ln(-x) + \varphi\right].$$

$$x \to +\infty: \quad u(x) \sim \pm\sqrt{\frac{x}{2}} \pm (2x)^{-1/4}\rho \times \cos\left[\frac{2\sqrt{2}}{3}x^{3/2} - \frac{3}{2}\rho^2 \ln x + \theta\right],$$

where ρ, θ, and s are defined respectively by

$$\rho^2 = \frac{1}{\pi} \ln \frac{1 + |s|^2}{2|\operatorname{Im} s|},$$

$$\theta = -\frac{3\pi}{4} - \frac{7}{2}\rho^2 \ln 2 + \arg \Gamma(i\rho^2) + \arg(1 + s^2),$$

$$s = \left(e^{\pi\alpha^2} - 1\right)^{1/2} \times \exp\left[i\frac{3}{2}\alpha^2 \ln 2 - \frac{i\pi}{4} - i \arg \Gamma\left(\frac{i\alpha^2}{2}\right) - i\varphi\right],$$

and Γ denotes the gamma function.

It was mentioned in Section I that the inverse spectral method gives rise to a matrix Riemann–Hilbert (RH) problem. A RH problem involves the determination of a function analytic in given sectors of the complex plane from the knowledge of the jumps of this function across the boundaries of these sectors. The algebraic–geometric method for solving the space-periodic initial value problem can be interpreted as formulating a RH problem which can be analyzed using functions defined on a Riemann surface. Also the isomonodromy method yields a novel RH problem (Fokas and Zhou, 1992). This implies the following interesting unification: Self-similar, decaying, and periodic initial value problems for integrable evolution equations in one space variable lead to the study of the same mathematical object, namely to the RH problem.

Every integrable nonlinear evolution equation in one spatial dimension has several integrable versions in two spatial dimensions. Two such integrable physical generalizations of the Korteweg–deVries equation are the so-called Kadomtsev–Petviashvili I (KPI) and II (KPII) equations. In the context of water waves they arise in the weakly nonlinear, weakly dispersive, weakly two-dimensional limit, and in the case of KPI when the surface tension is dominant. The nonlinear Schrödinger equation also has two physical integrable versions, known as the Davey–Stewartson I (DSI) and the DSII equations. They can be derived from the classical water wave problem in the shallow-water limit and govern the time evolution of the free surface envelope in the weakly nonlinear, weakly two-dimensional, nearly monochromatic limit. The KP and DS equations have several other physical applications.

A method for solving the Cauchy problem for decaying initial data for integrable evolution equations in two spatial dimensions emerged in the early 1980s. This method is sometimes referred to as the $\bar{\partial}$ (d-bar) method. Recall that the inverse spectral method for solving nonlinear evolution equations on the line is based on a matrix RH problem. This problem expresses the fact that there exist solutions of the associated x part of the Lax pair which are sectionally analytic. Analyticity survives in some multidimensional problems: it was shown formally by Manakov and by Fokas and Ablowitz that KPI gives rise to a *nonlocal RH problem*. However, for other multidimensional problems, such as the KPII, the underlying eigenfunctions are nowhere analytic and the RH problem must be replaced by the $\bar{\partial}$ (d-bar) problem. Actually, a $\bar{\partial}$ problem had already appeared in the work of Beals and Coifman, where the RH problem appearing in the analysis of one-dimensional systems was considered as a special case of a $\bar{\partial}$ problem. Soon thereafter it was shown (Ablowitz *et al*., 1983) that KPII required the essential use of the $\bar{\partial}$ problem. The situation for the DS equations is analogous to that of the KP equations.

The analysis of integrable singular integrodifferential equations and of integrable discrete equations, although conceptually similar to the analysis reviewed above, has certain novel features (Ablowitz and Clarkson, 1991).

The fact that integrable nonlinear equations appear in a wide range of physical applications is *not* an accident but a consequence of the fact that these equations express a certain physical coherence which is natural, at least asymptotically, to a variety of nonlinear phenomena. Indeed, Calogero and Eckhaus (see Fokas and Zakharov, 1993) showed that large classes of nonlinear evolution PDEs, characterized by a dispersive linear part and a largely arbitrary nonlinear part, after rescaling yield asymptotically equations (for the amplitude modulation) having a universal character. These "universal" equations are, therefore, likely to appear in many physical applications. Many integrable equations are precisely these "universal" models.

Related Developments

In the above review we concentrated on analytical aspects of integrable equations. However, it must be emphasized that the study of the rich mathematical structure associated with integrable equations has had a tremendous impact in modern mathematics. Examples include the following:

- The study of the space-periodic problem led to the discovery of new algebraic–geometric results such as the explicit construction of the so-called Baker–Akhiezer function. Furthermore, it has led to the explicit characterization of surfaces of constant mean curvature in connection with the Hopf conjecture.
- The study of space-periodic solutions of the KP equation has led to the solution of the Schottky classical problem of characterizing Riemann matrices in the theory of Riemann surfaces.
- The study of the quantum analogues of the classical integrable models led to beautiful connections between the Yang–Baxter equation and the Bethe ansatz, and also led to the discovery of the quantum groups.
- The study of the Hamiltonian and bi-Hamiltonian formulation of integrable equations led to novel Poisson structures and to beautiful connections with Kac–Moody–Virasoro algebras.

III. COHERENT STRUCTURES

Solitons are important *not* because they are exact solutions, but because they characterize the long-time behavior of integrable evolution equations in one space dimension. There exist two types of localized coherent structures associated with integrable evolution equations in two spatial variables: the *lumps* and the *dromions*. These solutions play a role similar to the role of solitons, namely they also characterize the long-time behavior of integrable evolution equations in two space dimensions (Fokas and Ablowitz, 1983; Fokas and Santini, 1990). The question of solving the initial value problem of a given integrable PDE and then extracting the long-time behavior of the solution is quite complicated and it involves spectral analysis and the formulation of either a RH problem or of a d-bar problem. On the other hand, having established the importance of solitons, lumps, and dromions, it is natural to develop methods for obtaining these particular solutions directly, avoiding the difficult approaches of spectral theory. There exist several such direct methods, including the so-called Bäcklund transformations, the dressing method of Zakharov and Shabat, the direct linearizing method of Fokas and Ablowitz, and the bilinear approach of Hirota.

A. Solitons

Using the bilinear approach, Hietarinta (in press) gave multisoliton solutions for a large class of integrable nonlinear PDEs in one space dimension. Here we only note that the one-soliton solution of the nonlinear Schrödinger equation (6), of the sine-Gordon equation (7), and of the modified KdV equation (8) are given, respectively, by

$$q(x,t) = \frac{p_R e^{i[p_I x + (p_R^2 - p_I^2)t + \eta]}}{\cosh[p_R(x - 2p_I t) + \eta]}, \qquad (10)$$

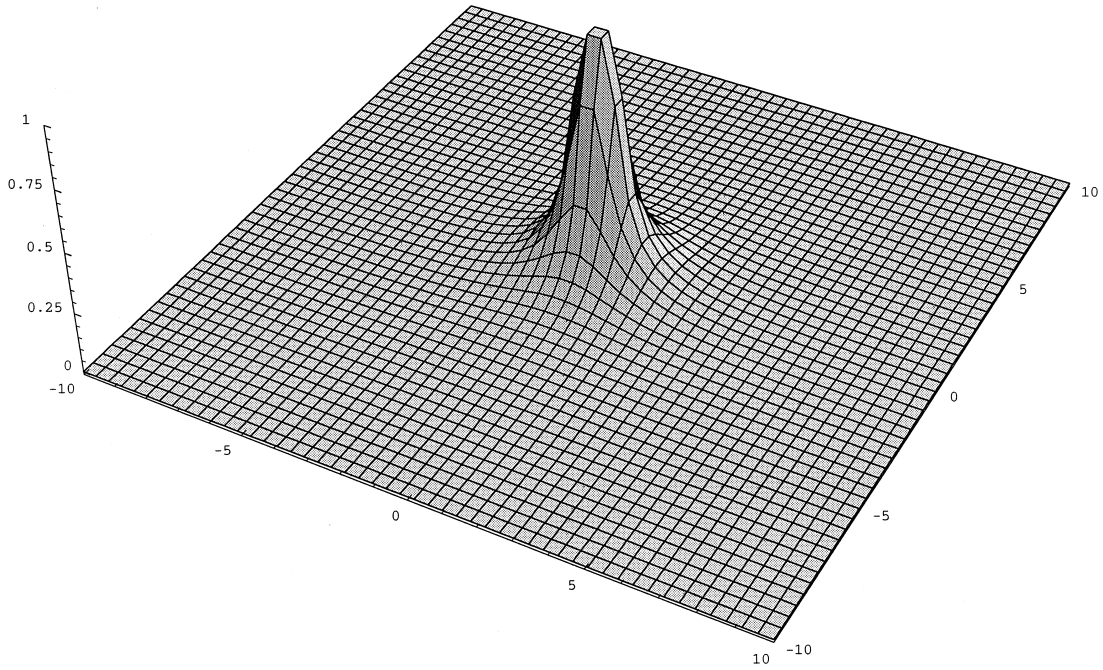


FIGURE 3 A one-lump solution for the DSII equation.

$$q(px+qt) = 4\arctan[e^{px+qt+\eta}], \quad p^2 = 1+q^2, \tag{11}$$

$$q(x-p^2t) = \frac{\pm p}{\cosh[px-p^2t+\eta]}, \tag{12}$$

where p_R, p_I, η, p, and q are real constants.

B. Lumps

The KPI equation is

$$\partial_x[q_t + 6qq_x + q_{xxx}] = 3q_{yy}. \tag{13}$$

The one-lump solution of this equation is given by

$$q(x,y,t) = 2\partial_x^2 \ln\left[|L(x,y,t)|^2 + \frac{1}{4\lambda_I^2}\right],$$
$$L = x - 2\lambda y + 12\lambda^2 t + a, \tag{14}$$
$$\lambda = \lambda_R + i\lambda_I, \qquad \lambda_I > 0,$$

where λ and a are complex constants. Several types of multilump solutions are given by Ablowitz and Villaroel (1997).

The focusing DSII equation is

$$iq_t + q_{zz} + q_{\bar{z}\bar{z}} - 2q\left(\partial_{\bar{z}}^{-1}|q|_z^2 + \partial_z^{-1}|q|_{\bar{z}}^2\right) = 0, \tag{15}$$

where $z = x + iy$, and the operator $\partial_{\bar{z}}^{-1}$ is defined by

$$\left(\partial_{\bar{z}}^{-1} f\right)(z,\bar{z}) = \frac{1}{2i\pi}\int_{\mathbb{R}^2} \frac{f(\zeta,\bar{\zeta})}{\zeta - z}\, d\zeta \wedge d\bar{\zeta}.$$

The one-lump solution of this equation is given by

$$q(z,\bar{z},t) = \frac{\beta e^{i(p^2+\bar{p}^2)t+pz-\bar{p}\bar{z}}}{|z+\alpha+2ipt|^2+|\beta|^2}, \tag{16}$$

where α, β, and p are complex constants. A typical one-lump solution is depicted in Fig. 3.

C. Dromions

The DSI equation is

$$iq_t + \left(\partial_x^2 + \partial_y^2\right)q + qu = 0, \qquad u_{xy} = 2\left(\partial_x^2 + \partial_y^2\right)|q|^2. \tag{17}$$

The one-dromion solution of this equation is given by

$$q(x,y,t) = \frac{\rho e^{X-\bar{Y}}}{\alpha e^{X+\bar{X}} + \beta e^{-Y-\bar{Y}} + \gamma e^{X+\bar{X}-Y-\bar{Y}} + \delta},$$
$$X = px + ip^2t, \qquad Y = qy + iq^2t, \tag{18}$$
$$|\rho|^2 = 4p_Rq_R(\alpha\beta - \gamma\delta),$$

where p and q are complex constants and α, β, γ, and δ are positive constants. Several types of multidromion solutions are given by Hietarinta (in press).

We conclude this section by noting that there exists another type of generalized soliton in two dimensions, namely the so-called *line-solitons*. These solutions can be constructed from the usual solitons by adding an appropriate y dependence. For example a line-soliton of the KPI equation is given by Eq. (2) where the argument of cosh

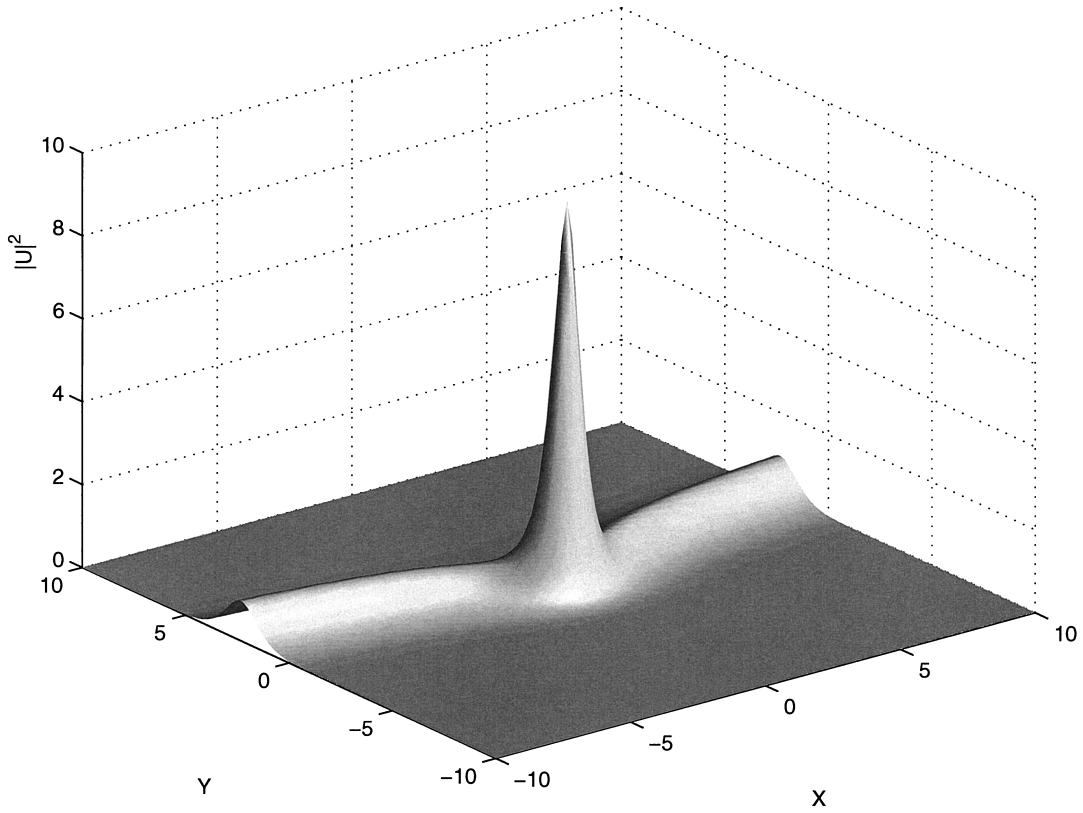


FIGURE 4 The interaction of a one-line-soliton and a one-lump for the KPI equation.

is replaced by $px + iqy + wt$, where $p^4 + pw + 3q^2 = 0$. However, there exist certain lines in the x–y plane where these solutions do *not* decay. A solution describing the interaction of a one-lump and a one-line-soliton for the KPI equation is given by

$$q(x, y, t) = 2\partial_x^2 \ln\left[|L(x, y, t)|^2 + \frac{1}{4\lambda_I^2} + bc + \frac{b}{2\lambda_I} e^{\theta(x,y,t)} + \frac{c}{2\lambda_I} e^{-\theta(x,y,t)}\right], \qquad (19)$$

where L and λ are defined in (14), b and c are nonnegative real constants, and

$$\theta(x, y, t) = 2\lambda_I\left[x - 2\lambda_R y - 4\left(\lambda_I^2 - 3\lambda_R^2\right)t\right].$$

The interaction of a one-lump and a one-line-soliton is depicted in Fig. 4.

IV. THE INVERSE SPECTRAL METHOD

The solution of the initial value problem of an integrable nonlinear evolution equation on the infinite line is based on the spectral analysis of the x part of the Lax pair. Thus for the KdV equation one must analyze Eq. (5a). This equation is the famous time-independent Schrödinger equation. We now give a physical interpretation of the relevant spectral analysis. Let the KdV equation describe the propagation of a water wave and suppose that this wave is frozen at a given instant of time. By bombarding this water wave with quantum particles, one can reconstruct its shape from knowledge of how these particles scatter. In other words, the scattering data provide an alternative description of the wave at fixed time. The mathematical expression of this description takes the form of a linear integral equation found by Faddeev (the so-called Gel'fand–Levitan–Marchenko equation) or equivalently the form of a 2×2 RH problem uniquely specified by the scattering data. This alternative description of the shape of the wave will be useful if the evolution of the scattering data is simple. This is indeed the case, namely, using Eq. (5b), it can be shown that the scattering data evolve linearly. Thus this highly nontrivial change of variables from the physical to scattering space provides a linearization of the KdV equation.

In what follows we will describe some of the relevant mathematical formulas. For simplicity we consider the NLS equation instead of the KdV equation. The associated Lax pair is given by

$$\psi_x + ik\sigma_3\psi = Q\psi, \tag{20a}$$

$$\psi_t + 2ik^2\sigma_3\psi = \left(-i|\lambda|q|^2\sigma_3 + 2kQ - iQ_x\sigma_3\right)\psi, \tag{20b}$$

where

$$\sigma_3 = \mathrm{diag}(1,-1), \qquad Q = \begin{pmatrix} 0 & q \\ \lambda\bar{q} & 0 \end{pmatrix},$$

and $\psi(x,t,k)$ is a 2×2 matrix-valued function.

If q is small, the NLS reduces to the linear equation

$$iq_t + q_{xx} = 0. \tag{21}$$

It was noted by Gel'fand and the author that linear equations also possess a Lax pair. Indeed, it can be verified that Eq. (21) is the compatibility condition of

$$\mu_x - ik\mu = q, \tag{22a}$$

$$\mu_t + ik^2\mu = iq_x - kq. \tag{22b}$$

Let $q(x,t)$ satisfy Eq. (21) in $-\infty < x < \infty$, $t > 0$, where $q_0(x)$ has sufficient smoothness and decay as $|x| \to \infty$. For example, $q(x,0) = q_0(x) \in L_1 \cap L_2$. The unique solution of this problem, which decays as $|x| \to \infty$ uniformly in t, is given by

$$q(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ikx - ik^2t}\hat{q}_0(k)\,dk, \tag{23a}$$

$$\hat{q}_0(k) = \int_{-\infty}^{\infty} e^{-ikx}q_0(x)\,dx. \tag{23b}$$

We will rederive this solution using the Lax pair formulation (22). This seems a complicated way of deriving a simple result. However, it has an important pedagogical advantage: The conceptual steps used here are similar to the conceptual steps used for the solution of the NLS equation.

We first assume that $q(x,t)$ exists and has sufficient smoothness and decay, and we perform the spectral analysis of Eq. (22a). This means that we construct a solution μ, which for fixed $-\infty < x < \infty$, $t > 0$, is bounded in k, $k \in \mathbb{C}$, and which is of $O(1/k)$ as $k \to \infty$. Actually such a solution is sectionally holomorphic and is given by

$$\mu = \begin{cases} \mu^+, & \mathrm{Im}\,k \geq 0, \\ \mu^-, & \mathrm{Im}\,k \leq 0, \end{cases} \tag{24}$$

where μ^+, μ^- are the following particular solutions of Eq. (22a):

$$\mu^+(x,t,k) = \int_{-\infty}^{x} e^{ik(x-x')}q(x',t)\,dx' \tag{25a}$$

$$\mu^-(x,t,k) = -\int_{x}^{\infty} e^{ik(x-x')}q(x',t)\,dx'. \tag{25b}$$

Indeed, since in (25a) $x > x'$, μ^+ is bounded and analytic if $\mathrm{Im}\,k > 0$; similarly for (25b). These equations are both valid if $\mathrm{Im}\,k = 0$; in this case subtracting Eqs. (25), we find

$$\mu^+(x,t,k) - \mu^-(x,t,k) = e^{ikx}\rho(k,t), \qquad k \in \mathbb{R}, \tag{26}$$

where

$$\rho(k,t) = \int_{-\infty}^{\infty} e^{-ikx'}q(x',t)\,dx', \qquad k \in \mathbb{R}. \tag{27}$$

Equations (25) imply that $\mu = O(1/k)$. This estimate and Eq. (26) define an elementary RH problem for the scalar function $\mu(x,t,k)$. The unique solution of this problem is

$$\mu(x,t,k) = \frac{1}{2i\pi}\int_{-\infty}^{\infty} \frac{e^{ilx}\rho(l,t)}{l-k}\,dl, \qquad k \in \mathbb{C}.$$

This equation and Eq. (22a) yield

$$q(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ikx}\rho(k,t)\,dk. \tag{28}$$

In summary, the spectral analysis of Eq. (22a) yields Eqs. (27) and (28), which are the direct and the inverse Fourier transforms, respectively.

In order to find the time evolution of the Fourier data, we use the t part of the Lax pair: Equation (25a) implies that

$$\rho(k,t) = \lim_{x\to\infty}[e^{-ikx}\mu^+(x,t,k)].$$

This equation, the assumption that $q \to 0$ as $x \to \infty$, and Eq. (22b), yield

$$\rho_t + ik^2\rho = 0.$$

Solving this equation in terms of $\rho(k,0)$, using Eq. (27) at $t = 0$, and denoting $\rho(k,0)$ by $\hat{q}_0(k)$, it follows that Eqs. (27) and (28) become Eqs. (23).

These equations can be used to construct the solution of the initial-value problem of Eq. (21) *without* the a priori assumption of existence: Given $q_0(x)$, *define* $\hat{q}_0(k)$ by Eq. (23b). Given $\hat{q}_0(k)$, *define* $q(x,t)$ by Eq. (23a). Since the dependence of $q(x,t)$ on x, t is of the form $\exp(ikx - ik^2t)$, it immediately follows that q satisfies Eq. (21). All that remains is to show that $q(x,0) = q_0(x)$, as a consequence of the inverse Fourier transform.

We now state the analogous result for the NLS equation. This result can be obtained by using Eqs. (20) instead of Eqs. (22) and following steps similar to the ones used above (Fokas, 2000).

Theorem. (Zakharov and Shabat, 1972). Let $q(x,t)$ satisfy the NLS Eq. (6) in $-\infty < x < \infty$, $t > 0$, with $q(x,0) = q_0(x) \in S(\mathbb{R})$. The defocusing NLS equation, i.e., Eq. (6) with $\lambda = 1$, has a unique global solution which

decays to zero as $|x| \to \infty$ uniformly in t. This is also the case for the focusing NLS equation, i.e., Eq. (6) with $\lambda = -1$, provided that the function $a(k)$, $\operatorname{Im} k > 0$, defined below, has at most a finite number of zeros all of which are simple and nonreal. In both cases this solution can be obtained as follows: Given $q_0(x)$, define the vector $v(x,k) = (v_1, v_2)^\tau$ as the unique solution of

$$v_{1_x} + 2ikv_1 = q_0(x)v_2, \qquad v_{2_x} = \lambda\bar{q}_0(x)v_1,$$
$$\operatorname{Im} k \geq 0, \qquad -\infty < x < \infty, \qquad \lim_{x\to\infty} v = (0,1)^\tau.$$

Given $v(x,k)$, define the functions $a(k)$ and $b(k)$ by

$$a(k) = \lim_{x\to-\infty} v_2(x,k), \qquad \operatorname{Im} k \geq 0;$$
$$b(k) = \lim_{x\to-\infty} \left(e^{2ikx} v_1(x,k)\right), \qquad k \in \mathbb{R}.$$

Given $a(k)$ and $b(k)$, define μ as the solution of the following 2×2-matrix RH problem:

1.

$$\mu^-(x,t,k) = \mu^+(x,t,k) \begin{pmatrix} 1 & -\dfrac{b(k)}{\bar{a}(k)} e^{-\theta} \\ \lambda\dfrac{\bar{b}(k)}{a(k)} e^{\theta} & 1 - \lambda\dfrac{|b|^2}{|a|^2} \end{pmatrix}, \quad k \in \mathbb{R},$$

where $\mu = \mu^+$ for $\operatorname{Im} k \geq 0$, $\mu = \mu^-$ for $\operatorname{Im} k \leq 0$, and $\theta = 2i(kx + 2k^2t)$.

2.

$$\mu = I + O\left(\frac{1}{k}\right), \qquad k \to \infty.$$

3.

$$\det \mu = 1.$$

4. If $\lambda = 1$, $\mu^\pm$ are holomorphic, and if $\lambda = -1$ the first column vector of μ^+ and the second column vector μ^- are meromorphic with poles at the zeros of $a(k)$ and of $\bar{a}(\bar{k})$, respectively; the residues of these poles are given by

$$\operatorname{res}_{k_j}(\mu^+)_1 = c_j e^{\theta_j}(\mu^+)_2(k_j), \qquad \theta_j = 2i\left(k_j x + 2k_j^2 t\right),$$
$$\operatorname{res}_{\bar{k}_j}(\mu^-)_2 = -\bar{c}_j e^{\bar{\theta}_j}(\mu^-)_1(\bar{k}_j),$$

where the subscripts 1 and 2 denote the first and second column vectors. This RH has a unique solution. Given μ^+, define $q(x,t)$ by

$$q(x,t) = -2i \lim_{k\to\infty} (k\mu^+_{12}),$$

where the subscript 12 denotes the 1, 2 component of the matrix μ^+.

V. A UNIFICATION

Recently a new method for studying boundary value problems for linear and for integrable nonlinear PDEs in two dimensions has been introduced. This method provides a unification as well as a significant extension of the following three seemingly different topics: (1) The classical integral transform method for solving linear PDEs and several of its variations such as the Wiener–Hopf technique. (2) The integral representation of the solution of linear PDEs in terms of the Ehrenpreis fundamental principle. (3) The inverse spectral (scattering) method for solving the initial value problem for nonlinear integrable evolution equations.

In what follows we briefly discuss this unification.

A. Transform Methods for Linear PDEs

Almost as soon as linear two-dimensional PDE's made their appearance, d'Alembert and Euler discovered a general approach for constructing large classes of their solutions. This approach involved separating variables and superimposing solutions of the resulting ODEs. The method of separation of variables naturally led to the solution of PDEs by a transform pair. The prototypical such pair is the direct and the inverse Fourier transforms; variations of this fundamental transform include the Laplace, Mellin, sine, and cosine transforms and their discrete analogues.

The proper transform for a given boundary value problem is specified by the PDE, the domain, and the given boundary conditions. For some simple boundary value problems, there exists an algorithmic procedure for deriving the associated transform. This procedure involves constructing the Green's function of a *single* eigenvalue equation and integrating this Green's function in the complex k plane, where k denotes the eigenvalue.

The transform method has been enormously successful for solving a great variety of initial and boundary value problems. However, for sufficiently complicated problems the classical transform method fails. For example, there does not exist a proper analogue of the sine transform for solving a third-order evolution equation on the half-line. Similarly, there do not exist proper transforms for solving boundary value problems for elliptic equations even of second order and in simple domains. The failure of the transform method led to the development of several ingenious but ad hoc techniques, which include conformal mappings for the Laplace and the biharmonic equations, the Jones method and the formulation of the Wiener–Hopf factorization problem, the use of some integral representation, such as that of Sommerfeld, and the formulation of a difference equation, such as the Malyuzhinet equation.

The use of these techniques has led to the solution of several classical problems in acoustics, diffraction, electromagnetism, fluid mechanics, etc. The Wiener–Hopf technique played a central role in the solution of many of these problems.

B. The Ehrenpreis Fundamental Principle

In 1950, Schwartz posed the problem of whether, given a polynomial P on $\mathbb{C}^n$, an elementary solution of the differential operator $P(i\partial/\partial t)$, $t \in \mathbb{R}^n$, always exists, i.e., if there exists a distribution E solving $P(i\partial/\partial t)E = \delta$, where δ is the Dirac delta function. The existence of such an elementary solution was established independently by Malgrange and Ehrenpreis; both of these proofs are nonconstructive. Thus in the same decade, techniques of functional analysis were used to try to construct this elementary solution explicitly. For example, if one considers the equation

$$P\left(\frac{i\partial}{\partial t}\right)q(t) = 0, \qquad t \in \Omega, \tag{29}$$

where Ω is a convex domain in $\mathbb{R}^n$, it follows that $q(t) = e^{-ik\cdot t}$, $k \in \mathbb{C}^n$, is a solution of (29) if $P(k) = 0$. For $n = 1$, the Euler principle states that every solution of (29) is a linear combination of exponentials. The generalization of Euler's principle for $n > 1$ was established by Ehrenpreis and Palamodov and was called by Ehrenpreis the *fundamental principle*.

An elementary implication of the Ehrenpreis principle is that for the general evolution equation

$$\left(\partial_t + i\sum_0^n \alpha_j(-\partial_x)^j\right)q(x,t) = 0 \tag{30}$$

formulated in $0 < x < \infty$, $0 < t$, there exists a measure such that

$$q(x,t) = \int e^{ikx - i\omega(k)t}\, d\mu(k), \qquad \omega(k) = \sum_{j=0}^{n} \alpha_j k^j. \tag{31}$$

We note that if $n = 2$, Eq. (30) can be solved by the sine or the cosine transform. In this case it is possible to rewrite the solution in the Ehrenpreis form (31), with $d\mu$ supported on the real axis and on the positive imaginary axis.

C. Integrable Nonlinear PDEs

Integrable equations, solitons, the inverse spectral method, and the associated algebraic–geometric machinery, which were briefly mentioned in Sections I–IV, have had an important impact on modern mathematical physics. However, in spite of this success, these methods are limited. Indeed, they are applicable only to the solution of the initial value problem with either decaying or space-periodic initial data. Thus an outstanding open problem in the analysis of integrable equations became the generalization of these methods to boundary value problems. The simplest such problem is formulated on the half-line. It was mentioned earlier that the inverse spectral method on the infinite line can be thought of as a nonlinearization of the Fourier transform. Thus a natural strategy for solving a problem on the half-line is to solve the associated linear problem by an x transform and then to nonlinearize this transform. However, this strategy fails: for the Korteweg–de Vries equation it fails immediately, since the associated linear equation is $q_t + q_{xxx} = 0$, for which there does not exist an appropriate x-transform. For the nonlinear Schrödinger equation, the associated linear equation can be solved by either the sine or the cosine transform, depending on whether $q(0,t)$ or $q_x(0,t)$ is given, but neither of these transforms nonlinearizes. It is the author's opinion that this failure reflects the fact that neither of these transforms is fundamental. The fact that they are limited only to second-order equations provides further support for this claim. Indeed, there exist a new formalism for solving any linear dispersive equation on the half-line, and this formalism *can* be nonlinearized.

D. A New Method

A general approach to solving boundary value problems for two-dimensional linear and integrable nonlinear PDEs is reviewed in Fokas (2000). This method can be applied to *linear PDEs with constant coefficients* and to *integrable nonlinear* PDEs in an arbitrary domain. It involves the following three steps: (1) Given a PDE, construct two compatible linear eigenvalue equations, called a Lax pair. (2) Given a domain, perform the simultaneous spectral analysis of the Lax pair. (3) Given appropriate boundary conditions, analyze the global relation satisfied by the solution and by its derivatives on the boundary of the domain.

We now discuss the relation of this method with the three topics presented earlier.

1. Suppose that $q(x, y)$ satisfies a linear PDE. Performing the spectral analysis of the x part of the Lax pair corresponds to constructing an x-transform; similarly, performing the spectral analysis of the y part corresponds to constructing a y-transform. The advantage of the Lax pair is that it provides the tool for performing the *simultaneous* spectral analysis. This gives rise to a new transform, which in contrast to both the x- and y-transforms, is "custom-made" for the given PDE and the given domain. In this sense the new method provides the synthesis of separation of variables.

2. Suppose that $q(x, y)$ satisfies a linear PDE in a convex polygon. In this case, step 1 yields for $q(x, y)$ an

integral representation in the complex k plane, which has an explicit x and y dependence and which involves a certain function $\hat{q}(k)$, $k \in \mathbb{C}$, which we call the spectral function. This function can be expressed in terms of an integral of q and of its derivatives along the boundary of the polygon. However, some of these boundary values are unknown. Thus in order to compute $\hat{q}(k)$ [and thus $q(x, y)$] one needs to determine the part of $\hat{q}(k)$ involving the unknown boundary values. An important aspect of the new method is the understanding that this can be achieved by analyzing the *global equation* satisfied by the boundary values of q and of its derivatives. For evolution equations and for elliptic equations with simple boundary conditions, this involves the solution of a system of algebraic equations, while for elliptic equations with arbitrary boundary conditions, it involves the solution of a Riemann–Hilbert problem. For simple polygons, this Riemann–Hilbert problem is formulated on the infinite line, thus it is equivalent to a Wiener–Hopf problem. This explains the central role played by the Wiener–Hopf technique in many earlier works.

3. An important advantage of the new method is that it can be nonlinearized. Indeed, the results valid for linear PDEs obtained by this method can be generalized to integrable nonlinear PDEs.

4. For linear equations, the explicit x, y dependence of $q(x, y)$ is consistent with the Ehrenpreis formulation of the solution. Thus this method provides the concrete implementation as well as the generalization to concave domains of this fundamental principle. For nonlinear equations, it provides the extension of the Ehrenpreis principle to integrable nonlinear PDEs.

ACKNOWLEDGMENT

This work was partially supported by the EPSRC.

SEE ALSO THE FOLLOWING ARTICLES

DIFFERENTIAL EQUATIONS, ORDINARY • DIFFERENTIAL EQUATIONS, PARTIAL • NONLINEAR OPTICAL PROCESSES • NONLINEAR PROGRAMMING

BIBLIOGRAPHY

Ablowitz, M. J., and Clarkson, P. A. (1991). "Soliton, Nonlinear Evolution Equations and Inverse Scattering," Cambridge University Press, Cambridge.

Ablowitz, M. J., and Segur, H. (1977). *Phys. Rev. Lett*. **38,** 1103.

Ablowitz, M. J., and Villaroel, J. (1997). *Phys. Rev. Lett*. **78,** 570.

Ablowitz, M. J., Kaup, D. J., Newell, A. C., and Segur, H. (1973). *Phys. Rev. Lett*. **31,** 125.

Ablowitz, M. J., Bar Yaakov, D., and Fokas, A. S. (1983). *Stud. Appl. Math*. **69,** 135.

Belokolos, E. D., Bobenko, A. I., Enolskii, V. Z., Its, A. I., and Matveev, V. B. (1994). "Algebro-geometric Approach to Nonlinear Integrable Equations," Springer-Verlag, New York.

Crighton, D. (1995). "Applications of KdV." *In* "KdV'95" (M. Hazewinkel, H. Capel, and E. de Jager, eds.), pp. 2977–2984, Kluwer, Dordrecht.

Flaschka, H., and Newell, A. C. (1980). *Commun. Math. Phys*. **76,** 67.

Fokas, A. S. (2000). *J. Math. Phys*. **41,** 4188.

Fokas, A. S., and Ablowitz, M. J. (1983). *Stud. Appl. Math*. **69,** 211.

Fokas, A. S., and Santini, P. M. (1990). *Physica D* **44,** 99.

Fokas A. S., and Zakharov, V. (eds.). (1993). "Important Developments in Soliton Theory," Springer-Verlag, New York.

Fokas, A. S., and Zhou, X. (1992). *Commun. Math Phys*. **144,** 601.

Gardner, C. S., Greene, J. M., Kruskal, M. D., and Miura, R. M. (1967). *Phys. Rev. Lett*. **19,** 1095.

Hietarinta, J. (in press). "Scattering of solitons and dromions." *In* "Scattering" (P. Sabatier and E. Pike, eds.), Academic Press, New York.

Korteweg, D. J., and deVries, G. (1895). *Phil. Mag*. **39,** 55.

Lax, P. D. (1968). *Commun. Pure Appl. Math*. **21,** 467.

Russell, J. S. (1844). "Report on Waves," J. Murray, London.

Wadati, M. (1972). *J. Phys. Soc. Japan* **32,** 1681.

Zabusky, N. J., and Kruskal, M. D. (1965). *Phys. Rev. Lett*. **15,** 240.

Zakharov, V., and Shabat, A. (1972). *Sov. Phys. JETP* **34,** 62.



Spatial Objective Analysis

H. J. Thiébaux
National Centers for Environmental Prediction

I. Introduction
II. The Fundamental Problem
III. Optimal Statistical Objective Analysis
IV. Kriging
V. Empirical Interpolation Techniques

GLOSSARY

Autoregressive Dependent on its own history, or determined by it up to random innovations.
Diagnostics Measures whose values describe the state, energy, or distribution of a property of an observed system.
Earth-oriented system One whose natural coordinates are designated with respect to the earth, in latitude, longitude, and depth, altitude, or pressure.
Global description Designation of the state of a system or the value of one of its variables at each point of its domain.
Minimum variance estimation Determination of a function or parameter values for a designated function of stochastic variables, which guarantee its having the least statistical variance among all functions or sets of parameters.
Spatial coherence Continuity of values of variables, such as temperature, salinity, or wind velocity, in a spatially continuous environment.
Unbiased Property of an estimate for a variable or parameter of having the same statistical mean value as the variable or parameter estimated.

SPATIAL OBJECTIVE ANALYSIS is a body of techniques of applied mathematics that has been developed to provide global descriptions of spatially coherent variables with data from observation points that are sparse and inhomogeneous in space. Thus, in the specific, *a spatial objective analysis* is a tensor estimate of the state of a system for each gridpoint of the domain of interest, that has been constructed from a set of irregularly spaced observations. The word *objective* identifies the construction as the outcome of applying a mathematical algorithm, in distinction to a *subjective* construction which admits the influence of individual scientific intuition.

I. INTRODUCTION

Historically, the techniques of spatial objective analysis have been developed as descriptive and diagnostic techniques of science, beginning with K. F. Gauss's pioneering work in astronomy and extended by people who are now distinguished as founders of the discipline of mathematical statistics. Areas of scientific endeavor in which spatial objective analysis is currently used include geology, atmospheric and oceanic sciences, and environmental studies.

The relative isolation of research in these different fields has led to the assignment of credits to a spectrum of scientists, each of whom had the acumen to appreciate the importance of these techniques and applied them within their own fields with outstanding results. Prominent on this roster are G. Matheron, L. S. Gandin, and R. E. Kalman.

Scientific study of large-scale, earth-oriented systems requires capabilities for constructing global descriptions with algorithms built from assumptions about spatial interrelationships of their component variables. These requirements must confront the *de facto* configuration of our habitat, which precludes spatially uniform data coverage. In fact, actual observations of global-scale phenomena, including those of remote sensing technology, may provide information that is highly nonuniform in coverage. Nonetheless, global descriptions of spatially dimensioned states of the system which accurately reflect its properties and interrelationships of its component parts must be constructed.

The techniques developed to meet these requirements may be classified by the characteristics of the information available to describe the system under study. There are two clear divisions. One presumes no history of information, of which an example would be an isolated, geological field study, producing a single "snap shot" of observations at selected locations, for a system whose properties are presumed to be relatively fixed in time. The other assumes an accumulated record of observations on a time-evolutionary system for which either the statistical relationships among states of the system at separate locations remain the same throughout the record or whose changes with time can be determined independently.

II. THE FUNDAMENTAL PROBLEM

A spatial objective analysis algorithm is a formula for providing estimates of the values of a spatially coherent variable where each gridpoint estimate is constructed from a set of "surrounding" observations. If observations recorded true values perfectly, then the objective might be to fit them perfectly with a surface whose intermediate values would provide estimates of the variable everywhere else. However, the premise is incorrect. Also, it is generally true that the signal component of the field we seek to estimate differs from the instantaneous state, or "true value," at each instant of time and point of space.

Since the distinction between the signal component of the variable at a specific location and instant of time and the value of an observation at that same location and instant is basic to the theory from which all spatial objective analysis algorithms are derived, we consider this first. The *signal component* is the value one would like to know and which the estimate is constructed to approximate. It is not the same as the value of an observation, were it possible to obtain one because an observation is a composite of the instanteous state, errors caused by imprecision of sensors and recorders, and mistakes made in the transmission of information. Furthermore, the instantaneous state at a specific location and instant of time generally reflects microscale variability, the influence of which is regarded as obscuring the signal. Accordingly, the signal component has a theoretical definition as the average of the instantaneous states over a localized region and short time interval. The construction of an algorithm for estimating the signal from actual observation reports takes into account the unwanted, noisy constituents, so that the algorithm will provide a filter for separation of signal from noise, as well as an interpolater.

The coherence in space and in time that is assumed by the theory of objective analysis is equivalent to the mathematical concept of the continuity of functions. Thus the goal of objective analysis is the creation of a spatially continuous estimator of the signal in the field variable: an algorithm that may be evaluated at any point, effectively interpolating only the signal components of the observations going into it.

Traditionally, the *signal*, $\mu(\lambda, \phi, p, t)$, and the *noise*, $\varepsilon(\lambda, \phi, p, t)$, components of observations are assumed to be additive:

$$Z(\lambda, \phi, p, t) = \mu(\lambda, \phi, p, t) + \varepsilon(\lambda, \phi, p, t);$$

and the further assumption is made that the noise is independent of the signal, in the sense that the magnitude or strength of the signal does not influence the level of noise and vise versa. The indexing parameters denote earth-oriented coordinates, respectively: longitude, latitude, pressure or altitude or depth, and time. Here $\varepsilon(\lambda, \phi, p, t)$ represents the collective influence of small-scale variability, of no inherent interest, together with instrument and transmission system errors.

A general expression for the observed value of a field variable is that of the output of a *generalized Kalman filter*, or *GKF* (see Thiébaux, 1997, for a practical discussion of this concept). In this, the indices of distinct location/time points are subsumed by the notation s and $s + \Delta$. The formulation of the GKF assumes that the true value of the field variable at any location/time may be written as a composite of its deterministic evolution and a stochastic innovation,

$$X_{s+\Delta} = \Phi_s^* \bigcirc X_s + W_{s+\Delta},$$

and that a recorded observation is a linear combination of the true value and noise from sensing, transmitting, and recording devices:

$$Z_{s+\Delta} = X_{s+\Delta} + V_{s+\Delta}.$$

Φ^* denotes the algorithm for the deterministic component of the change in X from one location/time to another, W denotes the stochastic component in the evolution of the true field, and V denotes the collective impact of the noise elements. GKF output is composed of a prediction from the proximal/previous estimate, modified by the weighted discrepancy between what is observed and predicted:

$$\hat{X}_{s+\Delta} = \Phi_s \bigcirc \hat{X}_s + K_{s+\Delta}(Z_{s+\Delta} - \Phi_s \bigcirc \hat{X}_s)$$

A distinction is made here between the true *system operator* of the field variable, Φ^*, and the *model system operator* used to generate predictions, Φ. Generally, the true operator will not be known precisely and making the assumption that they are the same will lead to erroneous deductions from a formulation that equates them. Customarily K is called the *gain matrix of the filter*, for which the formulation includes the possibility of its dependence on time and space, like that of the system operators. The generality of this representation for a spatial objective analysis is mathematically impressive. However, it is impractical to apply to the analysis of most physical systems in its full generality because the components are not known at this level of detail.

Within the present framework, an *optimal estimator* is one for which the average squared departure of values of the estimates it produces from the corresponding signal values is smallest among all possible algorithms. Since the signal is a conceptual quantity that cannot be observed, the average is the statistical mean-squared departure, and thus optimal objective analysis is *minimum variance estimation* in the classic statistical sense. Whether or not the minimum can be guaranteed depends on whether or not the multivariate statistical distribution of the field variable or of its increment from some known background field can be specified. When the relevant multivariate statistical distribution can be specified, the optimal estimator for any specified location is unique, and it is the conditional mean value of the variable at that location with respect to its joint distribution with the values of the variable at all the locations at which it is observed. We refer to this as the *ultimate objective analysis* and reserve the definition *optimal objective analysis* for what is practical in terms of our knowledge of statistical properties of the field variable. The most general way of representing the ultimate objective analysis value for any point, say a point indexed by 0 constructed with observations from m locations indexed $1, \ldots, m$, is in terms of the cumulative distribution function for Z_0 conditioned on the observed values of its m covariates $Z_1, \ldots, Z_m$. Specifically, in terms of

$$F(z_0 \mid Z_1, \ldots, Z_m) = \text{Prob}(Z_0 \leq z_0, \text{ given the values of } Z_1, \ldots, Z_m),$$

the conditional mean is

$$E(Z_0 \mid Z_1, \ldots, Z_m) = \int_{-\infty}^{+\infty} z_0 \, dF_0(z_0 \mid Z_1, \ldots, Z_m),$$

and this can be shown to provide the unique minimum of

$$E\left\{[Z_0 - g(Z_1, \ldots, Z_m)]^2\right\}$$

for all possible functions g of the set of observations. However, it clearly depends on knowing the family of multivariate distribution functions

$$F(z_0, z_1, \ldots, z_m) = \text{Prob}(Z_0 \leq z_0, Z_1 \leq z_1, \ldots, Z_m \leq z_m),$$

from which all conditional distribution functions and their conditional mean values can be derived. Many practical situations warrant the assumption that these distributions are reasonably approximated by multivariate normal distributions, and in these cases the objective analysis formulation assumes a particularly simple form.

In those circumstances in which the multivariate statistical distributions for the field variable cannot be specified, alternative approaches to objective analysis are employed. Again, the choice of a technique will depend on what assumptions can reasonably be made about boundary conditions on the variable for the region within which it is considered and characteristics of its spatial variability. The nonstatistical alternatives that will be cited in this article are referred to collectively as *empirical interpolation techniques*. There are several, each with advantages tuned to specific observation characteristics and requirements for the estimated field. *All share the disadvantage that no analysis algorithm can outperform the statistical objective analysis in circumstances in which the latter is a practical alternative*.

III. OPTIMAL STATISTICAL OBJECTIVE ANALYSIS

Here the underlying assumption is that the situation warrants application of the family of multivariate normal distributions to the differences between the true/observed values and corresponding modeled mean values. The differences between the location-specific values and the modeled mean values may be called "analysis increments." The "modeled mean values" are called "first-guess values," and these may be derived from a complex algorithm that takes into account everything in our knowledge base about the system under study, or it may be simply an ensemble average of available information or "the climatology of a geophysical field." The model detail and its accuracy in representing the system will be reflected in the spatial scale and structure of the analysis increments, presenting a trade-off between the requirement for complexity

in the first-guess algorithm and the requirement for complexity of the analysis algorithm. *The closer the model corresponds to the operation of the true system, the less structure remains in the increment field; conversely, the simpler the model, the more residual statistical structure remains to be represented.*

In any case in which a multivariate normal distribution may be assumed for the differences between the location-specific values and the modeled mean values, the optimal estimator for an analysis increment for any unobserved location is a weighted linear combination of the observed increments that are within "influence range." Observed increments are derived from observed values by differencing with corresponding modeled means. Thus, denoting observed values by $Z_1, \ldots, Z_m$ and the modeled mean values by $Z_0^f, Z_1^f, \ldots, Z_m^f$, for each point of estimate 0 and the m observing locations, respectively, the objective analysis algorithm generates point estimates:

$$\hat{Z}_0 = Z_0^f + \sum_{i=1}^{m} w_i \left(Z_i^o - Z_i^f\right). \tag{1}$$

The coefficients that provide the weights bear a direct and simple relationship to the covariance parameters of the assumed underlying normal distribution function. With double subscripts identifying the locations for the variables to which each scalar covariance σ_{ij} pertains, the between-location covariance array may be written as

$$\sum = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \sigma_{m1} & & \sigma_{mm} \end{bmatrix}$$

and the array whose entries are the covariances between the values of the variable at the point of estimate and at the locations of observations as

$$\sum_0 = (\sigma_{11} \cdots \sigma_{1m}).$$

With this notation the array of weights for the analysis increments is given by

$$W = \sum_0 \sum^{-1}. \tag{2}$$

As a theoretical formulation, (1) and (2) describe the *ultimate objective analysis*. In practice, the formulation requires specification of the covariance structure, which must itself be estimated, to create a working, optimal statistical objective analysis (OSOA) algorithm. It is at this point that "optimal" takes on its practical connotation and the dependence of the accuracy of the analysis on the representation of the covariance structure assumes central importance. As the word suggests, *covariance* is a measure of the strength of the relationships, or *covariability*, of increments at different locations. Its tensor representation determines the weights assigned to observed increments in OSOA.

The accuracy of the foregoing representation is of considerable significance: *If the assigned covariance values are not representative of the true statistical relationships of the increment field, the weights and therefore the analysis will be less than optimal.* A body of research has been devoted to the derivation and study of the properties of covariance functions for the representation of spatial covariance structure. Some of this has started from first principles, in the following sense. The discrepancy between an observed field and an unbiased modeled mean field is described as a spatially coherent stochastic process, for which the covariance function may be derived analytically from the spatial analogue of an autoregressive representation for a time series. Selection of the order of the autoregressive representation and assignment of values to the parameters of the covariance function require comparative fitting of candidate functions to an array of covariances computcd from data, together with the study of implicit properties of these functions versus known or theoretical properties of the increment field. These derivations have assumed stationarity of the statistical properties of the increment field and required archives of observations from fixed arrays for the determination of spatial structure.

IV. KRIGING

In circumstances in which the increment field is known to be statistically nonstationary or for which an archive of observations is not available to provide guidance for selecting a representation for the between-location covariances of analysis increments, OSOA cannot be used and an alternative technique must be selected. One practical alternative that has been developed and used extensively in a number of natural and social science settings is *Kriging*. The term is derived from the name Krige, the man to whom G. Matheron ascribes its inspirational use in the field of geology. This technique was developed to produce estimates of values of spatial processes over a geographic region with a one-time set of data from a sampling array. Kriging is a multistage procedure that first fits a trend surface to the data, then estimates the covariance structure of the residual field from the differences between the observations and the fitted surface, and, finally, corrects the point predictions of the trend surface by a constrained multivariate interpolation of the observed residuals.

The trend surface is constructed by fitting a linear combination of basis functions $h_l(s)$ to the sample data in the manner of a regression analysis. Thus the coefficients, α_l

are determined by minimizing the sum of squared differences between the surface and the observed values

$$\sum_{i=1}^{m}\left[Z_i^0 - \sum_l \alpha_l h_l(s_i)\right]^2,$$

where $s_1, \ldots, s_m$ are the locations of the observations. This provides a trend surface estimate

$$m(s) = \sum_l \alpha_l h_l(s),$$

which may be evaluated at any point of the region. In Kriging, this is referred to as the *drift* and is somewhat analogous to the role of the modeled mean field of OSOA. One important departure from OSOA is accommodation for bias that may be present in $m(s)$. Since the minimization that parametrizes the fitted surface does not guarantee its unbiasedness, Kriging constrains the weights of the linear interpolation of residuals to sum to unity.

Other significant differences between the Kriging analysis and OSOA derive from assumptions that are made in the Kriging analysis about the joint distribution of the discrepancies between the trend surface and the data to which it has been fitted. This distribution is assumed to be a multivariate normal distribution that is both stationary and isotropic. The mean, which is the bias b in the fitted surface, and the variance σ^2 are assumed to be constant over the region; and the correlation between values of the field variable at separate locations is assumed to be a function only of scalar displacement, $\rho(\|\Delta\|)$, which approaches zero with increasing separation. With this notation, between-location covariances are

$$\sigma_{ij} = \sigma^2 \varrho(\|\Delta\|), \qquad \text{with} \quad \|\Delta\| = \|s_i - s_j\|.$$

With the constraint of having only one observation at each location, the requirement for a representation of the covariances for interpolation of residuals to combine with the trend surface estimate is satisfied by constructing a "variogram." To illustrate, we let $Y(s)$ denote the residuals, which are the discrepancies between the trend surface and the data to which it has been fitted:

$$Y(s_i) = m(s_i) - Z_i^0 \qquad \text{for} \quad i = 1, \ldots, m.$$

With this notation and the foregoing assumptions about the distribution of the residuals, the expected value of the square of the difference between each pair of residuals may be written in terms of the variance and the correlation function as

$$E\left\{[Y(s_i) - Y(s_j)]^2\right\} = E\left\langle\{[Y(s_i) - b] - [Y(s_j) - b]\}^2\right\rangle$$
$$= 2\sigma^2[1 - \varrho(\|\Delta\|)].$$

Thus the expected value of $[Y(s_i) - Y(s_j)]^2/2$ is $\sigma^2[1 - \varrho(\|\Delta\|)]$. Since $\varrho(\|\Delta\|)$ approaches 0 at large separations, a plot of $[Y(s_i) - Y(s_j)]^2/2$ against the distances between observing locations $\|\Delta\| = \|s_i - s_j\|$ may be fitted with a function $\gamma(\|\Delta\|)$ whose asymptote provides an estimate of the variance σ^2 and whose normalized values provide a representation for the correlation function:

$$\hat{\varrho}(\|\Delta\|) = 1 - \gamma(\|\Delta\|)/\sigma^2.$$

V. EMPIRICAL INTERPOLATION TECHNIQUES

Spatial continuity of the variables for which analyzed fields are constructed is basic to spatial objective analysis. However, having sufficient knowledge of the properties of the stochastic elements of the observed field for a statistical objective analysis is not a requirement for the construction of a spatially continuous estimated field for the region from which observations are obtained. Under circumstances in which it is unreasonable to make explicit assumptions concerning either the definition of a mean field or the form of the statistical distribution of observed residuals from a defined mean field, other options are available. Alternatives to statistical objective analysis are referred to here as empirical interpolation techniques. These include (1) distance-weighted interpolation, (2) spline fitting, (3) function surface fitting, and (4) neural network analysis.

The formal appearance of the first item on this list is similar to a statistical objective analysis:

$$\hat{Z}_0 = \sum_{i=1}^{m} a_i Z_i^0.$$

However, the coefficients that assign weights to observed values of the analysis increments are functions only of the spatial/temporal separation between the point/time of the estimate and the locations/times of the observations. They do not take into account the possibility of significant correlations among the component observations. Nonetheless, in circumstances in which observations are fairly uniformly spaced in the region of the analysis and when the covariance structure of the analysis increments is not known, this technique can provide reasonable estimates. Distance-weighted interpolation has a long history of use in atmospheric science, where it is known as Cressman or Barnes analysis; it is the technique used to produce large-area grid-point analyses of surface temperature from satellite reports and has recently been applied in a four-dimensional objective analysis of wind profiler data (Spencer *et al.*, 1999).

Spline fitting has become a scientific art in recent years (Luo and Wahba, 1997; Luo *et al.*, 1998). It is a special case of *function surface fitting* in which several-dimensional polynomials are fitted to data within

subregions. The polynomials have fewer parameters than the number of observations, so that they perform a smoothing role as well as data fitting, and they are matched at the boundaries of the subregions for continuity. Control for excursions of the piecewise continuous surface beyond reasonable limits on the variability of the analyzed field is achieved by putting the spline under tension. Spline surface fitting has considerable flexibility for tuning the surface to match theoretical properties or response to orography. However, it requires considerable insight and attention to mathematical and physical detail, and, like all constructions made by fitting functions to data, the resulting analysis algorithm cannot be used for estimation beyond the region of the observations with confidence in the result.

Function surface fitting has been considered in outline as the first step of Kriging in which a *trend surface* is obtained by choosing the parameters for a linear combination of basis functions to be those parameter values that minimize the sum of the squared departures of the fitted surface from the observations. Basis functions may be selected to satisfy implicit spectral properties known from physical theory or an analog field, or they may be sets of orthogonal functions selected for their suitability as statistical diagnostics of the analyzed field. Recent literature on this subject is extensive in its treatment of both the theory of function surface fitting and high-speed computer techniques for data fitting. For examples, see Borzelli and Ligi (1999) and Emery and Thomson (1997).

Neural network analysis, as a practical technique for objective estimation of spatial fields, may be said to be in its infancy. However, it holds promise as a powerful technique to augment traditional linear statistical methods in data analysis and forecasting. Schematically, neural network models insert "hidden layers," referred to as *neurons*, between the input layer, i.e., the observations, and the output layer, i.e., the field estimate; and the intermediate, multiply linked pathways are modeled with nonlinear functions. This generic structure gives the technique great versility. Its practical application to data analysis and forecasting can be a rather complex process, as discussed by Hsieh and Tang (1998).

SEE ALSO THE FOLLOWING ARTICLES

GEOSTATISTICS • KALMAN FILTERS AND NONLINEAR FILTERS

BIBLIOGRAPHY

Aunon, J., and Gomez-Hernandez, J. J. (2000). "Dual kriging with local neighborhoods: Application to the representation of surfaces," *Math. Geo.* **32,** 69–85.

Borzelli, G., and Ligi, R. (1999). "Empirical orthogonal function analysis of SST image series: A physical interpretation," *J. Atmos. Oceanic Techno.* **16,** 682–690.

Emery, W. J., and Thomson, R. E. (1997). "The spatial analyses of data fields." *In* "Data Analysis Methods in Physical Oceanography," Chapter 4, Elsevier Science, New York.

Hsieh, W. W., and Tang, B. (1998). "Applying neural network models to prediction and data analysis in meteorology and oceanography," *Bull. Am. Meteoro. Soc.* **79,** 1855–1870.

Luo, Z., and Wahba, G. (1997). "Hybrid adaptive splines," *J. Am. Stat. Assoc*. **92,** 107–116.

Luo, Z., Wahba, G., and Johnson, D. R. (1998). "Spatial–temporal analysis of temperature using smoothing spline ANOVA," *J. Climate* **11,** 18–28.

Spencer, P. L., Janish, P. R., and Doswell III, C. A. (1999). "A four-dimensional objective analysis scheme and multitriangle technique for wind profiler data," *Monthly Weather Rev*. **127,** 279–291.

Thiébaux, H. J. (1997). "The power of the duality in spatial–temporal estimation," *J. Climate* **10,** 567–573.



Sphere Packing

Gábor Fejes Tóth
Hungarian Academy of Sciences

Wlodzimierz Kuperberg
Auburn University

I. The Sphere Packing Problem—Its Statement and History
II. Lattice Packing of Balls—Motivation from Number Theory
III. Packing of Convex Bodies
IV. Sphere Packings and Codes

GLOSSARY

Convex set A set in Euclidean n-dimensional space E^n containing every line segment connecting two points of the set.

Convex body A compact convex set with nonempty interior.

Convex disk A convex body in the Euclidean plane E^2.

Density of a packing Intuitively, the percentage portion of space occupied by the solids arranged in the packing, a precise definition involves the notion of limit.

Dirichlet–Voronoi cell (D–V cell) A convex polyhedron associated with a ball in a ball packing, consisting of points whose distance from the center of the ball does not exceed the distance from the center of any other ball.

Face centered cubic lattice (fcc lattice) The lattice consisting of all vertices and facet centers of the cubes in a face-to-face tiling of E^3 with cubes.

Lattice The collection of vertices in a face-to-face tiling of Euclidean n-dimensional space E^n by congruent parallelepipeds, or, equivalently, the collection of integer–coefficient linear combinations of a basis for R^n.

Local density The ratio between the volume of a ball and the volume of its Dirichlet–Voronoi cell.

Lattice packing A packing whose members are translates of each other with the corresponding translation vectors forming a lattice.

Packing A collection of sets with mutually disjoint interiors.

Regular dodecahedron The 12-faceted Platonic solid—the convex polyhedron whose facets are regular pentagons.

Rhombic dodecahedron The 12-faceted convex polyhedron circumscribed about a sphere, whose facets are congruent rhombi with diagonal ratio of $\sqrt{2}$-to-1.

Sphere packing A collection of congruent solid spheres (balls) with mutually disjoint interiors.

THE CLASSICAL sphere packing conjecture, also known as the Kepler conjecture, asserts that the maximum density of a packing of Euclidean three-dimensional space E^3 with congruent balls is attained in the fcc lattice packing. The conjecture remained open for almost 400 years, until the relatively recent announcement of a

computer-aided proof by Thomas C. Hales, confirming its veracity. The interest in sphere packings, especially of the lattice type, was spurred first by their connections to number theory. More recently, the development of computer science gave a boost to the entire field of discrete geometry that includes the topics of packing, covering, and tiling. In particular, various designs of efficient error-correcting codes for electronic transmission of information, via a geometric interpretation of such codes, generated an array of dense sphere packings in Euclidean spaces of high dimension.

I. THE SPHERE PACKING PROBLEM—ITS STATEMENT AND HISTORY

In its simplest, classical form, the *sphere packing problem* asks for the most efficient way of filling space with congruent, nonoverlapping solid spheres (balls). The readily anticipated, natural solution, seen often as stacks of cannonballs in old paintings or as stacks of oranges or apples in supermarkets, is nicely described in Johannes Kepler's treatise published in year 1611, and it goes as follows. Build a triangular pyramid of congruent balls by placing a single ball (A) on top of 3 mutually tangent ones (B), then place this small stack upon a triangular layer of 6 balls (C), then again place this larger stack upon a triangular layer of 10 balls (D), and so on, as shown in Fig. 1 reproduced here from Kepler's booklet. The pattern of the stack can be extended in all directions, resulting in an arrangement

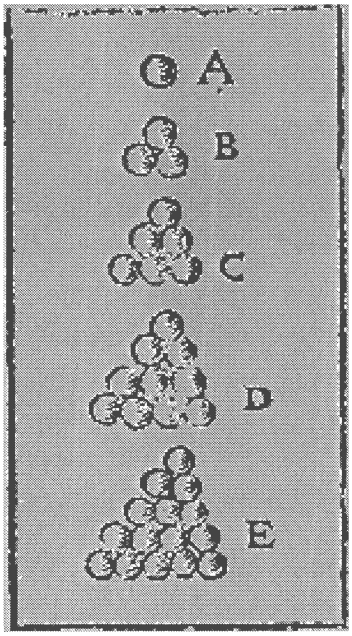


FIGURE 1 Kepler's drawing: a triangular stack of pellets.

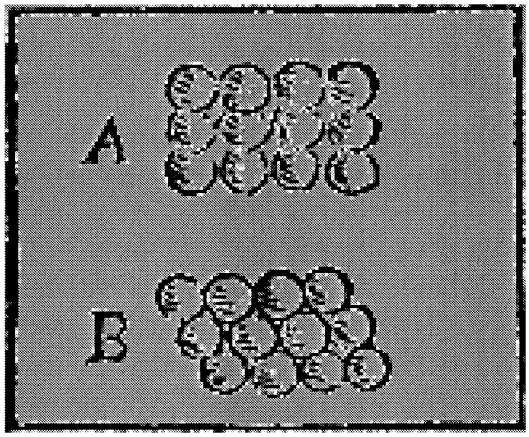


FIGURE 2 Kepler's drawing: a square layer and a triangular layer.

of mutually nonoverlapping balls (a packing) uniformly distributed throughout the whole space. This packing consists of flat layers of balls, in which the balls' centers are coplanar and form the familiar triangular pattern of points. Each ball in the packing is tangent to 12 other balls: 6 in its own layer and 3 in each of the two adjacent layers.

Kepler also describes the square pyramid pattern of packing balls, consisting of flat layers in which the balls' centers form a square grid instead of the triangular one (see Fig. 2). In this arrangement each ball is tangent to 12 other balls as well: 4 in its own layer and 4 in each of the two adjacent layers. This is not just a coincidence. Kepler observes that the two patterns, triangular and square, when extended to fill the whole space, produce essentially the same (congruent) packings, as one can be transformed into the other by a rotation.

This packing, Kepler asserts, is the "tightest possible," but neither does he specify what he precisely means by "tightness," nor does he offer even a hint of rigorous, mathematical proof of this assertion. Moreover, some of his pronouncements, such as the one he makes about partitioning space: "*space in the round cannot be divided without remainder except into cubes and rhomboids*," are simply incorrect.

Nowadays, the concept of *density* of a packing is used to measure its efficiency, or tightness, and it is defined in the following way. Let $\mathcal{P}$ be a packing consisting of solids $\{B_1, B_2, B_3, \ldots\}$, in this case congruent balls, and let $B(r)$ denote the ball of radius r centered at the origin. (The use of the term *packing* includes the implicit assumption that the interiors of the solids are mutually disjoint.) Then the density of $\mathcal{P}$ is defined as

$$d(\mathcal{P}) = \limsup_{r\to\infty} \left(\frac{1}{|B(r)|} \sum_{i=1}^{\infty} |B_i \cap B(r)| \right),$$

where the volume of a set S is denoted by $|S|$.

The density of the ball packing described by Kepler is easily computed to be $\pi/\sqrt{18} = 0.74048\ldots$. Hence, in terms of density, his assertion on the maximum tightness of a ball packing can be stated as

The density of a packing with congruent balls cannot exceed $\pi/\sqrt{18}$.

and in this form it became known as the *Kepler conjecture*.

For the sake of historical accuracy, it should be mentioned that Kepler's interest in sphere packings as a tool for investigations of the atomic structure of matter was preceded and influenced by the work of Thomas Harriot, a mathematical assistant to Sir Walter Raleigh.

Chronologically, the first rigorous result in the direction of the Kepler conjecture came in 1831, in a comment of C. F. Gauss concerning properties of positive quadratic forms. In that context, Gauss considers a special kind of ball packings, called *lattice packings*, in which the centers of the balls form a *lattice* of points, *i.e.*, the collection of vertices of congruent parallelepipeds tiling space in a face-to-face manner (for example, the *cubic lattice* is the collection of vertices in a face-to-face space tiling by cubes). Gauss proves that among all possible lattice ball packings, the densest one is unique, and that it is precisely the one described by Kepler. The lattice underlying this packing is called the *face-centered cubic lattice* or, in abbreviation, the *fcc-lattice*, since it can be described by adding to the points of a cubic lattice the centers of the faces of the cubes. The connection between lattice ball packings and number theory is discussed in the next section of this article.

The general case of the Kepler conjecture turned out to be extremely difficult to resolve, despite the common belief in its correctness and numerous attempts at a proof. Among partial results is a succession of upper bounds on the density of ball packings, gradually getting closer to, but all falling short of (above, actually) the ultimate bound of $\pi/\sqrt{18}$.

The first of such bounds was given by H. F. Blichfeldt in 1929. Blichfeldt considers the more general problem of ball packings in n-dimensional Euelidean space E^n. Analytically, the unit ball in R^n is the set B^n of points $\mathbf{x}$ with $\|\mathbf{x}\| \leq 1$, where $\|\mathbf{x}\|$ denotes the distance from $\mathbf{x}$ to the origin. Blichfeldt shows that the number $(n+2)2^{-(n+2)/2}$ is an upper bound for the density of such packings. In dimension $n = 3$, this produces an upper bound of $0.88388\ldots$, a seemingly modest result considering the conjectured least upper bound of $0.74048\ldots$, but the significance of Blichfeldt's work lies in its breakthrough nature, its generality, and the simplicity of his ingenious method. Following is a short description of Blichfeldt's idea.

Given a packing $\{B_i\}$ of R^n with unit balls, replace each ball with a concentric one of radius $\rho > 1$ (these may now overlap, like spherical clouds) and furnish each enlarged ball with a mass distribution, variable from point to point, but translation invariant from ball to ball. If the total mass density at each point of space is bounded above by a constant C, then the density of the original packing of unit balls is at most $C\omega_n/M(n)$, where ω_n is the volume of the unit ball and $M(n)$ is the mass of the enlarged ball. A suitable choice of ρ and of mass distribution can produce a meaningful density bound for ball packings in R^n. Blichfeldt obtains his bound by taking $\rho = \sqrt{2}$, setting the mass distribution as $m(p) = 2 - \text{dist}^2(p, c_i)$, where c_i is the center of B_i, and by estimating that the (total) mass density does not exceed 2 at any point.

Elaborating on Blichfeldt's technique, R. A. Rankin improved his result in 1947 and gave a bound of $0.828\ldots$ for ball packings in R^3. Rankin obtained this result by suitably refining the mass distribution function and by utilizing the newly invented at that time electronic computing machines. The computational algorithm designed by Rankin was used to obtain improvements of Blichfeldt's bound in dimensions higher than 3 as well, but by tedious, computer-aided calculations, one dimension at a time, in contrast to the compact, analytical formula given by Blichfeldt.

C. A. Rogers established in 1958 a better bound that can be described in very natural geometric terms. Consider the regular simplex of edge length 2 in E^n and the collection of $n+1$ unit balls centered at the vertices of the simplex. Let b denote the total volume of the $n+1$ sectors of the balls contained in the simplex, and let s be the volume of the simplex. Then the ratio $\sigma_n = b/s$ is the Rogers bound on the density of ball packings in n-dimensional space E^n.

For $n = 2$, the bound $\sigma_2 = \pi/\sqrt{12}$ is sharp, as it is attained in the familiar triangle-patterned circle packing of the plane (see Fig. 3). Thus, this special case of the Rogers bound provided another proof of the optimality of this circle packing, originally established by A. Thue in 1910. For $n = 3$, the Rogers bound of $\sigma_3 = 0.7797\ldots$ is considerably better than the corresponding Blichfeldt bound.

Rogers' technique is based on estimating the volume of the *Voronoi region* enclosing a ball in a packing, a useful concept that goes back to P. G. L. Dirichlet and is sometimes called the *Dirichlet cell*, or the *Dirichlet–Voronoi cell* (*D–V cell*, in abbreviation). The D–V cell of a ball consists of all points whose distance from the center of the ball does not exceed the distance from the center of any other ball of the packing. Each D–V cell is a convex polyhedron. All D–V cells of balls in a packing have

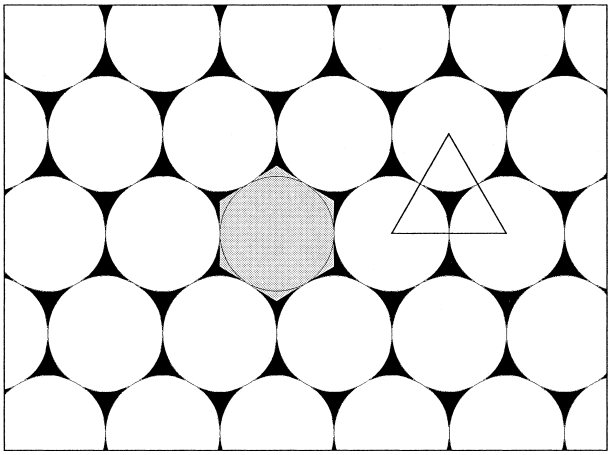


FIGURE 3 Densest circle packing. One D–V cell shown (hexagon), density's reaching Rogers bound indicated (triangle).

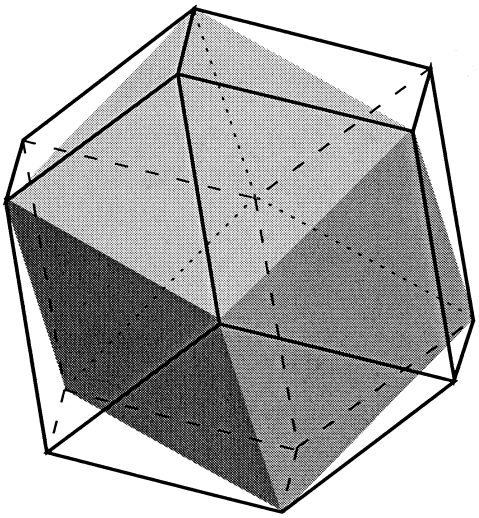


FIGURE 4 Rhombic dodecahedron. (The short diagonals of the rhombic facets form the skeleton of a cube.)

mutually disjoint interiors, and their union is the entire space. A lower bound on the volume of each D–V cell in a packing of unit balls immediately yields an upper bound on the density of the packing. The ratio between the volume of a ball and its D–V cell is called the *local density* at the ball, and the (global) density of a packing cannot exceed the maximum among its local densities. Rogers established his bound by partitioning the D–V cell radially from the ball's center into simplices and analyzing the ratio between the volume of the sector of the ball corresponding to such a simplex and the volume of the simplex.

By a refinement of Rogers' technique for $n = 3$, J. H. Lindsay improved in 1983 the upper bound to $0.77844\ldots$ and in 1993 D. J. Muder lowered it further down to $0.7731\ldots$.

Regarding the relation between the local and global densities, in the Euclidean plane E^2 the situation is relatively simple: the minimum-area D–V cell of a circular disk is the circumscribed regular hexagon, and in the triangle-patterned circle packing in which each circle is tangent to six others (see Fig. 3) each D–V cell is such a minimal regular hexagon. Thus, in E^2, the maximum global density of a circle packing is attained by maximizing the local density at each circle.

Things are not so simple in E^3: in the fcc-lattice ball packing described by Kepler, each D–V cell is a rhombic dodecahedron, a 12-faceted polyhedron circumscribed about the ball, whose every facet is a rhombus with a diagonal ratio of $\sqrt{2}$-to-1 (see Fig. 4), while D–V cells of smaller volume are easily attainable in ball packings.

For instance, the configuration of 12 unit balls tangent to a central unit ball at the midpoints of the facets of a circumscribing *regular* dodecahedron can occur, making the regular dodecahedron the D–V cell of the central ball. While the volume of the rhombic dodecahedron circumscribed about the unit ball is $4\sqrt{2} = 5.656\ldots$, the volume of the regular dodecahedron is only $5.548\ldots$. Thus, the local density of a ball packing in E^3 can be, at some balls, greater than the conjectured global maximum density.

This raises the question of the maximum local density in a ball packing. L. Fejes Tóth conjectured in 1942 that

The minimum volume D–V *cell in ball packings in* E^3 *is the regular dodecahedron.*

and this became known as the *dodecahedral conjecture*. (The upper bounds established by Rogers, Lindsay, and Muder should thus be considered as partial results toward the dodecahedral conjecture rather than the Kepler conjecture.)

Compounding the difficulty of the Kepler conjecture is the fact that, while the conjectured densest, fcc-lattice packing has local density of $\pi/\sqrt{18}$ at each of the balls, it is not the only ball packing with this property. Recall the triangle-patterned layers of this lattice packing and observe that each such layer has "deep holes" into which the balls from an adjacent layer are accommodated. But the adjacent layer occupies only one of the two triangular grids of holes, and it is possible to choose the other one for the placement of the adjacent layer instead (see Fig. 5).

These two possibilities give rise to a multitude of noncongruent layered ball packings, among which the lattice one is generated only if the shift vector from layer to layer is repeated consistently. Each of these "laminated" ball packings is of the same density, as the distance between

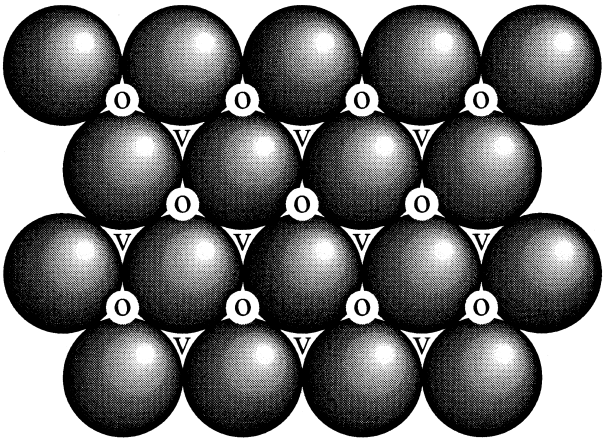


FIGURE 5 The deep holes in a triangular layer: occupied (o) and vacant (v).

the adjacent layers is the same in all of them, but in the nonlattice laminated packing the D–V cells consist, in most cases, of two congruence classes: some of them are rhombic dodecahedra, and the remaining ones form another congruence class. The exception is one particular laminated packing in which all D–V cells are of the second type (the shifts from layer to layer alternate each time). The existence of laminated packings, although not mentioned by Kepler, had already been discovered by Harriot with whom he communicated extensively on the matters of ball packing.

In each laminated packing every ball is touched by 12 balls, but the configurations of the 12 balls tangent to the central one, extracted from the laminated packings, are of two noncongruent types. Also, recall the configuration of 12 balls tangent to a central ball at the midpoints of the facets of a circumscribing regular dodecahedron. In this regular configuration, the 12 balls, while tangent to the central one, are not in contact with each other. This is so because the dihedral angle of the regular dodecahedron is smaller than 120°. Such a multitude and apparent looseness of the configurations of the 12 balls around the central one suggest the intriguing question: *What is the maximum number of congruent nonoverlapping balls that can come in contact with one ball of the same size?* This question was the subject of a famous discussion between Sir Isaac Newton and David Gregory. Newton's opinion was that 12 should be the maximum, while Gregory thought that 13 might be possible. One hundred eighty years passed before the question was settled, by R. Hoppe. Newton was right, although experiments and computations seem to give some credence to Gregory's opinion. The 12 tangent balls can be rearranged so as to make room for a 13th ball to come very close to tangency.

Realizing that local density alone is not the key to the solution of the Kepler conjecture, Fejes Tóth proposed in 1953 to consider a cluster of up to 12 D–V cells surrounding one cell, and to estimate a certain weighted average of the cell's volumes. He thereby presented a program that could potentially reduce the problem to the determination of the minimum value of a function of a finite number of variables. Considering the large number of variables involved and the overwhelming complexity of the function to be analyzed, he suggested that computers be used to carry out the astronomical amount of calculations required to complete his program.

It now appears that, due in great part to the recent development of computer technology and computing methods related to it, a confirmation of the Kepler conjecture along a program similar to that of Fejes Tóth became possible. In 1998, Thomas C. Hales announced that he completed a proof of the Kepler conjecture following years of investigations accompanied by an enormous amount of numerical calculations performed by computers. Part of the work was done by Samuel Fergusson who assisted Hales in resolving one of the critical cases. Hales' approach to the problem resembles the program of Fejes Tóth in the idea of building a space partition associated with a given ball packing and considering, along with one central ball, only a limited number of balls, not too distant from the central one. But the partition he constructs is neither of the Voronoi, nor of the dual to it, Delone type, but rather a hybrid of the two, whose definition is quite intricate, involving modifications and exceptions in several special cases.

Equally involved is the way Hales manages to balance the weights (through his elaborate "scoring system," as he calls it) to average the densities corresponding to the pieces of the partition. Even the classification of the about 5000 planar maps (i.e., the combinatorially equivalent configurations of the neighboring ball centers), too complex for a human brain to perform, was done by a computer before each case could then be analyzed and eliminated as a potential counterexample. The case analysis involved thousands of inequalities and equations, and it required a careful design of interval arithmetic to ensure the necessary computational accuracy and avoid false conclusions.

The voluminous manuscript of Hales' work describing his approach in detail and the theory behind the computations has been submitted for publication, and, at the time of the writing of this article, while believed to be essentially correct, is undergoing extensive editorial revisions. In the meantime, Hales and his student, Sean McLaughlin, announced and produced jointly a preprint with a proof of the dodecahedral conjecture, in which they utilize the strategy and computing machinery developed by Hales in his proof of the Kepler conjecture.

II. LATTICE PACKING OF BALLS—MOTIVATION FROM NUMBER THEORY

A lattice in R^n is the collection of all integer–coefficient linear combinations of a basis for R^n. Due to its periodicity, the density of the lattice packing of balls whose centers lie at the points of Λ is easily computed as the ratio between the volume of the ball and the volume of the parallelepiped spanned by the basis of Λ. The volume of the parallelepiped spanned by the vectors $\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{in})$, $i = 1, \ldots, n$ is given by $|\det(\mathbf{A})|$, the absolute value of the determinant of the matrix $\mathbf{A} = (a_{ij})$. Among all lattice packings of R^n with unit balls there exists one of maximum density, and that density, denoted by $\delta_L(B^n)$, is called the *lattice packing density* of the ball.

Applications of lattice packings of balls to number theory were discovered already by Gauss, but it was H. Minkowski who by a systematic study of the theory of lattice packings developed a new branch of mathematics which he named the *geometry of numbers*. The early development of the theory of packing was motivated to a great extent by arithmetic problems. As an illustration we discuss the problem of arithmetic minimum of homogeneous positive definite quadratic forms.

Consider a lattice Λ generated by the vectors $\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{in})$, $i = 1, \ldots, n$. With Λ we associate the quadratic form

$$f(\mathbf{x}) = f(x_1, \ldots, x_n) = \sum_{i=1}^{n} (a_{i1}x_1 + \cdots + a_{in}x_n)^2.$$

We investigate the minimum f_0 of $f(\mathbf{x})$ taken over nonzero integer vectors:

$$f_0 = \min_{\mathbf{x} \in Z^n} f(\mathbf{x}).$$

Observe that f_0 is nothing else but the square of the length of the shortest nonzero vector of Λ. Thus the system of balls of radius $\sqrt{f_0}/2$ centered at the elements of Λ form a packing. The density of this lattice packing is $f_0^{n/2}\omega_n/|\det(\mathbf{A})|$. Hence we get for f_0 the upper bound

$$f_0 \leq \left(\frac{\delta_L(B^n)|\det(\mathbf{A})|}{\kappa_n} \right)^{2/n}.$$

This inequality is sharp. Equality is attained if the vectors $\mathbf{a}_i$ form a basis for a densest lattice packing with balls of radius $\sqrt{f_0}/2$.

Any homogeneous positive definite quadratic form

$$g(\mathbf{x}) = g(x_1, \ldots, x_n) = \sum_{i,j=1}^{n} s_{i1}x_i x_j \quad (s_{ij} = s_{ji})$$

can be written in the form

$$g(\mathbf{x}) = \sum_{i=1}^{n} (a_{i1}x_1 + \cdots + a_{in}x_n)^2$$

with appropriate constants a_{ij}. The representation is not unique but we have for any such representation that the *discriminant* $D = \det(s_{ij})$ of $g(\mathbf{x})$ is equal to $\det(\mathbf{A}^2) = (\det(a_{ij}))^2$. Thus we get the following theorem.

For each homogeneous positive definite quadratic form $g(\mathbf{x})$ with discriminant D there exists a nonzero integer vector $\mathbf{v}$ such that

$$g(\mathbf{v}) \leq \left(\frac{\delta_L(B^n)D^{1/2}}{\kappa_n} \right)^{2/n}.$$

The packing density $\delta_L(B^n)$ is known for $n \leq 8$. We have

$$\delta_L(B^2) = \frac{\pi}{\sqrt{12}}, \quad \delta_L(B^3) = \frac{\pi}{\sqrt{18}}, \quad \delta_L(B^4) = \frac{\pi^2}{16},$$

$$\delta_L(B^5) = \frac{\pi^2}{15\sqrt{2}}, \quad \delta_L(B^6) = \frac{\pi^3}{48\sqrt{3}}, \quad \delta_L(B^7) = \frac{\pi^3}{105},$$

$$\delta_L(B^8) = \frac{\pi^4}{384}.$$

The cases $n = 2$ and $n = 3$ were solved by J. L. Lagrange and Gauss, respectively. A. Korkin and G. Zolotarev found the values of $\delta_L(B^4)$ and $\delta_L(B^5)$ and Blichfeldt settled the cases $n = 6$, 7, and 8. There is an extensive literature dealing with the construction of efficient ball packings. We describe some of these constructions in Section IV where we give also the lattice packings of balls realizing the extreme densities $\delta_L(B^n)$ for $n \leq 8$.

III. PACKING OF CONVEX BODIES

Generally, the packing theory deals with packings and lattice packings of *convex bodies*, a class of solids that includes balls. A convex body is a compact convex set with nonempty interior. Given a convex body K in E^n, a *packing with replicas of K* is a family of solids $\{K_1, K_2, K_3, \ldots\}$, each congruent to K, whose interiors are mutually disjoint. If all of the K_i's are parallel translates of each other, and if the corresponding translation vectors form a lattice, then the packing is a *lattice packing of K*. The density of a packing with replicas of K is defined in the same way as the density of a ball packing (see Section I of this article), as the corresponding lim sup. The *packing density of K*, denoted by $\delta(K)$, is defined as the supremum of the densities of all packings with congruent copies of K. According to a theorem of H. Groemer the supremum is attained. Restricting ourselves to lattice packings we obtaine the analogously defined *lattice packing density* $\delta_L(K)$ of K.

The general theory of packing of convex bodies in n dimensions is still in its early stages of development, with many of its basic problems yet to be resolved. However, the case of $n = 2$ (the plane) has been explored quite extensively and successfully. From the multitude of results on packing the plane with congruent convex disks (planar convex bodies), we choose to mention here a theorem of L. Fejes Tóth because of its significance and substantial consequences: If K is a convex disk and H is the minimum area hexagon containing K, then $\delta(K) \leq |K|/|H|$. Since the minimum area hexagon containing the circular disk is the regular circumscribed hexagon, the theorem of Fejes Tóth implies immediately the theorem of Thue on the maximum density of circle packings. Another significant consequence of the theorem of Fejes Tóth is obtained through a theorem of C. H. Dowker on centrally symmetric convex disks. Dowker proved that if K is a centrally symmetric convex disks, and $k \geq 4$ is an even integer, then among all minimum area k-sided polygons containing K there is a centrally symmetric one. Since every centrally symmetric hexagon admits a lattice tiling (*i.e.*, a lattice packing without residue) of the plane, it follows that

$$\delta(K) = |K|/|H| = \delta_L(K)$$

for every convex centrally symmetric disk.

No result analogous to the theorem of Fejes Tóth for centrally symmetric convex bodies K in dimensions higher than 2 exists, and the identity $\delta(K) = \delta_L(K)$ for such bodies does not hold true already for $n = 3$.

In 1905 Minkowski proved that for all n

$$\delta_L(B^n) \geq \zeta(n)/2^{n-1},$$

where

$$\zeta(n) = \sum_{i=1}^{\infty} i^{-n}.$$

This is a special case of the more general inequality

$$\delta_L(K) \geq \zeta(n)/2^{n-1}$$

valid for all centrally symmetric convex bodies K. This inequality bears the name of the *Minkowski–Hlawka theorem*, since it was stated implicitly by Minkowski in 1898 and proved first in 1944 by E. Hlawka. Many alternative proofs and refinements of the theorem were given, although no essential improvement is known for large n. All proofs make use of an averaging argument and thus do not say how lattices satisfying the inequality may be found. The best refinement of the Minkowski-Hlawka theorem is due to W. Schmidt who established the inequality

$$\delta_L(K) \geq cn2^{-n}$$

for every centrally symmetric convex body K and for every sufficiently large n provided $c \leq \log 2$. K. Ball used a variational argument to show the existence of lattice ball packings with a slightly greater density. He established

$$\delta_L(B^n) \geq 2(d-1)\zeta(n)/2^n,$$

which is the presently best lower bound known for $\delta_L(B^n)$.

Rogers conjectures that for sufficiently high dimensions, $\delta(B^d) > \delta_L(B^d)$. Thus far, this inequality has not been confirmed in any dimension. Moreover, except of dimensions 10, 11, and 13 (see the next section) all the best packings known are lattice packings. Also in the case of upper bounds the seeming disadvantage of lattice arrangements over general ones could not be exploited so far.

We mentioned Blichfeldt's upper bound

$$\delta(B^n) \leq \frac{n+2}{2} 2^{-n/2}$$

and its improvements by Rankin and Rogers. Although these improvements are significant in E^3, asymptotically they are only by a constant factor smaller than Blichfeldt's bound. An improvement in the exponential order of magnitude was given first by V. M. Sidelnikov in 1973. The presently best asymptotic upper bound known,

$$\delta(B^n) \leq 2^{-(0.599+o(1))d} \qquad (\text{as } d \to \infty),$$

was proved by G. A. Kabatjanskiĭ and V. I. Levenštein.

It is worth noting that the proofs of these bounds use analytic methods borrowed from coding theory. An analogue of the linear programming bound for error-correcting codes could be established first only for the ball packing problem in spherical geometry, from which the Euclidean case follows, e.g., by a Blichfeldt type argument.

Let S^n denote the n-dimensional spherical space, that is the boundary of B^{n+1}. The angular distance between two points x and y on S^n is the angle between the radii ox and oy. Let $M(n, \varphi)$ be the maximum number of caps of diameter φ forming a packing on S^n. Equivalently, $M(n, \varphi)$ is the maximum number of points on S^n all of whose angular distances are greater than or equal to φ. Still another interpretation of $M(n, \varphi)$ is that it is the maximum number of points different from the origin in E^{n+1} such that any pair of them spans an angle at the origin greater than or equal to φ.

Kabatjanskiĭ and Levenštein established the bound

$$M(n, \varphi) \leq (1-\cos\varphi)^{-n/2}\, 2^{-d(0.099+o(1))} \qquad (\text{as } d \to \infty)$$

for all $\varphi \leq \varphi^* = 62.9974\ldots^\circ$. For large values of n this bound is an essential improvement on earlier bounds by Rankin and K. Böröczky. It implies the bound for $\delta(B^n)$ as follows. In a packing of unit balls in E^n replace each ball by a concentric one of radius λ, $1 < \lambda < 2$. Consider an arbitrary point $p \in E^n$ and two centers of balls, x and y

within distance λ from p. Using the fact that the distance between x and y is at least 2, it is easy to check that the size of $\angle xpy$ is at least arccos $(1–2/\lambda^2)$. Thus, there are at most $M(n-1, \arccos(1-2/\lambda^2))$ centers of balls within distance λ from p, that is p is covered by at most $M(n-1, \arccos(1-2/\lambda^2))$ enlarged balls. Hence we obtain

$$\delta(B^n) \le \lambda^{-n} M(n-1, \arccos(1-2/\lambda^2)) \le 2^{-(0.599+o(1))d}.$$

A finite set of points on S^n is often referred to as a *spherical code*. The use of this term is justified by the following. Suppose that an information source emits words from an alphabet of size M. In order to transmit the information through a radio channel we encode the alphabet into analog symbols: with each letter we associate a function $f_i(t)$ $(0 \le t \le t_0, i = 1 \ldots, M)$, the amplitude of the radio wave to be transmitted. We assume that the radio signals have equal energy, say 1 unit and furthermore the frequency of the signals is bounded. This means that in the Fourier expansion of the signals all coefficients, except say the first $n+1$ are zero. Remembering that under right choice of the units the energy of a wave is equal to the sum of the squares of its Fourier coefficients, the signals $f_i(t)$ can be viewed as points on S^n. In order to distinguish the signals we have to keep the minimum distance among them as big as possible. We are lead to the problem considered here, just this time in the dual formulation: for given n and M we are looking for the maximum number $\varphi(n, M)$ such that M points with mutual angular distances at least $\varphi(n, M)$ can be placed on S^n.

Quite recently, H. L. Cohn and N. Elkies announced that they developed an analog of the linear programing bound for error-correcting codes to sphere packings in E^n, and used it to calculate upper bounds for the density of sphere packings, obtaining the best bounds known for dimensions 4–36. They expect that their approach can be used to solve the sphere packing problem in dimensions 8 and 24, where their bound thus obtained are extremely close to the densities of the best packings known, described in the next section as the E_8 and the Leech lattice packings, respectively.

IV. SPHERE PACKINGS AND CODES

In this section we consider further connections of sphere packings to coding theory. A *binary code* is a set of n-dimensional binary vectors (or $\{0, 1\}$-*words* of length n). The *weight* of a word is the number of its coordinates that differ from zero. The *Hamming distance* between two such vectors (or *codewords*) is the number of coordinates at which they differ. The greater the distance between two codewords, the easier it is to distinguish one from the other. Thus, reliable and efficient storage and transmission of information requires an explicitly constructed code with a large number of codewords as far from one another as possible. This can be interpreted as the sphere packing problem in the Hamming space.

A code consisting of M words of length n and with minimum Hamming distance d between any two of its words is called an (n, M, d)-code. The set of all binary words of length n is a linear space over the binary field GF(2). A code which is a subspace of that linear space is called a *linear code*, which codes are of special importance in coding theory and turn out to be interesting from a geometric point of view.

Geometrically, an (n, M, d)-code is a subset of the set of vertices of the n-dimensional unit cube. The balls of radius $\frac{1}{2}\sqrt{d}$ centered at the codewords form a packing in E^n. In a suitably chosen manner, this local sphere packing can be used to generate a periodic packing in the whole space E^n. We describe two basic constructions devised by J. Leech and N. J. A. Sloane to generate such packings.

Construction A The point $x = (x_1, x_2, \ldots, x_n)$ is a center of a sphere if x is congruent (modulo 2) to a codeword. Geometrically, this means that the unit cube, with the spheres attached to it, is translated to new positions, filling a "checkerboard" pattern in E^n. In this construction, the radii of the spheres are $\min\{1, \frac{1}{2}\sqrt{d}\}$.

Construction B This construction applies to codes in which every codeword is of even weight. The point $x = (x_1, x_2, \ldots, x_n)$ is a center of a sphere if x is congruent (modulo 2) to a codeword and $\sum_{i=1}^{n} x_i$ is divisible by 4. The radius of the spheres is $\min\{\sqrt{2}, \frac{1}{2}\sqrt{d}\}$.

Each of the constructions A and B yields a lattice packing if and only if the code on which it is based is linear.

The densest lattice packing of balls in E^n for $n = 3, 4$ and 5 is obtained by applying Construction A to the $(n, 2^{d-1}, 2)$-code consisting of all words of even weight. The densest lattice packing of balls in E^7 and E^8 can be obtained by construction A using codes obtained from an Hadamard matrix. An *Hadamard matrix* H is an $n \times n$ matrix with entries ± 1 such that $HH^T = nI$. It is known that an Hadamard matrix can exist only for $n = 1, 2$ and multiples of 4. In the case when n is a power of 2, an $n \times n$ Hadamard matrix H_n can be easily obtained by induction, setting $H_1 = (1)$ and

$$H_{2k} = \begin{pmatrix} H_k & H_k \\ H_k & -H_k \end{pmatrix}$$

for $k = 1, 2, 4, \ldots$. We consider the matrix H_8 so obtained, and observe that its first row and column contain only 1-s. Next we consider the matrix $\widetilde{H}_8$ obtained from H_8 upon replacing the $+1$'s by 0's and the -1's by 1's. The rows of $\widetilde{H}_8$ with the first column deleted form a $(7, 8, 4)$-code. The

codewords of this code are the vertices of a 7-dimensional regular simplex. We obtain the densest lattice packing of balls in E^7 by applying Construction A to this code. The densest lattice packing of balls in E^8 is obtained through Construction A used on the (8, 16, 4)-code consisting of the rows of H_8 together with the complements of the rows. The underlying lattice is denoted by E_8. The same lattice of balls can be obtained also by Construction B applied to the trivial (8, 2, 8)-code consisting of the all 0's and all 1's words. We mention that the densest lattice packing of balls in E^6 cannot be obtained directly by Construction A or B, but is obtained as an appropriate section of the densest lattice ball packing in E^7.

A remarkably dense lattice packing of balls in E^{24}, playing a fundamental role not only in the theory of ball packings but in group theory as well, was constructed by Leech. The construction of the Leech lattice is based on the *Golay code* G_{24}, a $(24, 2^{12}, 8)$-code, which can be described as follows: Let $C_{12} = (c_{ij})$, $i, j = 1, \ldots, 12$ be defined by

$$c_{ij} = \begin{cases} 0, & \text{if } i = j = 1 \text{ or } i, j \geq 2 \text{ and } i + j - 4 \\ & \text{is a quadratic residue mod 11;} \\ 1, & \text{otherwise} \end{cases}$$

and let I_{12} be the unit matrix of order 12. Then G_{24} is the 12-dimensional linear space (over $GF(2)$) spanned by the rows of the matrix (I_{12}, C_{12}). Each codeword in the Golay code is of even weight; in fact the weight of each codeword is divisible by 4.

An application of Construction B to G_{24} results in a lattice packing of balls of radius $\sqrt{2}$, which makes possible to join this packing, without overlap, with a translate of itself through the vector $(-\frac{5}{2}, \frac{1}{2}, \ldots, \frac{1}{2})$ thereby to form a new lattice packing, called, after its inventor, the *Leech lattice*. The density of this packing is $0.001930\ldots$, which compares quite favorably to the Rogers bound $\delta(B^{24}) \leq \sigma_{24} = 0.002455\ldots$. Each ball in the Leech lattice is touched by 196560 other balls, a number which happens to be the maximum possible contact number in any packing of congruent balls in E_{24}. Each of the densest known lattice packings in dimensions below 24 can be obtained as an appropriate section of the Leech lattice.

We mentioned the conjecture of Rogers according to which for sufficiently high dimensions, $\delta(B^d) > \delta_L(B^d)$. Although this inequality has not been verified for any dimension so far, in dimensions $n = 10, 11$, and 13, nonlattice ball packings denser than any *known* lattice packing have been found. These examples have been produced by applying Construction A to certain nonlinear codes.

Rush and Sloane used a modification of construction A to establish the existence of dense packings of balls and certain other convex bodies. They used to codes over larger alphabets and replaced the Hamming metric a metric generated by the convex body. For "superballs" $\{(x_1, x_2, \ldots, x_n) \in E^n : \sum_{i=1}^n x_i^\sigma \leq 1\}$ for the sequence $d_j = p_j^\sigma$ of dimensions, where p_j is an odd prime their bound exceeds the Minkowski-Hlawka bound. In a series of papers Rush improved these results further. Rush's method yields the lower bound $\delta(B^n) \geq (\frac{1}{2})^{n+o(n)}$ for the packing density of balls in E^n, which is remarkably close to the bound of the Minkowski-Hlawka Theorem. Unfortunately, the codes used in this approach cannot be explicitly given, so the method cannot be considered as "constructive."

SEE ALSO THE FOLLOWING ARTICLES

CONVEX SETS • DESIGNS AND ERROR-CORRECTING CODES

BIBLIOGRAPHY

Ball, K. M. (1993). "A lower bound for the optimal density of lattice packings," *Duke J. Math.* **68,** 217–221.

Cohn, H. (2000). New Bounds on Sphere Packings, Doctoral Thesis, Harvard University, Cambridge, MA, April 2000; also: H. Cohn and N. Elkies, in preparation.

Conway, J. H., and Sloane, N. J. A. (1998). "Sphere Packings, Lattices and Groups," Third edition, Springer-Verlag, New York.

Fejes Tóth, G., and Kuperberg, W. (1993). Packing and covering with convex sets. *In* "Handbook of Convex Geometry" (P. M. Gruber and J. M. Wills, eds.), pp. 799–860, North-Holland, Amsterdam.

Ferguson, S. P. (1998). Sphere packings V. Preprint.

Ferguson, S. P., and Hales, T. C. (1998). A formulation of the Kepler conjecture. Preprint[1], `arXiv:math.MG/9811072`.

Gruber, P. M., and Lekkerkerker, C. G. (1987). "Geometry of Numbers," Elsevier, North-Holland, Amsterdam.

Hales, T. C. (1997). "Sphere packings I," *Discrete Comput. Geom.* **17,** 1–51.

Hales, T. C. (1997). "Sphere packings II," *Discrete Comput. Geom.* **17,** 1–51.

Hales, T. C. (1998). An overview of the Kepler conjecture. Preprint[1], `arXiv:math.MG/9811071`.

Hales, T. C. (1998). Sphere packings III. Preprint,[1] `arXiv:math.MG/9811075`.

Hales, T. C. (1998). Sphere packings IV. Preprint,[1] `arXiv:math.MG/9811076`.

Hales, T. C. (1998). The Kepler conjecture. Preprint,[1] `arXiv:math.MG/9811078`.

Hales, T. C. (2000). "Cannonballs and honeycombs," *Notices of the American Mathematical Society* **47**(4), 440–449.

Hales, T. C., and McLaughlin, S. (1998). A proof of the dodecahedral conjecture. Preprint[1], `arXiv:math.MG/9811079`.

Pach, J., and Agarwal, P. K. (1995). "Combinatorial Geometry," Wiley, New York.

Zong, Ch. (1999). "Sphere Packings," Springer, Berlin/Heidelberg/ New York.

[1]arXiv electronic preprints can be retrieved from the web site `http://arxiv.org/` or its front `http://front.math.ucdavis.edu`, as well as its several mirror sites.



Statistical Robustness

Bruce M. Hill
University of Michigan

I. Introduction
II. Robustness to the Prior Distribution
III. Precise Measurement
IV. Fiducial Argument
V. Procedural Robustness
VI. Robustness to Likelihood Function
VII. Examples of Nonrobustness
VIII. Conclusions

GLOSSARY

Bayesian hypothesis testing Theory initiated by H. Jeffreys and developed by L. J. Savage, in which the posterior probability for a hypothesis is obtained by integrating out parameters specified by the hypothesis, using an *a priori* distribution for such parameters to weight the likelihood function.

Exchangeable Probability distribution of a finite sequence $X_1, \ldots, X_n$ of random variables is said to be exchangeable if the joint distribution of the variables is invariant under permutation, that is, all permutations of the variables have the same probability distribution.

Fiducial probability Form of post-data probability evaluation devised by R. A. Fisher, in which the observed numerical values of the data do not alter the pre-data probability distribution for certain random quantities.

Permutation Any of the sequences into which a finite set of elements can be ordered.

Post-data robustness A property of a terminal statistical decision in which the post-data expected utility obtained under *reasonable* variation in the *a priori* distribution, likelihood function, or utility function differs from the post-data expected utility of the decision by only a small percentage.

Precise measurement or stable estimation Principle formulated by L. J. Savage giving conditions under which the posterior distribution based upon a likelihood function that is sharply concentrated relative to the *a priori* distribution can be approximated by the likelihood function standardized to have unit area.

Principle of insufficient reason Principle used by T. Bayes and P. Laplace that asserts that when there are no known reasons to differentiate among a finite set of possibilities with respect to their chance of occurrence, it *may* then be appropriate to give each possibility equal probability.

Probability distribution Assignment of a unit of probability mass to various possible alternative hypotheses or outcomes.

Procedural robustness Property of a statistical or decision-making procedure in which the pre-data

expected utility obtained under *reasonable* variation in the *a priori* distribution, likelihood function, or utility function differs from the pre-data expected utility of the procedure by only a small percentage.

Random An outcome is said to be random if it is not known to be deterministic and is the result of a process that is predictable at best only in a statistical sense. Such outcomes are said to be due to chance.

I. INTRODUCTION

The word *robust* means strong, hardy, healthy. In statistics it is used to suggest that some probability evaluation or property of a decision procedure is not very sensitive to alteration of the conditions under which the evaluation is made, so that the conclusions remain essentially correct even though the precise conditions for those conclusions to hold are not satisfied. It is implicit that the evaluation or procedure is in some sense appropriate for the circumstances under consideration, since otherwise some very poor but insensitive procedures would be robust. In this article, we shall trace some important ideas concerning robustness. We shall consider robustness with respect to the prior distribution and the likelihood function, both in the post-data decision making sense and in the pre-data procedural sense. Robustness is a property which may or may not exist in any particular situation, and one may speak more generally of sensitivity analysis. Thus, if in important respects the conclusions of a statistical or decision analysis are not very sensitive to *reasonable variations* of the prior distribution, likelihood function, and utility function, then we call the analysis robust. To be more precise, this will be the case if the variation in expected utility from an action or a procedure, due to reasonable variation of the input from that of some specific prior distribution, likelihood function, or utility function, is only a small percentage of the expected utility for that specific input. Global robustness almost never holds when there is randomness, since one can always find some prior distribution, likelihood function, or utility function for which the conclusions are substantially different from those of any specified procedure. Hence it must be understood that robustness concerns only limited and reasonable variation in the input. In practice, when such limited robustness occurs it is often due to substantial common information and values, and even this is relatively rare in many important real world problems of inference and decision making.

II. ROBUSTNESS TO THE PRIOR DISTRIBUTION

Historically, one of the first questions considered concerned robustness with respect to the choice of the *a priori* probability distribution for a parameter. Probability evaluations are made either through theory, past observed frequencies, judgment, or some combination of these. This is true for both the subjective and frequentist views of probability. One of the most important ways of making such evaluations is through the principle of insufficient reason, as implicitly used by Bayes, and formulated and used by Laplace. This principle was the basis for Laplace's use of uniform distributions to evaluate probabilities by counting or combinatorical analysis. To illustrate, suppose that there are K possible values of some parameter and there is no *known* reason to distinguish one value from another with regard to its likelihood of truth. Then the principle of insufficient reason would suggest taking the probability of each possible value to be $1/K$. In the subjective approach to probability of de Finetti (1937, 1974, 1975) and Savage (1972) this is particularly natural, since if one were to be awarded a prize subject to the truth of a particular value, then one would not be able meaningfully to distinguish between the various choices, and so would be indifferent as to which value was chosen. It must be understood that this indifference is not meant as a logical rule, that is, a rule asserting that one *must* or even that one *should* be indifferent, for this is patently false. Indeed, since any two different objects are logically distinct, it is absurd to assert such indifference as a logical rule. The sun may or may not appear to rise tomorrow, but most people do not take the probability to be $1/2$. All probability evaluations are conditional on the state of information or knowledge that obtains when they are made, so that in examples such as this no sensible person would be indifferent. On the other hand, when the state of information regarding a finite number of possibilities is subjectively regarded as the same, and if one must make a choice between them, then it is difficult to see any real alternative to the principle of insufficient reason.

Bayes motivated this choice by observing that with respect to independent trials resulting in either success or failure, with the same chance on each trial, that is, a Bernoulli sequence, and where one knows little about the value of this chance, it would be reasonable to regard all possible numbers of successes in n trials as equally likely. This is a compelling evaluation, since if one were instead to take some number of successes as more likely than others, then one must know something particular to differentiate the one number or proportion from the others. It has been proven that such a subjective judgment, for each n, is essentially equivalent to giving the conventional parameter P, representing the unknown chance of a success, a uniform *a priori* distribution, that is, a Beta distribution with both parameters equal to unity.

Suppose now that one makes this evaluation, and asks whether it is robust to variations in the *a priori* distribution.

For example, consider an urn which contains only red or white balls, a known number N in all, but with an unknown proportion of red balls. Let success (1) correspond to the draw of a red ball, and failure (0) correspond to the draw of a white ball. Let H_i denote the hypothesis that there are exactly i red balls in the urn, $i = 0, \ldots, N$. An urn is picked at random, that is, with equal probability for each urn, and then a simple random sample of size n is drawn from that urn. On the basis of this sample we want to draw inference as to which urn was drawn from, that is, the proportion of red balls in that urn, and also to predict the color of the next ball to be drawn from that urn. With regard to the probabilistic prediction of the color of the next ball, this example is a special case of the general problem of inductive reasoning, as formulated by David Hume. If the *a priori* probability for each H_i is $1/(N+1)$, then it is straightforward to make the evaluation, which is equivalent to that made by Bayes, and, as $N \to \infty$, corresponds to that based upon the uniform *a priori* distribution for P. See Feller (1968, p. 123). Now, however, let us ask what happens if the probability for the hypothesis H_i is p_i instead of $1/(N+1)$. To what extent will the evaluation that we have made be robust to such variation in the *a priori* probability distribution?

There is an extremely simple method of analysis due to de Finetti (1975, p. 221). A finite sequence $X_1, \ldots, X_N$ of random variables is said to be exchangeable if the joint distributions are invariant under permutations. In other words, it is exchangeable if for each $k < N$, and each subset of (distinct) indices $j_1, \ldots, j_k$, the distribution of $X_{j_i}, \ldots, X_{j_k}$ is the same as that for $X_1, \ldots, X_k$. An infinite sequence is said to be exchangeable if each finite subsequence is exchangeable. Exchangeability is a condition expressing the judgment that order is completely irrelevant for probabilistic evaluations. This condition can itself be viewed as a consequence of the principle of insufficient reason, where the finite collection of possibilities about which one is indifferent consists of the $n!$ permutations of a given sequence of n outcomes, and one cannot differentiate one such permutation from another with respect to its chance of occurring.

Mathematically, exchangeability is a far-reaching generalization of the concept of independent identically distributed random variables. According to a celebrated theorem of de Finetti, the class of infinite exchangeable sequences consists of all mixtures of independent identically distributed sequences of random variables. See de Finetti (1937; 1975, Chapter 11), Feller (1971, p. 228), Savage (1972, p. 50). In the finite case, the class of exchangeable distributions for 0–1-valued data can be represented in terms of sampling without replacement from urns such as described above.[1] See de Finetti (1975, p. 217) and Heath and Sudderth (1976).

Consider the (potentially) infinite exchangeable sequence of 0–1-valued variables X_j, where X_j takes on the value 1 if there is a success on trial j, and is 0 otherwise. Suppose that our probability distribution for the sequence is exchangeable, and that we have observed the data $X_j = x_j$ for $j = 1, \ldots, n$, where each x_j is either 1 or 0. Denote by $\omega_r^{(n)}$ the probability that there will be r successes in the first n trials. Because of exchangeability it follows immediately that

$$\omega_r^{(n)} = \binom{n}{r} \times P(X_1 = \cdots X_r = 1, X_{r+1} = \cdots X_n = 0). \tag{1}$$

Now let us consider the evaluation of $p_r^{(n)}$, which is by definition the conditional probability that the next observation is a success, given that there were r successes in the first n trials. For the case considered by Bayes and Laplace, where P is given the uniform *a priori* distribution, it is an elementary computation that $p_r^{(n)} = (r+1)/(n+2)$. We wish to examine the extent to which this computation is robust to the *a priori* distribution for the number of successes in n trials. It follows from the definitions of exchangeability and of conditional probability that $p_r^{(n)}$ does not depend upon which r of the values x_i are 1. So let $\mathcal{D}$ represent the data in which the first r trials yield success and the next $n - r$ trials yield failures. Then

$$\begin{aligned} P(X_{n+1} = 1 \mid \mathcal{D}) &= \frac{P(X_1 = \cdots X_r = X_{n+1} = 1, X_{r+1} = \cdots X_n = 0)}{P(\mathcal{D})} \\ &= \frac{\omega_{r+1}^{(n+1)} \times \binom{n}{r}}{\omega_r^{(n)} \times \binom{n+1}{r+1}} \\ &= \frac{[(r+1)/(n+2)]}{1 + [1 - (r+1)/(n+2)](a_{r,n} - 1)}, \end{aligned} \tag{2}$$

where $a_{r,n} = \omega_r^{(n+1)}/\omega_{r+1}^{(n+1)}$. In this evaluation, which is straightforward, apart from exchangeability it has been assumed only that $P(\mathcal{D}) > 0$.

This simple formula exhibits the robustness properties of the Bayes–Laplace evaluation. For any exchangeable distribution on the X_j such that $a_{r,n}$ is close to unity, the posterior probability that the next trial is a success

[1] For any such exchangeable sequence of length n, one must wind up with some number of successes, say i, between 0 and n. Under exchangeability all ordered sequences (or paths) leading to the same number i of successes are equally likely, and this is also true for sampling from an urn with exactly i red balls. It follows that if the probability p_i for sampling from the ith urn (which has i red balls) is taken equal to the probability of i successes in n trials for the given exchangeable sequence, then the distribution of observations from such urn sampling is the same as that for the given exchangeable sequence.

is necessarily close to $(r+1)/(n+2)$, and it is exactly equal to this value when $a_{r,n}=1$. But the condition $a_{r,n}\approx 1$ requires, from the subjective viewpoint, only the inability to distinguish sharply between the occurrence of r and $r+1$ successes in $n+1$ trials, and would often hold in practical applications, especially when n is large. Furthermore, for any fixed *a priori* distribution π for p, one can easily examine the extent to which this evaluation is robust, or insensitive to the choice of π. One obtains the classical Bayes–Laplace formula $(r+1)/(n+2)$, to a very good approximation, for any exchangeable distribution subject to the approximate truth of the condition. In this example, we have obtained a clear understanding of the nature and degree of robustness to the choice of the prior distribution for the parameter P of a Bernoulli process. Equation (2), when $r=n$ and $a_{r,n}=1$, is known as Laplace's rule of succession, which was put forth as a predictive probability for the next trial giving rise to a success when all previous trials have resulted in success and little is known beforehand about the chance of success. It is a completely valid and robust evaluation for the circumstances in which it was proposed, and agrees with common sense far more than does the maximum-likelihood estimate r/n. For example, if there has been one success in one trial, the Bayes–Laplace predictive probability that the next trial is a success is $2/3$, while the maximum-likelihood estimate is the unreasonably large value 1. Historically, there has been controversy about some applications of this formula, perhaps based upon misunderstandings regarding the intentions of Laplace.

III. PRECISE MEASUREMENT

The robustness to the prior distribution that we have exhibited in the special case of a Bernoulli process occurs in many problems in statistics. The general principle was formulated by Savage (1962) under the name *precise measurement* and again by Edwards *et al.* (1963) under the name *stable estimation*. The general description is as follows. Let Θ be a parameter (scalar or vector) for a family of probability distributions, with probability densities $f(x;\theta)$, say, relative to Lebesgue measure.[2] In other words, the parameter Θ identifies a particular probability distribution for the data. It is assumed that there is a true value for the parameter, say θ_0. We observe a sequence of random variables $X_1,\ldots,X_n$ which are conditionally independent, given θ_0, each having density $f(x;\theta_0)$. For a particular observed sequence of observations, say $X_i=x_i$ for $i=1,\ldots,n$, the realized likelihood function is the function

$$l(\theta)=\prod_{i=1}^{n} f(x_i;\theta).$$

According to all accepted statistical theory, given the truth of the model, this function carries all the information in the data and is a technically a sufficent statistic.

Under very weak conditions that often hold in practical applications, one can show that if the sample size n is sufficiently large, then it is probable that there will exist a neighborhood $\mathcal{N}$ of θ_0 in which the realized likelihood function will be unimodal with most of its mass concentrated in that neighborhood. In fact, often $l(\theta)$ can be approximated by a multivariate normal density function with center at $\hat{\theta}$, where $\hat{\theta}$ is the maximum-likelihood estimate of θ, that is, the value of the parameter that maximizes $l(\theta)$. The covariance matrix for this approximation to the likelihood function can be written as A/n, where A is a function of the observed data. For example, A is often obtained by using a quadratic approximation to $\log l(\theta)$ near $\hat{\theta}$.

Now suppose that the *a priori* density $\pi(\theta)$ for the parameter Θ is diffuse relative to $l(\theta)$, that is, $l(\theta)$ is much more concentrated than is $\pi(\theta)$, and that $\pi(\theta)$ is nearly constant in the neighborhood $\mathcal{N}$ of $\hat{\theta}$ in which almost all of the mass under the likelihood function is contained. Then it is clear that conditional upon Θ being in $\mathcal{N}$, the posterior distribution for Θ is essentially given by the likelihood function standardized to have total area 1 in $\mathcal{N}$. Furthermore, if $\pi(\theta)$ is nowhere too large outside $\mathcal{N}$, then the posterior mass to be attached to the complement of $\mathcal{N}$ will be negligible, so that the posterior distribution for Θ is well approximated by the conditional posterior distribution for Θ, given $\Theta\in\mathcal{N}$, which is essentially just the likelihood function standardized to have unit volume in $\mathcal{N}$, and so also is often well approximated by a normal density with center at $\hat{\theta}$ and covariance matrix A/n.

The collection of such results makes up the precise measurement principle of L. J. Savage. Details concerning the approximation are given in Savage (1962) and Edwards *et al.* (1963). This method of analysis constitutes one of the primary means by which robustness of the posterior distribution can be obtained, that is, substantial consensus as to the truth concerning the value of Θ, even when there are quite diverse initial opinions π among scientists. The argument does not always hold, since for individuals with relatively concentrated *a priori* distributions, the posterior mass outside the neighborhood $\mathcal{N}$ favored by the data can be large, or the *a priori* distribution may not be nearly uniformly distributed even within $\mathcal{N}$, or the integral of $l(\theta)$ may be infinite. But for large n the argument is often very useful, and sometimes even for small n. One may

[2] Measure theory is not essential here. The probability densities can be viewed as approximations to discrete mass distributions, and the integrals as approximations to the finite sums that arise in real-world scientific and decision-making problems.

view the situation as follows. The *a priori* distributions π for different individuals vary for many reasons, such as past experiences and exposure to data, different theoretical knowledge, genetic reasons, and so on. But the data of the experiment are common to all individuals and tend to bring their opinions closer together. When there is sufficiently such data the opinions of different scientists and/or decision makers tend to merge.

IV. FIDUCIAL ARGUMENT

If one assumes an infinite sequence of Bernoulli trials, Bayes and then Laplace would have in effect taken the *a priori* distribution for the Bernoulli parameter P to be uniform on the interval [0, 1]. An objection that was raised with regard to this choice was that if the parameter were transformed by a monotonic transformation, say to $\Theta = \log[P/(1-P)]$, then it would be different to take a uniform *a priori* distribution for P than to do so for Θ, and that this could lead to somewhat different results, at least for small samples, since uniformity for Θ is formally equivalent to taking the density for P to be $\pi(p) = 1/[p(1-p)]$. Similarly, if the parameter were $\log P$, then this would mean that the implicit density for P was $1/p$.[3] It was argued that each such choice for parameter was equally valid with the choice of P, that uniformity for the *a priori* distribution of each such parameter was equally valid with uniformity for P, and that therefore there was nothing special about taking the uniform distribution for P, as had been done by Bayes and Laplace.

Such arguments entirely ignore the motivation by Bayes and Laplace to avoid bias in predicting the number of successes in n trials, as well as robustness arguments such as that of de Finetti given above. However, they opened the door for the statistician R. A. Fisher initially to reject the Bayesian approach, in effect because the prior distribution for P could be chosen in many ways, with the implicit notion that the posterior distribution would not be robust with respect to the various meaningful choices for the *a priori* density of P. See Fisher (1959, p. 16).

When the sample size n is not large he was correct that in many problems of statistical inference the conditional distribution of the parameter, given the data, can be quite sensitive to the choice of the *a priori* distribution. As a means of avoiding such issues regarding the choice of the *a priori* distribution, Fisher then suggested his fiducial approach as an alternative to the method of Bayes. Initially he was under the impression that it was a totally different approach, which allowed one to go from a state of total ignorance about a parameter to a state of probabilistic knowledge concerning the parameter, based upon observational data. To illustrate the fiducial method, consider the important case of an observable variable $X \sim N(\mu, 1)$, where this notation means that, given μ, the distribution of X is the normal distribution with mean μ and variance 1. Before observing the value of X, one is therefore of the opinion that, given μ, the random quantity $X - \mu \sim N(0, 1)$. The fiducial argument asserts that after observing $X = x$, it is then the case that $x - \mu \sim N(0, 1)$, and hence also $\mu \sim N(x, 1)$, where x is the observed numerical value, and now μ is the unknown, given a probability distribution.[4] Fisher here used the older concept of probability of Bayes and Laplace, where probability is used to express a state of knowledge about an unknown quantity, which need not be random in the conventional sense, but merely unknown. Fisher never succeeded in supplying a rational argument for his evaluation but insisted that it was both plausible and logical in some unspecified sense. He thus appeared to have obtained a post-data probability distribution for the parameter without having input a prior distribution, and was therefore seemingly not subject to the various criticisms with respect to robustness of the Bayesian approach. However, precisely the same result is obtained if one uses the Bayesian approach with the *a priori* distribution for μ taken as Lebesgue measure interpreted as an improper probability distribution. Furthermore, similar relationships between the fiducial method and the Bayesian method with improper prior distributions hold in many important examples of statistical inference.

Late in his life Fisher (1959, p. 51) acknowledged:

> Probability statements derived by arguments of the fiducial type have often been called statements of 'fiducial probability'. This usage is a convenient one, so long as it is recognized that the concept of probability involved is entirely identical with the classical probability of the early writers, such as Bayes. It is only the mode of derivation which was unknown to them.

The argument that fiducial probability was in essence nothing more than a surreptitious Bayesian posterior evaluation had in fact been given by Jeffreys to Fisher in debates at the *Royal Statistical Society*. The quotation suggests that Fisher eventually consented to the arguments of Jeffreys. Fiducial analysis, when it could be justified at all, was equivalent to a special case of the Bayesian analysis with a uniform *a priori* distribution for the appropriate

[3] These last two densities are called improper because the integral over the domain is infinite. For the first, the predictive probability that the next trial is a success, given r successes in n trials, is the same as the maximum-likelihood estimate, r/n. For the second, it is $r/(n+1)$. For large n, all of the commonly recommended prior distributions for P lead approximately to r/n.

[4] A random quantity such as $X - \mu$, whose distribution remains the same after X is replaced by its observed value x, is sometimes called a pivotal quantity.

parameter. It is straightforward to prove that this is true for problems concerning location and scale parameters. Apart from this Bayesian interpretation, as an alternative method for obtaining the posterior distribution, the fiducial argument has not received much support in statistical theory.

Furthermore, an example of Edwards (1972, p. 207) shows that this argument is highly suspect as a logical argument, insofar as, if accepted, it would provide a logically valid proof for the principle of insufficient reason. But, as argued above, this principle is absurd as a *logical* rule. By *reductio ad absurdum*, the fiducial argument can then be rejected as a logical argument. Here is Edwards's example. Let there be two hypotheses concerning a number H, that it may have value $+1$ or -1. We assume in the Fisherian sense that there is no *a priori* information about the value of H. Let X be an observable random variable, also with the same two possible values. Let the probability model for the data be such that $P(X=+1 \mid H=+1)=p$ and $P(X=+1 \mid H=-1)=q$, where $p+q=1$. The probability distribution for the random quantity $Y=HX$ does not depend upon the value of H. Thus $HX=+1$ if and only if either both quantities are $+1$ or both are -1, and so

$$P(Y=+1 \mid H=+1)=P(Y=+1 \mid H=-1)=p.$$

This is a standard requirement in fiducial theory, as when the distribution of the quantity $X-\mu$ does not depend upon μ in the normality example. Now let $p=q=1/2$. In this case, the observation of X is totally uninformative with respect to the value of H. As Edwards says, "it will be universally agreed that the observation of X is so utterly uninformative about H that there is no point in even making it."

But if we apply the fiducial argument, under which we can substitute for the random variable X its observed value, say $X=+1$, it follows that the post-data or fiducial probability for the event $H=+1$ must be $1/2$. Edwards, arguing that such a conclusion cannot be valid, states "Thus, starting with absolutely no information about H, and conducting an absolutely uninformative experiment, we end up with a definite statement of probability."

There can be no more decisive demonstration than this that the fiducial argument, as a logical argument, is invalid. One can perhaps make technical objections to the example, but it seems clear that the very essence of the fiducial argument, as a logical argument, has been destroyed.[5] Of course, the example only disproves the fiducial argument as a logical argument, not as an openly *subjective* evaluation equivalent to a form of the principle of insufficient reason and only to be used in those circumstances where it is judged appropriate. In Edwards's example the Bayesian analysis would consist in taking the prior probability for each hypothesis to be .5, and then, since the experiment is totally uninformative, the posterior probability is of course also .5. Here the principle of insufficient reason *might* be employed to obtain the prior probability, whereas the fiducial argument would purport not to use the principle of insufficient reason at all, but rather to obtain the same post-data probability of .5 on the grounds of having absolutely no knowledge *a priori* as to which hypothesis was true. Just as in the examples of Jeffreys, there was an implicit *a priori* distribution involved in the use of the fiducial argument.

Despite this criticism of the fiducial argument as a logical argument, there is a sense in which it was an innovative and valuable contribution to statistical inference. To illustrate, one of the most important examples of the fiducial approach concerns the nonparametric predictive procedure discussed by 'Student' (W. S. Gossett) and Fisher, which the present author later restated and called A_n.[6] The procedure A_n asserts, for an exchangeable sequence of observable random variables in which ties have probability 0, that conditional upon the data $X_1, \ldots, X_n$, the next observation X_{n+1} is equally likely to fall in any of the open intervals between successive order statistics of the given data (Hill, 1968, p. 677). Note that unconditionally upon the data (or equivalently if the X_i for $1 \le i \le n$ were not observed) this would be immediate, since under exchangeability all orderings of the data are equally likely. Therefore X_{n+1} is unconditionally equally likely to be the largest or the smallest observation or have any other rank among the $n+1$ observations.

In this problem there is a high degree of sensitivity to the choice of the *a priori* distribution for the parameter, which is taken as the unknown distribution function in the fully nonparametric case. One can of course input a proper *a priori* distribution for the distribution function and mechanically perform a Bayesian analysis; but different reasonable choices for such an *a priori* distribution can lead to *very different* posterior distributions, and no single choice in which the conditional distribution of X_{n+1} depends on the data has been convincing. On the other hand, the unconditional probability that X_{n+1} is in each of the open intervals between consecutive order statistics is exactly $1/(n+1)$ since this probability is the same as the probability that X_{n+1} has each of the $n+1$ possible ranks, and this probability is objectively determined merely by the fundamental assumption of exchangeability with ties

[5] In the second edition of his text, Edwards indicated that he was not entirely satisfied with his own argument but gave no reasons. He may have had in mind some minor technical issues with respect to the precise definition of a pivotal quantity.

[6] See Aitchison and Dunsmore (1975) for a general discussion of predictive inference.

having probability 0. Hence this is a perfect example for the Fisherian criticism of the Bayesian approach since the unconditional probabilities are well determined, while the *a priori* distribution is highly subjective, and the conditional probabilities are quite sensitive to the choice of the *a priori* distribution.

Here is how the fiducial argument would apply. Consider a conventional formulation of statistical inference, in which the observations are conditionally independent, given Φ, with cumulative distribution function $F(x;\phi)$, where Φ is an unknown parameter. Assume that the distribution function is continuous in x for each ϕ, but is otherwise unknown. Let $X_{(i)}$ denote the ascending order statistics of the data for $i=1,\ldots,n$. Then let $\theta_i = F(X_{(i)};\phi) - F(X_{(i-1)};\phi)$ for $i=1,\ldots,n+1$, where by definition $X_{(0)}=-\infty$ and $X_{(n+1)}=\infty$. Before the data are drawn, the distribution of the θ_i is a uniform distribution on the n-dimensional simplex, that is, a special Dirichlet distribution in which all the parameters are equal to unity.[7] This is the fundamental frequentistic intuition with regard to A_n, which Fisher presumably used to put forth his proposed fiducial solution. Then Fisher (1939, 1948) applied the fiducial argument to suggest that even when the random variables $X_{(i)}$ are replaced by their observed values $x_{(i)}$, the uniform distribution for the Θ_i would still be appropriate. Fisher seemingly was unaware of the concept of exchangeability, or the close connection of this problem with the original example of Bayes, and did not address any of the criticisms or issues raised by the fiducial argument in this example.

Fisher's proposed fiducial distribution is an example of a posterior predictive distribution since whatever the rationale, it is posterior to the data and does partially specify a probability distribution for the future data. This predictive distribution is not completely specified since (in the case where ties have probability 0) what it does is to attach a probability of $1/(n+1)$ to each of the $n+1$ open intervals formed by the consecutive order statistics of the given sample and goes no further. The fiducial argument that Fisher gave for this evaluation depends upon one's willingness to persist with the pre-data evaluation of the distribution of $\theta_i = F(X_{(i)};\phi) - F(X_{(i-1)};\phi)$ *after* the $X_{(i)}$ are replaced by their observed numerical values. The question then arises as to whether there is a Bayesian analysis for which the same evaluation is also appropriate conditional upon the data for all or almost all such data, as in our discussion of other examples of the fiducial argument.

At first sight, there does not seem to be any implicit *a priori* distribution to justify this, although the conclusion is plausible when little is known about the population. However, Hill (1968, 1993) showed that A_n has both exact and approximate Bayesian justifications in the case of diffuse *a priori* knowledge about the shape of a finite population, and proposed A_n as a robust Bayesian procedure. It is exactly valid for merely simply ordered data, or for data measured on a 'rubbery scale,' as when the finite population values are distorted by some measuring instrument that is equivalent to making an unknown monotonic increasing transformation of the original values. For large n, A_n tends to agree with the empirical distribution function insofar as intervals containing a substantial number of order statistics are given very nearly the same predictive probability. Both the empirical distribution function and A_n are primary examples of robust statistical analyses, neither depending on any assumed special *a priori* information as to the shape of the finite population. A_n is the extension of the notion of a robust *a priori* distribution for a conventional parameter to the case where the parameter is the entire distribution function.

The A_n example is at the heart of some important issues concerning statistical inference. First, the fundamental objectivistic frequentist argument would support A_n on the basis that the evaluation is correct unconditionally, so that in repeated applications of the procedure the empirical frequencies would agree with the evaluation. For example, X_{n+1} would be the largest of the $n+1$ observations with a relative frequency of about $1/(n+1)$. Although one can input a proper *a priori* distribution and obtain a standard Bayesian updating for the conditional probabilities, rather than merely use the unconditional probabilities, the complexity of the problem and the sensitivity to the prior distribution dictate against doing so unless a convincing case can be made for one out of the myriad of such choices for the *a priori* distribution. This is not an argument against Bayesian updating in general, but such updating requires the greatest care in a nonrobust scenario such as this and should not be done merely for mathematical convenience.

Next, although this frequentist argument supports A_n, Hill (1968) proved for each $n \geq 1$ that there are no countably additive distributions on the observations for which the A_n evaluation holds almost surely, that is, for almost all $x_1,\ldots,x_n$. Thus either frequentists must abandon their fundamental argument regarding sensitivity and complexity or else abandon the special role of countable additivity in their theory. Furthermore, Hill (1993) showed that there do exist finitely additive processes, called splitting

[7]The Dirichlet distribution is the multivariate generalization of the Beta distribution, and the uniform Dirichlet distribution is the multivariate generalization of the uniform distribution on an interval. See DeGroot (1970, p. 49). A_n, which is a special case of H_n of Hill (1968, 1993), can be viewed as the generalization of the Bayes–Laplace analysis of binary data to the multinomial case, with an *unknown* number of categories or types. A_n is related to the Dirichlet process of Ferguson (1973). H_n allows for ties and smoothing in the framework of A_n. Both A_n and H_n are known to be coherent in the sense of de Finetti, so that a Dutch book or sure loss is not possible.

processes, for which A_n does hold almost surely, and which are thus in fact compatible with the fundamental frequentist arguments. Thus the use of A_n achieves the frequentist objective long run frequencies and also has the coherency properties of Bayesian decision procedures. Splitting processes are based upon the adherent mass distributions of de Finetti (1971, p. 240), which are in turn closely related to the Dirac functions as used by physicists and engineers. Finitely additive probability distributions provide a mathematically rigorous way to deal with conventional nonfinite measures such as Lebesgue measure.

The fiducial approach, as used in connection with A_n, can again be seen as a method of reasoning to derive a Bayesian posterior distribution by an indirect evaluation rather than in the traditional way. Whatever its defects, it was an important first step in what may be called Bayesian data analysis. See Hill (1990). It put forth a post-data probability distribution for the parameters, or for future observables, that was coherent in the sense of de Finetti, plausible in some circumstances, and robust in various respects. For example, in the case of A_1 with normal data, we have robustness to the *a priori* distribution for μ, in the sense that the fiducial conclusion holds to a good approximation whenever the prior distribution for μ is diffuse relative to the likelihood function, as is made clear by the stable estimation argument of Savage. In many examples the fiducial argument may be regarded as being equivalent to a robust Bayesian approach with a uniform *a priori* distribution for the appropriate parameter. The fiducial approach is oriented both toward scientific inference, which typically takes place in a unique context, and also to the policymaker problem of Hill (1990), which concerns repeated applications of a certain policy.[8] In circumstances where the fiducial argument is appropriate, it leads to a relatively uncontroversial and robust analysis, as is requisite in public policy issues.

V. PROCEDURAL ROBUSTNESS

Post-data robustness concerns the sensitivity to the *a priori* distribution, likelihood function, or utility function, of post-data expected loss due to terminal decisions. It is of particular importance in real-world scientific and decision-making problems, where the data are available and have been analyzed and one wishes to examine to what extent the optimal decision changes when one makes reasonable variations in the input. Procedural robustness, on the other hand concerns the pre-data evaluation of sensitivity to *a priori* distribution, likelihood function, or utility function, of the overall consequences of using a certain procedure, especially when done repeatedly under similar circumstances. In this context there is ordinarily no data analysis, and all probability distributions are prespecified before obtaining the data (see Berger, 1984, Section 3; Hill, 1990). Procedural robustness is of particular importance in assessing the value of a policy or strategy to be used repeatedly, as in bureaucratic organizations.

In connection with the controversy over the Behrens–Fisher distribution, it was discovered that the Fisherian fiducial method did not always agree with the confidence approach that became part of the Neyman–Pearson theory of statistics. See Fisher (1959, p. 95). The Neyman–Pearson departure from the Fisherian approach emphasized the behavioral properties of a procedure rather than scientific or inductive inference, and was oriented toward what may be called the policymaker problem. This theory is primarily concerned with the consequences of repeated applications of a procedure. For example, in the case of normally distributed data considered above, Neyman would argue that the interval $X \pm 1.96$ is a 95% confidence interval in the sense that under repeated application of the method, about 95% of the intervals would include the true μ. In this approach one is not concerned with individual instances of the truth or falsity of the assertions, but rather with the overall long-run consequences that follow from use of the procedure. Fuller understanding of the difference between the concepts of Neyman and Pearson and those of Fisher emerged from the work of A. Wald, who was working in the spirit of Neyman and Pearson. Based in part on earlier work of J. von Neumann dealing with game theory, he proposed a theory of admissible decision functions, as an attempt to rule out those procedures that were objectively defective. See Wald (1950).

In the conventional formulation of a statistical decision problem, there is a space of actions $\mathcal{A}$, a parameter space Θ, and observable data represented by a random variable X. A decision function d is a function that for each possible data value x specifies an action $a \in \mathcal{A}$ to be taken. The action space represents the various possible final or terminal actions that are contemplated and from among which one wishes to choose. The parameter space represents the most important aspects of reality about which one is uncertain that are believed to be relevant to the decision problem. There is also a utility function $U(\theta, a)$,[9] which represents the value to the decision maker of choosing act a when Θ is the true value of the parameter. If the true value of Θ were known, then one would need no data. One would

[8] In considering the greenhouse effect or the destruction of the environment, it is not natural (or desirable) to consider hypothetical repetitions of the overall experiment. One wants, instead, to understand the particular phenomenon in question, and to do something about it.

[9] Or alternatively, there is a loss function, $L(\theta, a)$, often taken as the negative of utility.

simply choose, for that Θ, the act that maximizes utility, $U(\theta, a)$, or minimizes loss, $L(\theta, a)$. When Θ is unknown, however, one may still have to choose an action. To help in such a choice, one can first observe relevant data, which reduces the degree of uncertainty about the value of Θ.

Wald's theory is closely related to the Bayesian theory, but not identical to it. In the Bayesian theory, at any state of knowledge about Θ, if an action must be taken, then an optimal action is to choose the value a that maximizes the expected utility, that is, that maximizes

$$\mathbf{U}(a) = EU(\theta, a),$$

where the expectation is taken with respect to the current probability distribution for Θ. Suppose the initial state of knowledge is represented by the distribution with density $\pi(\theta)$. In the Bayesian theory, after observing data $X = x$, the *a priori* distribution π is updated to become the posterior distribution π^* according to Bayes's theorem. Then one maximizes posterior expected utility by choosing an action a that maximizes $\mathbf{U}(a)$ with respect to π^*. In the decision theory of Wald one also allows the possibility that there is no prior distribution π, or if there is, it is unknown, and not to be specified subjectively. In this case one can define the concept of admissibility of a decision function without reference to the *a priori* distribution. Suppose that $L(\theta, a)$ is the loss function, and let $R(\theta) = E_{X|\theta} L(\theta, d(X))$ be the expected loss if decision function $d(X)$ is used when θ is the value of the parameter, that is, the usual risk function. We say that a decision function d_2 is inadmissible if there exists another decision function d_1 such that $R_1(\theta) \leq R_2(\theta)$ for all θ, with strict inequality for some θ. If a decision function is not inadmissible, then it is said to be admissible. One can reason as follows that inadmissible decision functions should be eliminated from consideration whenever possible.

Let Θ be a conventional parameter that determines the distribution of the random variable X and let $d_i(X)$, $i = 1, 2$, be two decision functions that depend on the value of X. Suppose that we are in a situation that is repetitive and that $L(\theta, d_i(X_j))$ is the loss to be sustained if d_i is used on the jth occasion, $j = 1, \ldots, M$. Then the expected losses, given the value of Θ, are

$$R_i(\theta) = E_{X|\theta} L(\theta, d_i(X)), \qquad i = 1, 2,$$

which are the risk functions. Here X is a generic random variable having the same distribution as each X_j. In general in this framework of repetitive situations there is the possibility of learning from one occasion to another, as for example, via data analysis. However, in some circumstances the decision functions will be mechanically implemented on a computer without the statistician or decision maker actually observing the X_j, so that no learning can take place from one occasion to another. In this case the theory is a pre-data theory of optimality, that is, before observing the data one desires an optimal policy to adopt for the given framework of repetitive decision making. We shall assume this is the case in the present discussion.

Suppose now that there is a referee who generates couples (θ_j, X_j) on a computer for $j = 1, \ldots, M$, using a probability distribution π to generate the θ_j, and some conditional distribution for X_j, given θ_j. Let the referee generate M independent couples in this way. Assume that the conditional distribution for X_j, given θ_j, is known to all concerned, but not π. In this case the loss associated with use of d_i on the jth occasion is $L(\theta_j, d_i(X_j))$. Summing over the M occasions, the actual increment in loss if d_2 were used on each occasion instead of d_1 would be $\sum_{j=1}^{M} [L(\theta_j, d_2(X_j)) - L(\theta_j, d_1(X_j))]$. The expectation of this increment, from the perspective of the referee who knows π, is then $K \times M$, where K is the π expectation of $K(\theta) = E_{X|\theta}[L(\theta, d_2(X)) - L(\theta, d_1(X))] = R_2(\theta) - R_1(\theta)$. If d_2 is dominated by d_1 in the sense of admissibility, then $R_2(\theta) - R_1(\theta) \geq 0$, with the inequality strict for some θ. So if π gives positive weight to the set of θ where the inequality is strict, then $K > 0$. This evaluation is mathematically valid for loss functions bounded from below.

If π has positive mass where $R_2(\theta) - R_1(\theta) > 0$, then from the perspective of the referee, who knows π, it would be imprudent to use d_2 in preference to d_1 because of laws pertaining to large M (laws of large numbers, central limit theorems) under which the overall loss would be regarded as likely to be large if M were large. The reference to a referee is made primarily to represent the situation where there is some mechanism which generates what might be called a 'true' distribution for Θ, so that one can assess the performance of the two decision functions from the perspective of such a distinguished distribution. It is not necessary that such a referee exist, but only the mechanism. This is the fundamental frequentist argument as to why inadmissible decision rules should be eliminated from consideration whenever possible in the case of repeated applications of a pre-data procedure. See Hill (1990) for further discussion.

Suppose that there is a referee (or mechanism) and that his distribution is π_0, which is unknown to the decision maker, who must then choose his own π without knowledge of the true π_0. A plausible method is to guess at π_0 and use the Bayes procedure for the guess. Formally, such a guess would be the decision maker's prior expectation of the referee's π_0. Because the Bayes risk is a continuous function of π, it follows that if the guess is close enough to the truth, then the risk for the procedure will be close to the risk of the optimal procedure, that is, the Bayes procedure with respect to π_0. See Blackwell and Girschick (1954, p. 147).

With respect to any statistical decision procedure we can consider robustness with respect to the choice of π and the likelihood function (or model) in the above framework. We first assume that the likelihood function is completely known, and then in the next section consider the case where it is unknown. When the parameter space and data space are finite, it is well known that if a procedure is Bayes with respect to an *a priori* distribution for which all prior probabilities are positive, then it is admissible, while if it is admissible, then it is Bayes (Blackwell and Girschick, 1954, p. 127). It follows that the Bayes procedures for such finite representations form a complete class, in the sense that for any other procedure there is a Bayes procedure that will dominate it. In fact, provided only that the data space and terminal action spaces are finite, it is possible to obtain by routine linear programming methods a Bayes procedure whose risk function is uniformly less than that of the original procedure by some $\epsilon > 0$ (Hill, 1998). Since all real-world statistical analyses are performed on a computer (or the equivalent) with finite memory, the *implementation* of any procedure is necessarily based upon a finite representation for the data, and such linear programming methods can be employed quite generally.

One of the arguments against the use of Bayes procedures concerns their lack of robustness in certain scenarios, such as that of nonparametric statistics. In these cases non-Bayesians have proposed certain relatively robust decision procedures, such as, for example, the empirical distribution function in the case of nonparametric statistics, and more generally the maximum-likelihood estimator. Typically such procedures, although robust, do not have the optimality properties of Bayes decision procedures. However, using the above-mentioned linear programming methods, one can replace such estimators or decision procedures by Bayes procedures that are uniformly better in risk than the original procedure. The resulting procedure is both robust and has the optimality properties of a Bayes procedure. When the original non-Bayes procedure can only be improved upon slightly (although uniformly) there is little reason to do so, but unfortunately many non-Bayes procedures in common use can be greatly improved upon.

VI. ROBUSTNESS TO LIKELIHOOD FUNCTION

A number of robustness studies have compared different procedures for estimation of the parameters of a distribution, for example, comparing a trimmed mean estimator with some standard estimator such as the mean, when the model under which the mean is appropriate is false. A trimmed mean is the average after deletion of a certain proportion of observations in the upper and lower tails of the sample. Suppose that the true likelihood function is a mixture of two normal distributions of the form

$$(1-\epsilon)\times n\left(x;\mu_1,\sigma_1^2\right)+\epsilon\times n\left(x;\mu_2,\sigma_2^2\right), \qquad (3)$$

where $n(x;\mu,\sigma^2)$ is the density of the normal distribution with mean μ and variance σ^2.

Such a distribution is sometimes said to be contaminated by the second component of the mixture. It is clear that in certain circumstances the trimmed mean procedure can be far superior to use of the ordinary sample mean as an estimate of μ_1, which is the mean of the uncontaminated population. This occurs because if μ_2 is very different from μ_1, then the ordinary mean will show a substantial 'bias' since it will be estimating $(1-\epsilon)\mu_1+\epsilon\mu_2$ rather than μ_1. This bias can be greatly reduced by using a trimmed mean since one can thereby eliminate many of the observations from the 'contaminating' population. On the other hand, if either $\epsilon=0$ or μ_1 is very close to μ_2, then one has greater precision by using the ordinary sample mean. Thus, there is the usual tradeoff between bias and precision in a statistical analysis, and which estimator is appropriate depends on what the true distribution is (Hill, 1980).

The basic problem is virtually identical with the policymaker problem of the previous section, except that here one must consider robustness with respect to the likelihood function as well as robustness with respect to the prior distribution. The improvement obtained by using one estimator versus another for repetitive situations is given by

$$\sum_{j=1}^{M}[L(\theta_j,d_2(\underline{X}_j))-L(\theta_j,d_1(\underline{X}_j))].$$

In this formula the d_i are any two specified estimators. For example, d_1 might be the ordinary sample mean and d_2 some trimmed mean. On the jth repetition, a sample $\underline{X}_j$ is taken from a population with true mean θ_j for the uncontaminated population, and we wish to estimate θ_j. The θ_j might themselves be a sample from some underlying population that arises in a production process, for example, θ_j might be a measure of the quality of the jth uncontaminated batch produced by the process. If the model for the jth repetition is the contaminated mixture (3) with $\mu_1=\theta_j$, then, as in the previous section, we can examine the referee's expectation of the above sum, using some prior distribution for the θ_j that reflects empirical knowledge as to how the parameter varies during the production process. In this example, the contamination might be due to a certain proportion of defective items in each batch.

Standard robustness studies often implicitly use a uniform prior distribution for the location parameters and then only consider robustness with respect to the model or

likelihood function. Empirical studies can then demonstrate the degree of improvement by using one procedure versus another. In an industrial control process where a great deal is known about the distributions, for example, about the percentage of observations that are contaminated, great improvements can be obtained by using trimmed means and other such estimators. See Andrews *et al.* (1972) for robustness studies. It would be desirable also to examine robustness to the utility or loss function, which is often of critical importance in real world problems. Although, as before, it is desirable to eliminate inadmissible or non-Bayesian procedures, sometimes the full Bayesian analysis is very difficult to carry out. Thus a full Bayesian analysis would require an *a priori* distribution for ϵ and the parameters, and perhaps even form, of the contaminating distribution. Presumably much is sometimes known from past experience about ϵ, but it might be the case that *a priori* information about the contaminating distribution was not extensive. The practicalities of the situation can then force one into approximating the Bayes procedure by conditioning on only part of the total data. See Hill (1975, 1980) for examples of such robust analyses when the full Bayesian analysis is not feasible. Even in such cases, the above method of comparison can still be used to decide between any two specified procedures.

VII. EXAMPLES OF NONROBUSTNESS

One of the primary examples of nonrobustness arises in the theory of Bayesian testing of hypotheses initiated by Harold Jeffreys and Dorothy Wrinch in the 1920s. This theory appears in Jeffreys (1961) and was developed by Savage *et al.* (1962) and Edwards *et al.* (1963).

Suppose that there are two hypotheses about reality, say H_i, for $i = 1, 2$, and we desire to choose between them. We assume that the two hypotheses form a partition so that exactly one is true. We are given data $\mathcal{D}$ to aid in the choice. Suppose that the hypothesis H_i specifies a probability distribution, or model, for the data. In particular, suppose that given this hypothesis, there is an unknown parameter Θ_i and also a probability distribution π_i for Θ_i. The parameter may be different for the two hypotheses, and the hypotheses need not be nested, that is, one need not be a special case of the other. It follows from the theory of coherency that the essential calculation, relevant to the choice of hypothesis, is that of the Bayes factor, or likelihood ratio in favor of H_1:

$$\mathcal{L} = \frac{\int l_1(\theta_1) \times \pi_1(\theta_1)\, d\theta_1}{\int l_2(\theta_2) \times \pi_2(\theta_2)\, d\theta_2}, \tag{4}$$

where $l_i(\theta_i) = P(\mathcal{D} \mid H_i)$ is the likelihood function under hypothesis H_i. One must multiply the Bayes factor by the prior odds in favor of H_1 to obtain the posterior odds, say $\mathcal{O}$. The posterior probability for hypothesis H_1 is then $\mathcal{O}/(1 + \mathcal{O})$.

In typical examples, the hypothesis H_1 can be taken to be a sharp null hypothesis in the sense that if this hypothesis were known to be true, it would specify a value for the parameter Θ_1, or at least a tightly concentrated distribution π_1. The integral in the numerator of this equation can then generally be evaluated relatively easily. The nonrobustness enters in because under the hypothesis H_2, one would ordinarily use some relatively diffuse distribution π_2 for the parameter Θ_2. As discovered by Jeffreys, it is not possible simply to use a uniform distribution over Euclidean space for π_2 since then one would always accept H_1. Jeffreys suggested the use of a Cauchy distribution for π_2 to obviate this difficulty. In some real world problems of hypothesis testing, however, there is great sensitivity to the precise choice of π_2, and so the evaluation of posterior odds may not be robust. A simple concrete example that arises in forensic science is discussed by Shafer (1982).

Suppose that a crime is committed in which there is broken glass at the scene of the crime. A suspect is picked up, and it turns out that there are shards of glass on his clothes. The refractive index of the glass at the scene of the crime is θ_1, which we take as known. It is supposed that the critical evidence against the suspect is that the measured refractive index of some glass found on his clothes, x, is close to θ_1. It may be observed that it would be common sense that if glass with refractive index θ_1 is quite rare, and if x is sufficiently close to θ_1, then there can be strong evidence that the suspect is guilty. The two hypotheses are the hypothesis H_1, which asserts that the suspect is guilty of the crime, and the hypothesis H_2, which asserts that he is innocent. We will assume that one or the other of these is true, where implicitly there is some precise definition of guilt. If the suspect is guilty, we have in mind that the refractive index of the glass on his clothes, if there should be such glass, is likely to be close to θ_1; if he is innocent, then we would use some distribution, say π_2, to represent our opinions about the refractive index of glass on such a person, without any special reason to anticipate that this index might be close to θ_1. For example, one might take π_2 to be the distribution that one would use if the person were merely sampled at random from some appropriate subpopulation.

In problems of this type it is certainly not appropriate simply to choose a uniform *a priori* distribution under H_2 and mechanically perform a Bayesian analysis. The use of the Bayesian methodology serves to illustrate the basic point discovered by Jeffreys, that the conclusions can be very sensitive to the *a priori* distribution π_2. For example, if we take a sufficiently large interval and use a uniform prior distribution over this interval for the parameter, given the alternative hypothesis, then the data will support

the sharp null hypothesis. It is clear that depending upon the specification of π_2, one can obtain almost any conclusion that one wishes, and that for the analysis to have any real force there must be some reasonable grounds for the choice of π_2, for example, based upon the empirical distribution of refractive indices in some relevant subpopulation. When there are no such grounds, and with such sensitivity to the choice of π_2, no 'objective' solution is possible. *Any approach that pretends to obtain an* '*objective*' *solution in such circumstances is merely disguising the basic uncertainties involved in the problem.* When there are such grounds, then the practical force of the Bayesian approach is via sensitivity analysis. For example, taking a set of *a priori* distributions that can be agreed upon by most reasonable people, we may find that the conclusions are robust, that is, that there is overwhelming evidence of guilt or innocence. A virtue of this theory is that it makes very clear what would be requisite for a consensus among rational people.

There are numerous other examples of nonrobustness in statistics, such as in the analysis of the random effects model in Hill (1980) and, more generally, in problems with many parameters. Such nonrobustness is typically the case in important societal issues. If one asks questions concerning economic policy, or concerning the effect of capital punishment in the deterrence of crime, or concerning the existence of the greenhouse effect, or concerning methods of preventing the destruction of the environment and loss of species, or merely concerning how to dispose of our garbage, there is an extremely large number of underlying parameters of the various models, and the conclusions from even the most careful Bayesian analysis will be highly sensitive to the specification of the various *a priori* probability distributions and likelihood functions. It is really only in the simplest of examples that anything approaching an 'objective' conclusion emerges. When this occurs, it usually stems from the fiducial argument of Fisher or the stable estimation analysis of Savage, and sometimes from the Jeffreys–Savage analysis of the hypothesis testing problem. In these cases we may speak of a robust analysis, but rarely otherwise.

VIII. CONCLUSIONS

The problem of robustness in statistics has been formulated from a decision-theoretic point of view. The standard examples where such robustness occurs and leads to a consensus of opinions have been discussed. These include cases where the fiducial argument is compelling and is a robust Bayesian argument, where the principle of stable estimation holds, and where there is a consensus as to the choice of π_2 in the Jeffreys–Savage evaluation of posterior odds. To obtain a convincing statistical analysis, it is necessary to learn about real-world mechanisms that generate both parameter values and data, to obtain as nearly as possible optimal procedures to make use of such information, and finally to perform a sensitivity analysis under reasonable variation of the input. Because in real-world problems there is typically an enormous number of unknowns, it is important to recognize the limitations of all statistical procedures and to learn how to base inference and decisions upon those aspects of the problem for which there is the solidest evidence, whether about the *a priori* distribution, the likelihood function, or the utility function, rather than mechanically to implement some procedure.

SEE ALSO THE FOLLOWING ARTICLES

PROBABILITY • STATISTICS, BAYESIAN • STATISTICS, FOUNDATIONS • STATISTICS, MULTIVARIATE • STATISTICS, NON-PARAMETRIC

BIBLIOGRAPHY

Aitchison, J., and Dunsmore, I. R. (1975). "Statistical Prediction Analysis," Cambridge University Press, Cambridge.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). "Robust Estimates of Location," Princeton University Press, Princeton, NJ.

Berger, J. (1984). The robust Bayesian viewpoint (with discussion). *In* "Robustness of Bayesian Analysis" (J. Kadane, ed.), pp. 321–372, North-Holland, Amsterdam.

Blackwell, D., and Girshick, M. A. (1954). "Theory of Games and Statistical Decisions," Wiley, New York.

De Finetti, B. (1937). "La prévision: Ses lois logiques, ses sources subjectives," *Ann. Inst. Henri Poincaré* **7,** 1–68.

De Finetti, B. (1974). "Theory of Probability," Vol. 1, Wiley, New York.

De Finetti, B. (1975). "Theory of Probability," Vol. 2, Wiley, New York.

DeGroot, M. (1970). "Optimal Statistical Decisions," McGraw-Hill, New York.

Edwards, A. W. F. (1972). "Likelihood," Cambridge University Press, Cambridge.

Edwards, W., Lindman, H., and Savage, L. J. (1963). "Bayesian statistical inference for psychological research," *Psychol. Rev.* **70,** 193–242 [Reprinted in "Robustness of Bayesian Analysis" (J. Kadane, ed.), pp. 1–62, North-Holland, Amsterdam (1984)].

Feller, W. (1968). "An Introduction to Probability Theory and its Applications," 3rd ed., rev., Wiley, New York.

Feller, W. (1971). "An Introduction to Probability Theory and its Applications," Vol. 2, 2nd ed., Wiley, New York.

Ferguson, T. (1973). "A Bayesian analysis of some nonparametric problems," *Ann. Stat.* **1,** 209–230.

Fisher, R. A. (1939). "Student," *Ann. Eugen.* **9,** 1–9.

Fisher, R. A. (1948). "Conclusions fiduciare," *Ann. Inst. Henri Poincaré* **10,** 191–213.

Fisher, R. A. (1959). "Statistical Methods and Scientific Inference," 2nd ed., Hafner, New York.

Heath, D., and Sudderth, W. (1976). "De Finetti's theorem for exchangeable random variables," *Am. Statistician* **30,** 188–189.

Hill, B. M. (1968). "Posterior distribution of percentiles: Bayes theorem for sampling from a finite population," *J. Am. Stat. Assoc*. **63,** 677–691.

Hill, B. M. (1975). "A simple general approach to inference about the tail of a distribution," *Ann. Stat.* **3,** 1163–1174.

Hill, B. M. (1980). "Robust analysis of the random model and weighted least squares regression." *In* "Evaluation of Econometric Models" (J. Kmenta and J. Ramsey, eds.), pp. 197–217, Academic Press, New York.

Hill, B. M. (1990). "A theory of Bayesian data analysis." *In* "Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard" (S. Geisser, J. Hodges, S. J. Press, and A. Zellner, eds.), pp. 49–73, North-Holland, Amsterdam.

Hill, B. M. (1993). "Parametric models for A_n: Splitting processes and mixtures," *J. R. Stat. Soc. B* **55,** 423–433.

Hill, B. M. (1998). "Conditional probability." *In* "Encyclopedia of Statistics," Wiley, New York.

Jeffreys, H. (1961). "Theory of Probability," 3rd ed., Oxford University Press, Oxford.

Savage, L. J. (1972). "The Foundations of Statistics," 2nd rev. ed., Dover, New York.

Savage, L. J. *et al*. (1962). "The Foundations of Statistical Inference," Methuen, London.

Shafer, G. (1982). "Lindley's paradox," *J. Am. Stat. Assoc*. **77,** 325–351.

Wald, A. (1950). "Statistical Decision Functions," Wiley, New York.



Statistics, Bayesian

Bruce M. Hill
University of Michigan

I. Introduction
II. Decision Making
III. Criticism
IV. Conclusion

GLOSSARY

Likelihood function Function that specifies how the probability of a given realized set of data depends upon unknown parameters or hypotheses under consideration.
Posterior distribution Probability distribution obtained by means of Bayes's theorem and used to represent the beliefs, knowledge, or opinions of a person after the analysis of a specific set of data.
Prior distribution Probability distribution used to represent the beliefs, knowledge, or opinions of a person prior to analysis of a specific set of data.

BAYESIAN STATISTICS is named after Thomas Bayes (1702–1761), whose posthumously published article, "An essay towards solving a problem in the doctrine of chances," initiated the modern-day subject. A biographer of Bayes, Holland (1962), writes of this article that "it ranks as one of the most famous, least understood, and controversial contributions in the history of science. . . . Thomas Bayes . . . expressed ideas in science and in the humanities which will endure as long as man retains his freedom of thought and the power to express it." Indeed, Bayes's theorem and postulate are central to the theory of Bayesian statistical inference and decision making, which in turn are central to understanding of the process of scientific induction and of the nature and limitations of human knowledge. I shall here give some history of Bayesian statistics and a guide to its key precepts. Three main topics will concern us: induction, coherency, and utility.

I. INTRODUCTION

Let us begin with induction. The British philosophers Bacon, Locke, and especially David Hume (1748) proposed notions pertaining to the way in which human beings absorb information or data, in other words, to the process of learning from experience. The theory of probability was in its infancy in the 18th century and was still largely concerned with simple games of chance. It was Bayes who first put forth a probabilistic approach to learning from experience. He carefully analyzed an example in which it was reasonable to presuppose a form of symmetry with respect to the position of balls on a level table, as described below in Section D. There are two separate aspects to the analysis of Bayes: one is the use of Bayes's theorem, which follows easily from the conventional definition of conditional probability; the other is Bayes's postulate,

which amounts to the choice of a particular *a priori* or prior distribution. Bayes's theorem states that if one has a finite partition of the sure event, that is, a finite collection of mutually exclusive and exhaustive events, say H_i, then provided that $Pr\{\text{Data}\} > 0$,

$$Pr\{H_i \mid \text{Data}\} = \frac{Pr\{\text{Data} \mid H_i\}Pr\{H_i\}}{Pr\{\text{Data}\}}, \quad i = 1, \ldots, k.$$

The prior or initial probabilities $Pr\{H_i\}$ represent beliefs, knowledge, or opinions prior to the specific data being analyzed, while the probabilities $Pr\{H_i \mid \text{Data}\}$ represent the posterior or post-data probabilities for the same hypotheses, given the data at hand.[1] The quantities $Pr\{\text{Data} \mid H_i\}$, viewed as a function of i, constitute what is today called the likelihood function. This function specifies how the probability of the observed data, which is considered fixed and given, depends upon the unknown quantity or parameter, here the hypothesis H_i. The likelihood function is central to all modern-day statistical theory (Savage, 1962, pp. 15–20) and is viewed by some as having objective validity, for example, because it may be based upon an established probability model for the data. On the other hand, the initial probabilities are often viewed as being more controversial since they may represent subjective judgment as to the truth of the various hypotheses. No one questions the validity of Bayes's theorem as a mathematical theorem, and the controversy about this theorem concerns its use for scientific inference and decision making. Those who object to its use sometimes argue against the explicit introduction of subjective judgment as represented by the $Pr\{H_i\}$. Those who support its use view subjective or personal judgment as indispensable, not only for the prior probabilities, but even with respect to the choice of a model or likelihood function for the data, and question whether any other concept of probability has real-world meaning.

It should be noted that use of Bayes's theorem is nearly tantamount to the notion that statistical inference consists in the rational updating of initial opinions, as represented by the $Pr\{H_i\}$, to become post-data opinions, as represented by the posterior probabilities $Pr\{H_i \mid \text{Data}\}$. It is in this sense that Bayesian statistics is inductive, that is, it proposes a way to learn from experience. One goes from the state of knowledge represented by the initial probabilities to that represented by the posterior probabilities, using a method which incorporates in a coherent fashion the data being analyzed. The posterior probabilities based upon a particular set of data may then be used as prior probabilities for the analysis of a new set of data. Thus the procedure is ongoing, with opinions continually being revised as new data are assimilated. It can be argued that such a method of updating probabilities is part of a formal logic for dealing with uncertainty.

In the modern Bayesian approach of de Finetti (1937, 1974) and Savage (1972), the probabilities are all interpreted as subjective or personal probabilities, representing *judgment* rather than merely past frequencies, although such frequencies may be instrumental in the evaluation of the initial probabilities. Probability itself has been a highly controversial subject, with the three primary approaches being the formalistic approach of Kolmogorov, the frequentistic approach based on the assumption of limiting frequencies, and the subjective approach. In the celebrated exchangeability theorem of de Finetti (1937) a result, scarcely less important than Bayes's theorem itself, was proved, that gave the first clear understanding of the relationship between the subjective approach and the frequency approach, and which some believe to constitute a solution to the problem of induction as formulated by Hume.

The problem that de Finetti posed and largely solved was that of justifying from a subjective point of view the importance of statistical frequencies. Suppose that we have a sequence of related events E_i for $i = 1, \ldots, N$, each of which may or may not occur, for example, where E_i is the result of the ith flip of a coin. For convenience, we can assign the value 1 to heads and 0 to tails. It is natural sometimes to view the order of the flips as irrelevant, so that any permutation of the same sequence of results is given the same probability. This leads to the notion of exchangeability, which is a symmetry condition, specifying that joint distributions of the observable random quantities are invariant under permutations of the indices. Sequences of independent and identically distributed random variables are necessarily exchangeable, but the collection of exchangeable distributions is much more general, including in addition mixtures of such sequences, and sequences that arise from sampling without replacement. De Finetti proved a representation theorem which implies that if one regards a long sequence as exchangeable, then it necessarily follows that one's opinions can be represented, approximately, by a mixture of Bernoulli sequences, that is, of sequences for which there is a 'true' (but typically unknown) probability P of heads, and such that conditional upon $P = p$, the observations are independent with probability p of giving a head (success) on each toss. The parameter P is then said to have the mixing distribution given by de Finetti's theorem. Such a representation is exact for infinite sequences, and a slight modification holds also for finite sequences, the difference amounting to that between sampling with and without replacement from an urn.

[1] It is customary to use the symbol '|' to indicate that the information to the right of this symbol is being conditioned upon.

If one observes the outcomes in a sufficiently long such sequence of exchangeable events, say E_i for $i \leq n$, where $n < N$, then it easily follows that the remaining events E_i for $n < i \leq N$ are also conditionally exchangeable, given this data, and that to a good approximation one will be led to base opinions about a future trial on the past empirical frequency with which heads (or success) has occurred; this constitutes de Finetti's justification for induction. In his work de Finetti emphasized prediction of observable quantities, rather than what is usually called inference about parameters. In other words, he was primarily interested in the posterior distribution of future observable variables, rather than in the posterior distribution for the parameter P.

The exchangeability theorem of de Finetti in fact shows that the conventional Bernoulli parameter P can be interpreted as the limiting proportion of successes, and the above mixing distribution, sometimes called the de Finetti measure, is identical with the *a priori* distribution in the sense of Bayes. Consider, for example, a large, finite number k of trials, such as flips of a coin, each of which can result in success (1) or failure (0). Assume the trials are exchangeable, and let P_k be the proportion of successes to be obtained in the k trials. The possible values for P_k are then the fractions i/k, with $0 \leq i \leq k$. Now introduce the hypothesis H_i that asserts that the proportion P_k will have the value i/k. One and only of these hypotheses must be true. Let $\pi_k(\cdot)$ represent the *a priori* probability distribution for the hypotheses, so that $\pi_k(i/k)$ is the *a priori* probability for hypothesis H_i. It can be shown that as k goes to infinity, π_k converges to the mixing distribution π for P that arises in de Finetti's representation theorem.

Now consider the sequence of observable trials X_j for $j = 1, \ldots, n$, where $X_j = 1$ if there is a success on trial j, and $X_j = 0$ if there is a failure on trial j. If k is substantially larger than n, then conditional upon the value of $P_k = p$, this sequence is approximately a Bernoulli sequence with parameter $P = p$. Since P_k is itself random, this produces, to a good approximation, a probabilistic mixture of Bernoulli sequences, with π as the weighting function for the mixture. Suppose that we now observe the results of the first n trials, so that we are given the data $X_j = x_j$ for $j = 1, \ldots, n$, where x_j is 1 or 0 depending upon the outcome of the jth trial. The likelihood function for these data is therefore approximately

$$L(i/k) = (i/k)^r \times (1 - i/k)^{n-r},$$

where $r = \sum_{i=1}^{n} x_i$ is the number of successes in the n trials.[2] If the data consist of the results of the first n trials, it follows that the posterior distribution for the parameter P_k given by Bayes's theorem is

$$\pi_k^*(i/k) = Pr\{H_i \mid \text{Data}\} \approx \frac{\pi_k(i/k) \times L(i/k)}{\sum_{i=0}^{k} \pi(i/k) \times L(i/k)}.$$

Given the outcomes of the first n trials, the posterior predictive distribution for the results of the future trials $X_{n+1}, \ldots, X_k$ can therefore be obtained by first updating the prior distribution π_k for P_k to yield the posterior distribution π_k^* for P_k and then using this posterior distribution as a mixing distribution to represent opinions about future events. For example, to predict the next observable variable X_{n+1}, we have

$$\begin{aligned} &Pr\{X_{n+1} = 1 \mid \text{Data}\} \\ &\quad = \sum_{i=0}^{k} Pr\left\{X_{n+1} = 1 \mid \text{Data}, P_k = \frac{i}{k}\right\} \times \pi_k^*(i/k) \\ &\quad = \sum_{i=0}^{k} \frac{i}{k} \times \pi_k^*(i/k). \end{aligned}$$

The same basic method can be used for the joint posterior predictive distribution of several future observations. See Savage (1972, pp. 50–55).

Thus the process of induction proceeds as follows. One first decides, using subjective judgment, whether one regards the sequence as approximately exchangeable. If so, one specifies one's prior distribution π_k for the proportion of heads in a large number k of trials. Finally, given the data, one updates this prior distribution to become a posterior distribution by means of Bayes's theorem and then uses the posterior distribution to obtain the posterior predictive probabilities for future observations. See Section D for further discussion. Since de Finetti's theorem holds not only for events, but for the most general random quantities, such a procedure is theoretically available in all problems of statistical inference based upon exchangeable sequences of observations. Although one may object for one reason or another to such a procedure, it is clearly a proposed solution to the problem of the revision of human knowledge based upon experience.

Now let us consider a justification for such a procedure based upon the notion of coherency. In the work of de Finetti (1937, 1974) the primary aim is to avoid the possibility of a Dutch book, that is, to avoid sure loss on a finite collection of gambles. Thus, suppose one has taken on a finite collection of gambles of any type whatsoever, including such things as investments, insurance, and so on. One can prove that a necessary and sufficient condition for the avoidance of sure loss of a positive amount greater than some $\epsilon > 0$ on a finite number of gambles (whether unconditional or conditional) consists in consistency with

[2] The exact likelihood function is given by a hypergeometric distribution. See Feller (1968, p. 43).

the axioms of finitely additive probability theory. In other words:

1. The probability of an event is a nonnegative number.
2. $Pr\{S\} = 1$, where S is the sure event.
3. The probability of a finite union of disjoint events is the sum of the probabilities.
4. Conditional probabilities are such that

$$Pr\{A \mid B\} = Pr\{A \text{ and } B\}/Pr\{B\},$$

provided that $Pr\{B\} > 0$.

These axioms for subjective probability theory are identical with those for the conventional probability theory of Kolmogorov (1950), except that in the latter the finite additivity of Axiom 3 above is replaced by the stronger condition of countable additivity, so that also the probability of a *denumerable* union of disjoint events is the sum of the probabilities.[3] Countable additivity has never been demonstrated within the rigorous approaches to probability theory, and is viewed by subjectivists as merely a particular continuity condition that one may or may not wish to adopt in a particular context, just as with the use of continuous functions in physics.[4] In de Finetti's theory a probability is interpretable as a price one pays for a gamble on the occurrence of an event. A conditional probability on A given B is interpreted as a price for a called-off gamble, that is, a gamble which one either wins, loses, or is called off, according as to whether both A and B occur, B but not A occurs, or B does not occur, respectively. de Finetti proposes the use of gambles involving small amounts of money to assess probabilities, so as to avoid utility complications. His theory is consistent with that of Savage (1972), who simultaneously develops subjective probability and utility.

II. DECISION MAKING

The next subject that I wish to discuss concerns the relationship of the Bayesian approach to decision making. In the work of Blaise Pascal and Daniel Bernoulli it was recognized that when one is faced with a decision problem, one is implicitly dealing with questions of judgment of truth, as represented by probability, and judgments of value, as represented by utility. A utility function is a numerical valued function defined for various 'consequences' or 'rewards.' The modern Bayesian approach to utility was first formulated by Ramsey (1926), and culminated in the theory of Savage (1972). Savage shows that one can derive both subjective probability and utility by means of the same underlying concepts and methods, which he formulated axiomatically. He begins with qualitative probability, and then shows under what circumstances it can be extended to quantitative probability. He then derives the existence of a utility function, which is such that one action is preferred to another if and only if the first yields a higher expected utility. Within his theory optimal decision-making requires specification of a prior distribution, a likelihood function, and a utility function. Given the data, one uses Bayes's theorem to update the prior probabilities to posterior probabilities and then acts in a specified decision problem so as to maximize posterior expected utility. In some contexts it is convenient to employ loss functions instead of utility functions, in which case one minimizes posterior expected loss. The utility function can be either a personal utility function or some group utility function, however, chosen. Decision procedures of this sort are called Bayes procedures. See DeGroot (1970, Chapter 8).

Although some would argue against the use of utility or loss functions in problems of statistical inference (as opposed to decision problems), others have observed that without some attempt to measure utility or loss there would be no punishment for even the most absurd procedures. It is true, however, that sometimes the utility or loss function is not well known or that there is disagreement with regard to its choice. As observed by Savage, in a formal sense one can deal with unknown parameters of a loss function just as one deals with other statistical parameters. See Savage (1972, p. 15).

There are a number of variations of the basic de Finetti coherence theory that differ with respect to details of the exact definition of sure loss. For example, some statisticians of the school of A. Wald take admissibility as defining the concept of coherence, and others take extended admissibility. Admissibility of a decision procedure means that it is not possible to find another procedure that does at least as well for every possible value of the parameter or unknown quantitity and strictly better somewhere, where performance is measured with respect to expected loss, given the parameter, for some specified loss function. Extended admissibility means that it is not possible to do better everywhere uniformly by some fixed $\epsilon > 0$. If a procedure is admissible, then it is plainly also extended admissible, so extended admissibility is a less restrictive condition. The main theorems on these subjects are to the effect that either a statistical procedure is a Bayes procedure, or a limit of Bayes procedures, or else it is

[3] In the Kolmogorov formulation Axiom 4 provides the definition of conditional probability.

[4] In his book that founded the modern measure-theoretic approach to probability, Kolmogorov (1950, p. 15) states that the axiom of countable additivity cannot be justified except by expediency. Later he developed a theory of complexity to replace some or all of conventional (non-Bayesian) probability theory.

inadmissible, or even inadmissible in the extended sense. See DeGroot (1970, p. 133). The upshot of the theory of Wald, which was formalized outside of the Bayesian theory, is that reasonable objective criteria for performance of statistical estimators or decision procedures dictate that optimal procedures are necessarily Bayes procedures.

In fact, any real-world decision procedure must necessarily be implemented on a computer with finite memory. If a specified such procedure is not a Bayes decision procedure, then it is possible to improve its risk function uniformly in the parameter by Bayes decision procedures. In other words, there exists an $\epsilon > 0$ such that no matter what the true value of the parameter or unknown, the expected loss (or risk) will be lowered by at least that ϵ. Furthermore, this can be done using simple and standard algorithms for linear programming (Hill, 1998). This result demonstrates that any non-Bayesian decision procedure is logically defective, and according to the frequency theory of probability, will lead to huge and unnecessary costs in repeated applications.

Nowadays, many statistical procedures have the word 'Bayes' attached to them, for example 'empirical Bayes' and 'intrinsic Bayes.' The use of the word 'Bayes' may be intended to suggest the desirable properties and character of genuine Bayes procedures, whether or not this is in fact the case. However, in all such examples either the procedure is in fact a genuine Bayes procedure, in which case there is no need for such adjectives as 'empirical' or 'intrinsic,' or else it can easily be improved upon uniformly by genuine Bayes procedures. Roughly, one can say that any procedure, proposed from any point of view whatsoever, is either a Bayes procedure or else is defective in an objectively verifiable sense.[5]

III. CRITICISM

Let us now turn to some of the criticisms of the Bayesian approach. Bayes's theorem was nearly lost until it was rediscovered by Laplace, who was responsible for extremely innovative uses of Bayesian methods and applications to many different areas of science and human affairs. Bayesian methods were also employed by Gauss in his theory of least squares. Later in the 19th century, however, there was criticism of some of the uses to which Bayesian methods had been put by Laplace and his disciples. This view culminated in the work of the statistician R. A. Fisher, who in his early work was a severe critic of the Bayesian approach. Fisher introduced the fiducial approach to statistical inference, which he initially believed was more objective than the Bayesian. It is one of the many ironies concerning the Bayesian approach that Fisher eventually understood that his fiducial argument was at best only a special case of Bayes's theorem, with an appropriate uniform prior distribution and using a different method to derive the posterior distribution.

This takes us back to Bayes's postulate. His example concerned the relative positions of balls on a level table relative to a specified side wall. An initial ball is rolled and then n additional balls by the same mechanism, the rolls being mutually independent. Each of the n additional balls is scored a success (1) or a failure (0) according as it is or is not closer to the specified side than the initial ball. If the data consist of r successes in n rolls following the roll of the initial ball, then Bayes obtains $(r+1)/(n+2)$ for the conditional probability that the next roll will be a success. He assumes that there is no information as to the positions of the balls, only knowledge as to which rolls gave successes and failures. Although the original rolls are independent, the comparison with the initial ball introduces exchangeable dependency, and the sequence of successes and failures in fact forms the first nontrivial example of an exchangeable sequence.[6]

Bayes's arguments were along the following lines. First, he assumed that the levelness of the table implied a uniform distribution for the distance of the initial ball from the specified side. Over a century later Poincaré's study of the roulette wheel problem provided a justification for the uniform distribution, at least if the balls were rolled with a high velocity. See Feller (1971, p. 62). In the case of a uniform distribution for the position of the initial ball relative to the specified side, if X denotes this random position and if the length of the table is taken as unity, then conditional upon $X = x$, the sequence of n successes and failures based upon the n additional rolls becomes a Bernoulli sequence with unknown parameter $P = x$. Since $P = X$ is itself random (with a uniform distribution), this generates a probability mixture of Bernoulli sequences and is exchangeable. Next, even apart from his explicit example, Bayes suggested that it would be reasonable in the case of n trials of a similar structure, about which one knew little, to take all numbers of successes in n trials as equally likely, that is, with probability $1/(n+1)$ for each of the $n+1$ possibilities. From a modern point of view, based upon de Finetti's exchangeability theorem, this approach would not require the introduction of any unknown parameters such as the P of a Bernoulli sequence. However,

[5] This need not, however, entirely rule out non-Bayesian procedures, since they sometimes provide reasonable approximations to Bayes procedures in situations which are too complex to allow for an exact Bayesian solution to be derived and implemented, at least in the present state of science and technology. However, to justify the usage of such a non-Bayesian procedure it would be necessary to examine the adequacy of the approximation in each application.

[6] The concept of exchangeability was not explicitly formulated until the 20th century.

it can be proved that if such an assignment of probability is made for each n, then implicitly one is acting as though there were a P and one had a uniform prior density function π for P, where now P can take on any value between 0 and 1.

Much of the criticism of Bayes's postulate revolves about the appropriateness of such a prior distribution in situations other than the one he explicitly considered, where it was assumed that the table was level and each position of each ball equally likely. Fisher argued that for a Bernoulli sequence, if one transforms the original parameter P to a different parameter, say $\Theta = f(P)$, where f is a nonlinear monotonic transformation mapping the unit interval into itself, then one cannot in the usual theory of probability have both P and Θ uniformly distributed. Hence he questioned the appropriateness of taking a uniform *a priori* distribution for P as opposed to a uniform *a priori* distribution for such a Θ. However, it is clear that if the *a priori* distribution is to represent a state of ignorance with regard to the proportion of successes in n trials, then any nonuniform distribution for this proportion would constitute a bias towards some particular proportions as opposed to others (Edwards, 1978).

Fuller understanding of this phenomenon revolves around the theory of precise measurement (or stable estimation) of Savage (1962, pp. 20–25) and DeGroot (1970, pp. 198ff) which built upon the work of Laplace and Harold Jeffreys. The likelihood function for the parameter P is the function $L(p) = p^r \times (1-p)^{n-r}$, where r is the number of successes in n trials and p is between 0 and 1. One can show that for sufficiently large n this function tends to be sharp or highly concentrated relative to the prior distribution π. Since by Bayes's theorem the posterior density for P is proportional to the product of the likelihood function and prior density for P, it follows under some very mild conditions that if the prior density is nearly constant in an interval of length $c/\sqrt{n}$ about the maximum-likelihood estimate,[7] for some suitable constant c, then for practical purposes the posterior distribution can be taken as just the likelihood function, standardized so as to have unit area.

This constitutes the main sense in which statistical inference can be objective, at least within the Bayesian theory. Thus, although one need not have a uniform prior distribution for P, the posterior distribution, to a good approximation, may be as though this were the case. This argument in fact provides a justification based upon robustness for Bayes's postulate in many examples. The related work of Jeffreys (1961) is based upon choices of prior distributions, sometimes called invariant or uninformative, which are not necessarily uniform in the original parameter. For example, when the parameter is positive Jeffreys ordinarily takes the logarithm of the parameter to have a uniform distribution, with related modifications for bounded parameters and other special cases. When the conditions for precise measurement obtain, this typically makes only a negligible change in the posterior distribution. Many Bayesians refer to all such prior distributions as diffuse relative to the likelihood function. A diffuse prior distribution for the empirical distribution in a finite population leads to Bayesian nonparametric statistical prediction and inference as in Hill (1968), which generalizes the original work of Bayes.

The phenomenon of precise measurement is completely general, although particularly in high-dimensional problems prior distributions are often not diffuse relative to the likelihood function. Jeffreys also initiated the development of the theory of Bayesian tests of hypotheses, which involves nondiffuse prior distributions under the alternative hypothesis. See also Savage (1962, pp. 29–33). Here the notion is that given a hypothesis, there is a prior distribution for the parameters which that hypothesis specifies. One then obtains the posterior probability for the hypothesis, given the data, using Bayes's theorem as before, except that now the probability for the data given the hypothesis is expressed as a mixture or integral, integrating out with respect to unknown parameters. Such a procedure is quite different from conventional tests of hypotheses and resolves some difficulties associated with the latter.

Another criticism of the Bayesian approach concerns its use for prediction of future observables. From its inception with Bayes and Laplace, the Bayesian approach was concerned not only with conventional statistical inference about unknown parameters, but also with the prediction of future observables. Examples include the prediction of weather, interest rates, survival times for patients given a treatment for cancer, earthquakes, or the collapse of a bridge. The Bayesian approach allows one to obtain a posterior predictive distribution for future observables as well as a posterior distribution for conventional parameters. In fact, one can first obtain the posterior distribution for any unknown parameters and then integrate out the parameters with respect to this posterior distribution to obtain the predictive distribution for future observables. For example, consider a Bernoulli sequence with (unknown) parameter P, and suppose the available experimental data are that in n trials there were exactly r successes. Let $\mathcal{D}$ represent this data, and $L(p) = p^r \times (1-p)^{n-r}$ be the likelihood function. Then

$$P(X_{n+1} = 1 \mid \mathcal{D}) = E[P \mid \mathcal{D}] = \int_0^1 p \times \pi^*(p)\, dp.$$

The above equation gives the posterior probability that the next trial is a success, given the data $\mathcal{D}$, where

[7] The maximum-likelihood estimate of a parameter is the value of the parameter that maximizes the likelihood function.

$$\pi^*(p) = \frac{\pi(p) \times L(p)}{\int_0^1 \pi(p) \times L(p)\, dp},$$

is the posterior density for P and the expectation is taken with respect to the posterior distribution.[8]

Under some mild regularity conditions, $P(X_{n+1} = 1 \mid \mathcal{D})$ has the approximate value $(r+1)/(n+2)$ when n is sufficiently large. When $r = n$, so that one has observed n successes in n trials, this rule is known as Laplace's rule of succession. See Feller (1968, p. 124) for an example of the type of criticism of this rule that used to be common before the optimal character of Bayes decision rules was fully understood. Note that for small n this rule is much more reasonable than the maximum-likelihood estimate, which is r/n. For example, when $n = 1$, the maximum-likelihood estimate is either 1 or 0, depending upon whether or not a success occurred, whereas the rule of Laplace gives either 2/3 or 1/3. The predictive point of view was particularly emphasized by de Finetti (1937, 1974), who insisted that statistical procedures should be operationally meaningful and proposed the use of scoring rules to assess the performance of individuals making predictions. Aitchison and Dunsmore (1975) provide a careful discussion of the predictive viewpoint. They argue that conventional non-Bayesian methods of prediction, which simply plug in the maximum-likelihood estimate of the parameter and make predictions as though this estimate was exactly equal to the true value, are often seriously inadequate because they do not allow for realistic posterior uncertainty about the parameter; they do not perform very well, especially in typical real-world problems, where sample sizes are not large.

In the case of prediction of future observations, few would imagine that there can be any truly objective way to make such predictions. Yet by virtue of the exchangeability theorem of de Finetti, the Bernoulli parameter P can be interpreted as simply the long-run proportion of successes, so that inference about P is equivalent to a prediction about such a future proportion. Similarly for other conventional statistical parameters. The most persistent *philosophical* objection to Bayesian procedures concerning lack of objectivity can therefore be seen in a new light. While it is true that the choice of the *a priori* distribution π is often subjective, this merely reflects the fact that in typical real-world decision problems there is a diversity of different opinions and background knowledge, so when the experimental data are not sufficiently informative this can lead to substantially different posterior distributions. There is, however, no way to avoid this, other than by restricting attention only to problems where sample sizes are so large that the *a priori* distribution more or less washes out and everyone is left with essentially the same posterior distribution, as in the Bayesian precise measurement theorem of Savage. Unfortunately it is rarely the case that the data are so informative. Since conventional 'non-Bayesian' statistical methods are typically equivalent to the use of certain special diffuse prior distributions π for the parameters of the model, they do not offer any help with respect to the question of objectivity, but merely implicitly dictate use of specific *a priori* distributions without any serious attempt at justification.

When the data are not sufficiently informative, some would argue that the arbitrariness of the *a priori* distribution prevents any useful conclusions to emerge from the analysis of the data. This may or may not be the case. It is true that in certain types of problems, particularly high-dimensional problems, the conditional probabilities given by Bayes's theorem are often highly sensitive to the *a priori* distribution. Furthermore, in such examples it may be extremely difficult to perform either the necessary mathematical analysis or appropriate computer simulations. In such cases some statisticians choose models and *a priori* distributions which make the analysis mathematically tractable, but without any other attempt at justification. This makes their conclusions both unrealistic and unconvincing to others, and is the type of pseudo-Bayesian analysis that Fisher and J. Neyman were rightly critical of. In scientific work it is not appropriate to choose models and *a priori* distributions in such arbitrary ways.

In scientific work Bayesian methods are most effective either in examples such as that of Bayes where there is cogent reason for both the model and the *a priori* distribution, or in problems where the likelihood function is sharp relative to the *a priori* distribution as in the precise measurement theory of Savage. Even when neither of the above conditions holds, in some real-world decision or scientific problems there is at least a rough consensus among knowledgeable people as to the choice of the model and *a priori* distribution, and this consensus may already largely dictate the answer to a decision problem via sensitivity analysis. Bayesian methods are most controversial when there is extreme sensitivity to both the model and the *a priori* distribution, as in nonparametric statistics, so that one can obtain virtually any answer one wants. In such examples it is better to start with some conventional non-Bayesian solution to the problem and then improve it uniformly by a Bayes procedure, using linear programming. In this case the *a priori* distribution need not reflect subjective judgment, but one at least obtains the logically desirable properties of Bayes procedures.

The modern Bayesian standpoint is that objectivity consists in the explicit consideration of specific prior distributions, likelihood functions, and utility functions;

[8] Although given here in terms of a Bernoulli sequence, the above method of evaluation is quite generally available in problems of prediction.

by analyzing sensitivity to these choices, one can delineate those problems for which a consensus can be obtained. When opinions and preferences vary considerably among different individuals, then the Bayesian theory also makes it clear that this is in fact the case, so that objectivity is not possible. This is especially important in the area of hypothesis testing, where typically the prior distribution under the alternative hypothesis cannot be chosen in an 'objective' fashion, and so the conclusions of such an analysis are highly subjective. See Borel (1965, p. 165) for an illuminating general discussion of 'objectivity' in statistics. Borel sides with H. Poincaré in making the distinction between different degrees of subjectivity, and viewing 'objectivity' as arising when subjective probabilities have the same value for many people.

Finally, a serious criticism of the Bayesian approach concerns the question of time coherency and whether the Bayesian approach requires that *all* learning should take place via Bayes's theorem. Some modern Bayesians do not believe this to be the case, and have argued for a version of Bayesian statistics in which computational complexity plays an important role, as, for example, in Bayesian data analysis. Thus realistically complex data may first be analyzed graphically or using other data-analytic techniques; new models, parameters, and hypotheses may be formulated on the basis of such techniques; and finally Bayesian methods may be used to obtain the post-data probability distribution for the parameters of such models as have been introduced via the data analysis. In this process the overriding concern must be maximization of post-data expected utility, but Bayes's theorem continues to play a fundamental role (Hacking, 1967; Hill, 1990).

IV. CONCLUSION

As in the quotation at the outset of this article, since its origin the Bayesian approach has been highly controversial, and there have been many historical curiosities and ironies concerning it. These include the fact that both Bayes's theorem and de Finetti's theorem were nearly lost, although nowadays they are considered two of the most important results in the history of science. Under the impetus of Fisher and Neyman (who did not themselves see eye to eye) the Bayesian approach was subjected to a virulent attack during the early part of the 20th century.[9] Recently, however, there has been a strong and spirited resurgence of interest in the Bayesian approach and greater understanding of its role in all of the sciences and in human decision making.

[9]See Gillman (1992) for an amusing example of how the rejection of the Bayesian theory of conditional probability led to a public controversy in the game-show problem of Marilyn vos Savant.

SEE ALSO THE FOLLOWING ARTICLES

DATA MINING AND KNOWLEDGE DISCOVERY • DATA MINING, STATISTICS • PROBABILITY • RELIABILITY THEORY • STATISTICAL ROBUSTNESS • STATISTICS, FOUNDATIONS • STATISTICS, MULTIVARIATE • STATISTICS, NONPARAMETRIC

BIBLIOGRAPHY

Aitchison, J., and Dunsmore, I. R. (1975). "Statistical Prediction Analysis," Cambridge University Press, Cambridge.

Bayes, T. (1764). "An essay towards solving a problem in the doctrine of chances," *Phil. Trans*. **53,** 269, 370–418.

Borel, E. (1965). "Elements of the Theory of Probability," Prentice-Hall, Englewood Cliffs, NJ.

De Finetti, B. (1937). "La prévision: Ses lois logiques, ses sources subjectives," *Ann. Inst. Henri Poincaré* **7,** 1–68.

De Finetti, B. (1974). "Theory of Probability," Vol. 1, Wiley, New York.

DeGroot, M. (1970). "Optimal Statistical Decisions," McGraw-Hill, New York.

Edwards, A. W. F. (1978). "Commentary on the arguments of Thomas Bayes," *Scand. J. Stat*. **5,** 116–118.

Feller, W. (1968). "An Introduction to Probability Theory and its Applications," Vol. 1, 3rd ed., Wiley, New York.

Feller, W. (1971). "An Introduction to Probability Theory and its Applications," Vol. 2, 2nd ed., Wiley, New York.

Gillman, L. (1992). "The car and the goats," *Am. Math. Monthly* **9,** 37.

Hacking, I. (1967). "Slightly more realistic personal probability," *Phil. Sci*. **34,** 311–325.

Hill, B. M. (1968). "Posterior distribution of percentiles: Bayes' theorem for sampling from a population," *J. Am. Stat. Assoc*. **63,** 677–691.

Hill, B. M. (1990). "A theory of Bayesian data analysis," *In* "Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard" (S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, eds.), pp. 49–73, North-Holland, Amsterdam.

Hill, B. M. (1998). "Conditional probability," *In* "Encyclopedia of Statistics," Wiley, New York.

Holland, J. D. (1962). "The Reverend Thomas Bayes, F. R. S. (1702–61)," *J. R. Stat. Soc. A*. **1962,** 451–461.

Hume, D. (1748). "An Enquiry Concerning Human Understanding," London.

Jeffreys, H. (1961). "Theory of Probability," 3rd ed., Oxford University Press, Oxford.

Kolmogorov, A. N. (1950). "Foundations of Probability," Chelsea, New York.

Ramsey, F. (1926). "Truth and probability," *In* "The Foundations of Mathematics and Other Logical Essays" (R. B. Braithwaite, ed.), Press, New York.

Savage, L. J. (1962). "The Foundations of Statistical Inference, A Discussion," Methuen, London.

Savage, L. J. (1972). "The Foundations of Statistics," 2nd rev. ed., Dover, New York.



Statistics, Foundations

D. A. S. Fraser
University of Toronto

I. Background
II. Overview
III. Probability Model
IV. Statistical Model
V. Statistical Theory
VI. Foundations
VII. Principles
VIII. Likelihood Asymptotics

GLOSSARY

Ancillary A statistic with a fixed distribution, free of the parameter; for applications a physical interpretation is appropriate.

Conditioning A principle that recommends conditioning on an appropriate or relevant ancillary.

Likelihood A principle that recommends using only the observed likelihood function for statistical inference.

Probability model A mathematical construct for describing the long-run behavioral properties of a system that is actually or conceptually performable or repeatable under essentially constant conditions.

Statistical model A probability model with free parameters such that certain particular values for the parameters provide a probability model that is a good approximation to the process being examined.

Statistical reduction Sufficiency can reduce the dimension of the appropriate variable and lead to a marginal model. Conditioning can reduce the dimension of the free variable and lead to a conditional model.

STATISTICS seeks patterns and relationships within information collected concerning real world contexts. The contexts can range from sharply defined scientific issues, to social science matters, to wide-ranging industrial, commercial, or financial issues, to the contents of massive data storage devices. The information can be the collected results, tabulations, or electronic records obtained, or to be obtained, concerning the real world contexts. The patterns and relationships are the simplified descriptions including cause–effect relationships concerning the real world contexts being investigated.

Statistics is the methodology and theory for this process of extracting the important patterns and relationships from the collected information. The foundations of statistics are the basic elements of theory and principle that structure the planning and collection of information and the extraction of patterns and relationships.

I. BACKGROUND

Some elements of statistical thought were present even in Biblical times. But it was money and gamblers in 16th century France that brought mathematical skills to the calculation of probabilities for games of chance, with the intent of increasing financial gains. Also, astronomers had many precise measurements but measurement error created deviations from theoretical patterns, and the mathematical skills of Gauss, Laplace, and many others were applied to this departure of data from theory. More recently the social sciences collect data where underlying patterns are heavily masked by variation, both variation in the subject material and variation due to measurement error. In addition, in agricultural experimentation the long delay from planning to data collection provides a strong incentive to carefully plan an investigation and to carefully analyze the results in order to maximize the quantity and quality of the conclusions obtained from the long process from initial planning to final presentation. By now statistics has entered into almost all areas of human endeavor.

II. OVERVIEW

Statistics involves the planning of investigations, the carrying out of the investigation, the collection of data, the processing of the data, and the analysis through to conclusions from the investigation and data. Some of these components, however, can exist almost in isolation from the others; for example, the analysis of very large datasets on consumer choices can in practice be almost totally separate from the original collection of the data.

Early activity focused on variation in results from quite well-defined processes such as the games of chance in 16th century France. The variation or randomness in the results was formalized as probability theory, which is now a well-established component of statistical modeling.

Early formulations of this theory gave the structure needed to describe the measurement error in traditional physics and astronomy. In these areas there would be an underlying law or theory that would prescribe deterministic results and then the actual data would depart from this due typically to small measurement errors. Probability theory provided a way to model the error and this combined with the deterministic pattern led to statistical models. These were probability models but with free parameters corresponding to characteristics of the underlying deterministic process.

This type of statistical model, however, has much broader application than the traditional measurement error context. It could describe variation in the underlying material being investigated and even variation in the quantities that were of direct interest. This provides applications to the social sciences and to industrial, commercial, financial, and other areas.

Statistical methodology, however, extends far beyond this probabilistic modeling of error and variation. Small and large aggregations of data can be examined to get summary presentations, elicit patterns for further investigation, seek anomalies relative to previous patterns, and generally obtain guidance for future analysis and development often with a distinct profit motive not unrelated to that of the 16th century origins of the discipline.

As indicated above, a traditionally central part of statistics is concerned with investigations that lead to a statistical model, a probability model with free parameters corresponding to unknowns in the investigation. The primary statistical problem is then the analysis of data that are viewed as coming from this statistical model. Some areas of application may have special types of model form and be developed within related disciplines often with quite special notation. For example, actuarial science examines the lifetime of people or objects, obtaining probabilities of living, dying, or accidents in given time intervals. More recently this has extended into general insurance, financial, and related areas. Other specialized areas include operations and management research, control theory, econometrics, and many biological, industrial, management, and financial research areas. Interpreted broadly, statistics is the theory and methodology of knowledge acqusition, the use of this knowledge to manage and control, and the source of structure for many substantive areas of intellectual activity.

III. PROBABILITY MODEL

As indicated above, the development of probability methods provided the major ingredient for the development of statistics and continues as a significant core component.

A pure context involves a process or system that can be operated or performed repeatedly under essentially constant condition. For example, an ordinary six-faced die can be shaken or tossed and the upward face observed, or the time to failure can be observed when an electronic component from a stable manufacturing process is tested. The elementary probability model involves a *sample space* S of possible outcomes values, a class $\mathcal{A} = \{A\}$ of *events* or *subsets* A of S that are of interest, and a *probability* $P(A)$ for each event A. For the die $S = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{A}$ consists of all possible subsets of S such as $\{2, 4, 6\}$, $\phi = \{\cdot\}$, or S itself; for a symmetrical die with implied symmetric probabilities $p(A) = \#A/6$, where $\#A$ is the number of points in A. Figure 1 shows the amount of probability at each point in S for this symmetric case. For the lifetime example, S consists of the positive real line or

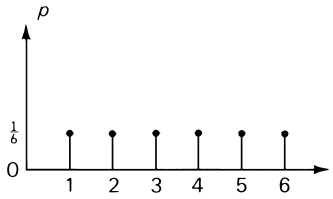


FIGURE 1 The uniform distribution on {1, 2, 3, 4, 5, 6}; the behavior for a symmetric die.

perhaps for convenience the whole real line R, $\mathcal{A}$ consists of all intervals and things predicated by intervals under countable union and intersection, and P could have many forms depending on the context. One simple possibility is the normal distribution with $P(A) = \int_A f(y)dy$, where $f(y) = (2\pi\sigma^2)^{-1/2} \exp\{-(y-\mu)^2/2\sigma^2\}$, which is located at some point μ and scaled by σ. For a location $\mu = 0$ and $\sigma = 1$, Fig. 2 shows the density function labeled $n = \infty$. The other curves provide examples from the Student family, which provides longer tailed distributions than the normal. Of course these distributions are on the whole real line, in an absolute sense, whereas the lifetime variable is necessarily positive. Nonetheless they may provide good approximations in many applications, which is the objective of probability modeling. A somewhat different possibility is provided by the exponential (θ) model $f(y;\theta) = \theta^{-1} \exp\{-y/\theta\}$ on the positive line. This can describe a constant failure rate situation and θ is the mean lifetime. This model *is* on the positive real line. Of course, other models can describe lifetime depending on the underlying process being investigated.

The probability $P(A)$ of an event A is the proportion of occurrences of the event in a long sequence of repetitions; ideally it is the limit of the proportion as the number of repetitions goes to infinity. In applications it is an empirical phenomenon that the proportion $\hat{P}_n(A)$ of occurrence of A in a long sequence of repetitions goes to a limit, or otherwise some change in conditions has occurred that has modified the process or system. This empirical phenomenon for a particular process may be difficult to verify beyond some reasonable approximation, and its assumption is typically based on the degree to which initial conditions are seen to be constant under repetitions.

Probability theory constructs and examines very general models and develops techniques for calculating probabilities for various types of events and also techniques for calculating many important characteristics of probability models.

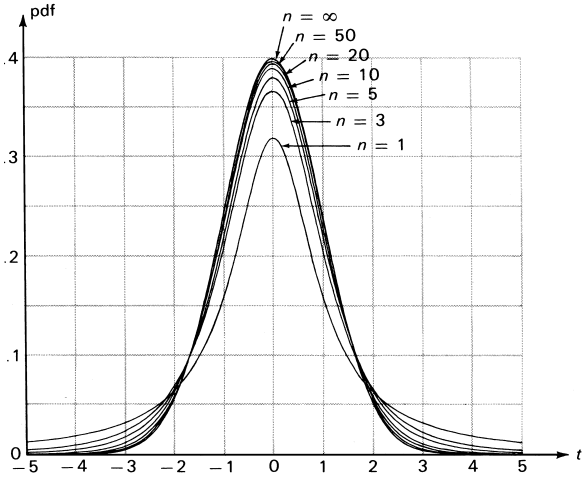


FIGURE 2 The student (n) distribution for $n = 1, 3, 5, 10, 20, 50, \infty$.

IV. STATISTICAL MODEL

A basic statistical model is a probability model with free parameters to allow a range of possibilities, one of which leads to a model that provides a reasonable approximation to a process or system being investigated. For example, in some context the normal (μ, σ^2) mentioned in Section III could be very appropriate. Also, the die example in Section III might be generalized to have probabilities $p_1, p_2, p_3, p_4, p_5, p_6$ at the six sample points 1, 2, 3, 4, 5, 6 where necessarily $p_i \geq 0$ and $\sum p_i = 1$, these being a consequence of probabilities viewed as proportions in a large aggregate.

Thus a general version of the basic model would have a space S, a collection $\mathcal{A}$ of subsets, and a probability measure $P(A;\theta)$ for each subset A in $\mathcal{A}$. The probability measure includes a free parameter θ in a space Ω that allows a range of possibilities for the measure; in an application it would be implicit that some one value of θ provided probabilities that closely approximated the behavior in the application.

The basic model as just described is concerned with frequencies or proportions on a space of possibilities. It does not distinguish component elements of structure or the practical significance of the various components. This basic model has been central to statistics throughout the 20th century.

Some more-structured models have been proposed. For example, in the measurement context one could model the measurement errors as $z_1, \ldots, z_n$ from some error distribution $f(z)$. The error distribution could be the standard normal in Fig. 2 or it could be one of the Student distributions with parameter value equal to, say, 6, which provides more realistic, longer tails. The statistical model would then be $y_i = \mu + \sigma z_i$ where μ and σ represent the response location and scaling. This alternative, more detailed model makes the process of statistical inference more straightforward and less arbitrary. For some details see Fraser (1979).

This raises the more general question of what further elements of structive from applications should reasonably be included in the statistical model. It also raises the question as to what modifications might arise in the statistical analyses using the more detailed or more specific models.

One modeling question was raised by Cox (1958). The context involved two measuring instruments, I_1 producing a measurement y that is normal (μ, σ_1^2), and I_2 producing a measurement y that is normal (μ, σ_2^2). A coin is tossed and depending on whether it comes up heads or tails we have $i = 1$ or 2 with probability 1/2 each; the corresponding instrument I_i is used to measure θ. The standard modeling view would use the model $f(i, y; \theta) = (1/2)g(y - \theta; \sigma_i^2)$, where $g(z, \sigma^2)$ designates a normal density with mean 0 and variance σ^2. In an application the data would be (i, y) with model $f(i, y; \theta)$. But clearly when the measurement is made the instrument that is used is known. This suggests that the model should be $g(y - \theta, \sigma_1^2)$ if $i = 1$ and $g(y - \theta, \sigma_2)$ if $i = 2$; this can viewed as the conditional model, given the value of the indicator i.

Traditional statistical theory would give quite different results for the first model than for the second model. A global model may be appropriate for certain statistical calculations. But in an application when one knows how the measurement of θ is obtained, it seems quite unrealistic for the purposes of inference to suggest that the instrument might have been different and that this alternative should be factored into the calculations.

V. STATISTICAL THEORY

Much of statistical theory through until the early decades of the 20th century was concerned with the analysis of data, say $\mathcal{D}$, together with a statistical model, say $\mathcal{M}$, representing the source of the data. Thus, what can we conclude from $(\mathcal{M}, \mathcal{D})$ concerning the unknown true value of the parameter, say θ, in the statistical model? This represented a substantial portion of the discipline until the mid-20th century and with ups and down remains a major portion of the core discipline.

Some major early workers in the 19th century such as Laplace promoted the idea of attaching a probability distribution, say $\pi(\theta)$, to the origins of the parameter θ. This might be something approximately uniform representing diffuse information concerning θ or something more local on certain ranges for θ as opposed to other ranges of values. This had been promoted earlier by Bayes (1763) and intermittently had periods of prominence through to the mid-20th century.

An alternative decision-theoretic approach was promoted by work of Neyman and Pearson (1933) and later generalized by Von Neumann and Morgenstern (1947) and Wald (1950). This approach viewed statistics as needing a decision function for any application; the data combined with the decision function would produce a decision concerning the unknown. This approach dominated most theoretical thought until 1955. There is of course the question of whether things can properly or should properly be mechanized in this way. There is also the issue of whether this methodological approach can produce sensible answers to the broad range of practical problems. By the mid 1950s there were substantial criticisms of the decision theory approach; in particular, there had been a major failure of the theory to produce reasonable statistical procedures for a broad range of problems.

In the mid-1950s publications by Fisher (1956) and Savage (1954) substantially altered the directions of statistics and opened wide areas for development. Fisher proposed insightful methods based on the earlier view of examining the model data combination $(\mathcal{D}, \mathcal{M})$. Savage favored the Bayesian approach emphasizing the use of personal priors to represent the latent views of the investigator concerning possible values for the parameters.

Both these directions opened new opportunities to a discipline that had become partially paralyzed by the decision-theoretic approach and by its inability to produce answers for wide-ranging problems.

VI. FOUNDATIONS

The foundations of statistics have changed and evolved with time. The early use of probability for statistical analysis was closely tied to the development of the least squares method, a widely used technique dating from Laplace and earlier. The Bayesian approach also comes from this same earlier period. Neither could be viewed at that time as an all-embracing foundation for statistics.

The decision theory approach, however, did present itself as an all-embracing theory: start with a model and a utility function and derive the optimum decision procedure for producing a decision from data. Its claims were all-emcompassing, but, as mentioned above, it failed to deliver for the broad needs of statistics.

The Bayesian approach has close ties to the decision-theoretic view but has more flexibility in allowing the personal prior or even in allowing many personal priors if many investigations are involved.

The frequentist approach as re-emphasized by Fisher (1956) reinvestigated the earlier approach of just analyzing a model together with data.

The Bayesian and frequentist approaches brought rather flexible and innovative directions to statistical theory after the rigidily of the decision theory approach. This was largely assisted by the increase in available computer

power. Both have since existed side by side with intermittent tensions and conflicts. The two approaches represent a major portion of statistics, have increasing overlap, and are growing closer in overall operation and objective.

VII. PRINCIPLES

As indicated above, the combination of a statistical model and data represents the core concern of statistics. Various methods and principles have evolved for the analysis and simplication of the model data combination.

A. Sufficiency Principle

Fisher (1922) defined a sufficient statistic $s(y)$ as a statistic with the property that the distribution of the basic variable y given $s(y)$ is independent of the parameter, say θ. The idea is that one would use the observed value of the sufficient statistic together with the marginal model, say $g(y;\theta)$, for that statistic, and otherwise ignore the original value of y. In a sense the actual value of y, given $s(y)$, can be viewed as an observation of pure error with no influence from the parameter θ, and as such can be viewed as irrelevant to any statistical assessment of θ. The *sufficiency principle* (S) is to use only the marginal model for the sufficient statistic in the inference process or assessment of data. As an example, consider a sample $y_1, \ldots, y_n$ from the normal (μ, σ_0^2) distribution: $\bar{y}$ is a sufficient statistic in the sense that the distribution of$(y_1, \ldots, y_n)$, given $\bar{y}$, does not depend on μ. Unfortunately, there are very few problems that admit a simple sufficient statistic that provides a dimension reduction such as the n to 1 found in this normal example. The concept and principle for sufficiency were widely accepted and adopted but turned out not to be available for most problems other than the familiar textbook cases. The detailed mathematical investigation of sufficiency thus unfortunately diverted statistical attention from seeking more fruitful directions for theory.

B. Likelihood Principle

Fisher (1922) introduced the concept of the likelihood function. The likelihood function $L(\theta; y^0)$ from data y^0 with model $f(y;\theta)$ is the density function $cf(y^0;\theta)$ examined as a function of θ for fixed data value. It would seem from a mathematical viewpoint to be a very obvious expression of what a function $f(y;\theta)$ would say about θ with $y = y^0$. It is usually examined relatively, that is, one θ value relative to another θ value, and this is expressed above by having an arbitrary constant c in the definition. The likelihood function $L^0(\theta) = L(\theta; y^0)$ or its logarithmic version $\ell^0(\theta) = \ell(\theta; y^0) = \log L(\theta; y^0)$ can be treated as a very informative numerical evaluation of one θ value relative to another. The *likelihood principle* (L) is to use only the observed value $\ell^0(\theta)$ of the likelihood function and not any other information from the original model–data combination. One could picture the plot of $\ell^0(\theta)$ on a monitor; the principle would say that this plot was the full allowable information concerning θ from the model–data combination. Thus, for the normal example in Section VII.A one would use only $\ell^0(\theta) = a - n(\theta - \bar{y}^0)/\sigma_0^2$ with a arbitrary and no other information from the original sample $(y_1^0, \ldots, y_n^0)$ or from the original model $\prod_1^n g(y_i - \theta; \sigma_0^2)$.

C. Weak Likelihood Principle

Sufficient statistics and the likelihood function as noted above date from Fisher's major 1922 paper. In Fisher (1925) he noted a very close connection between them: that the likelihood function as obtained from data was the best, later called minimal, sufficient statistic. Despite this asserted close equivalence the two concepts developed largely independently until the 1960s when informal discussion at meetings finally acknowledged that they were largely equivalent. Even now there is little general recognition of the equivalence and few introductory texts on mathematical statistics draw attention to the connection. The weak likelihood principle is to use only the observed likelihood function and the statistical model for the possible likelihood functions. Because of the link between the production of the likelihood function and the concept of minimal sufficiency it follows that the weak likelihood principle and the sufficiency principle are essentially equivalent. Note that the strong likelihood principle is to use the observed likelihood function but not any information concerning the possible likelihood function as would be given by the corresponding model.

D. Conditioning Principle

Fisher (1925) defined an ancillary statistic as one with a fixed distribution, free of the parameter of the model. In more recent language we would view an ancillary statistic as one that presents natural variation or error present in the model. The name suggests that it is supportive for a process of inference. Fisher recommended that one should condition on the value of the ancillary and thus use the conditional model given the observed value. The *ancillary principle* is to use only the conditional model and the observed value of the conditioned variable. The ancillary principle is often called the *conditioning principle* (C). As an example suppose y_1, y_2 come from the location model $f(y - \theta)$. Then it is easily seen that $a(y_1, y_2) = y_2 - y_1$ has a fixed distribution free of θ;

Fisher called it a configuration statistic. The conditional model for, say, $\bar{y}$, given $a(y_1, y_2) = a(y_1^0, y_2^0) = a^0$ is then $cf(\bar{y} - \theta - a^0/2) f(\bar{y} - \theta + a^0/2)$ and is a simple location model with parameter θ and with observed value $\bar{y} = \bar{y}^0 = (y_1^0 + y_2^0)/2$. The corresponding p value is

$$p(\theta) = \frac{\int_{-\infty}^{\bar{y}^0} f(\bar{y} - \theta - a^0/2, \bar{y} - \theta + a^0/2)\, d\bar{y}}{\int_{-\infty}^{\infty} f(\bar{y} - \theta - a^0/2, \bar{y} - \theta + a^0/2)\, d\bar{y}}$$

and it records the probability position of the observed $\bar{y}^0$ in its distribution conditional on the known error value a^0.

E. Connections among Principles

Birnbaum (1962) discussed the sufficiency (S), likelihood (L), and conditioning (C) principles from an equivalence or set-theoretic viewpoint. He showed that S plus C implied L. This caused some major stress in the statistical community, as S and C seemed well founded, while L seemed far too strong to most statisticians. Evans *et al.* (1986) then showed that C alone implied L. There seems to be no widely held resolution among members of the statistical community concerning these links. In fact many are unaware of key aspects of the conditioning principle or the arguments that support that principle. While Fisher (1925) certainly provided the key foundational approach and later (Fisher, 1934) analyzed the location and location scale models, it was the Cox (1958) example (see Section IV) concerning the measuring instruments that triggered a re-examination of the ancillary approach to inference.

An example with nonuniqueness of the ancillary statistic may be found in Fisher (1956, p. 47); also see Basu (1964) and Buehler (1982). This uncertainty kept the conditioning approach from general acceptance; indeed, it left a noticeable taint on the concept. More recent considerations, however, have emphasized a need for there to be some objective meaning for the ancillary to be used.

Objective ancillaries are central to the analysis of location-scale and transformation models when presented in terms of an error variable. Some general discussion may be found in Fraser (1968, 1979).

VIII. LIKELIHOOD ASYMPTOTICS

Recent likelihood asymptotic analysis developed from approximation theory, in particular, from the saddlepoint work of Daniels (1954) and Lugannani and Rice (1979). This led to the analysis of large sample likelihood functions (Barndorff-Nielsen, 1986; Fraser and Reid, 1995; Fraser *et al.*, 2000). With an increasing number n of coordinates and a fixed number of parameters it is found that with high accuracy, both theoretically $O(n^{-3/2})$ and empirically, there is an ancillary with an approximately fixed distribution and with a conditional distribution with the same dimension as the parameter. Also it is found that for a scalar component parameter there is a subsequent marginal distribution that provides a p value for assessing the scalar parameter. Under mild regularity the p value is unique.

In the early development of statistical inference, sufficiency was viewed as a prime focal concept, but its availability was extremely limited. The likelihood developments growing out of approximation theory have shown what the possible forms of a large sample model are and thus lead to the appropriate statistical analysis: first reduce by conditioning using the generally overlooked ancillary approach; then marginalize to obtain a p value free of the nuisance parameters. For some recent details, see Fraser *et al.* (1999, 2000).

SEE ALSO THE FOLLOWING ARTICLES

DATA MINING, STATISTICS • MATHEMATICAL LOGIC • MATHEMATICAL MODELING • PROBABILITY • STATISTICAL ROBUSTNESS • STATISTICS, BAYESIAN • STATISTICS, MULTIVARIATE • STATISTICS, NON-PARAMETRIC

BIBLIOGRAPHY

Barndorff-Nielsen, O. E. (1986). "Inference on full and partial parameters based on the standardized signed log likelihood ratio," *Biometrika* **73,** 307–322.

Birnbaum, A. (1962). "On the foundations of statistical inference," *J. Am. Stat. Assoc.* **57,** 269–332.

Basu, D. (1964). "Recovery of ancillary information," *Sankhyya A* **21,** 3–16.

Bayes, T. (1763). "An essay towards solving a problem in the doctrine of chances," *Phil. Trans. R. Soc.* **53,** 370–418.

Buehler, R. J. (1982). "Some ancillary statistics and their properties," *J. Am. Stat. Assoc.* **77,** 581–589.

Cox, D. R. (1958). "Some problems connected with statistical inference," *Ann. Math. Stat.* **29,** 357–372.

Daniels, H. C. (1954). "Saddlepoint approximations in statistics," *Ann. Math. Stat.* **25,** 631–650.

Evans, M., Fraser, D. A. S., and Monette, G. (1986). *Can. J. Statist.* **14,** 181–199.

Fisher, R. A. (1922). "On the mathematical foundations of theoretical statistics," *Phil. Trans. R. Soc. Lond. A* **222,** 309–368.

Fisher, R. A. (1925). "Theory of statistical estimation," *Proc. Camb. Phil. Soc.* **22,** 700–725.

Fisher, R. A. (1934). "Two new properties of mathematical likelihood," *Proc. R. Soc. A* **144,** 285–307.

Fisher, R. A. (1956). "Statistical Methods and Scientific Inference," Olwer and Boyd, London.

Fraser, D. A. S. (1968). "The Structure of Inference," Wiley, New York.
Fraser, D. A. S. (1979). "Inference and Linear Models," McGraw-Hill, New York.
Fraser, D. A. S., and Reid, N. (1995). "Ancillaries and third order significance," *Utilitas Math.* **47,** 39–53.
Fraser, D. A. S., and Reid, N. (2000). "Ancillary information for statistical inference," *In* "Proceedings of the Conference on Empirical Bayes and Likelihood" (E. Ahmed and N. Reid, eds.), Springer, New York.
Fraser, D. A. S., Reid, N., and Wu, J. (1999). "Regression analysis, nonlinear or nonnormal: Simple and accurate p-values from likelihood analysis," *J. Am. Stat. Assoc.* **94,** 1286–1295.
Fraser, D. A. S., Reid, N., and Wu, J. (2000). "A simple general formula for tail probabilities for frequentist and Bayesian inference," *Biometrika* **86,** 249–264.
Lugannani, R., and Rice, S. (1980). "Saddlepoint approximation for the distribution function of the sum of independent variables," *Adv. Appl. Math.* **12,** 475–490.
Neyman, J., and Pearson, E. S. (1933). "The testing of statistical hypotheses in relation to probabilities a priori," *Proc. Camb. Phil. Soc.* **29,** 492–510.
Savage, L. J. (1954). "The Foundations of Statistics," Wiley, New York.
Von Neumann, J., and Morgenstern, O. (1947). "Theory of Games and Economics Behaviour," Princeton University Press, Princeton, NJ.
Wald, A. (1950). "Statistical Decision Functions," Wiley, New York.



Statistics, Multivariate

David L. Banks
Stephen E. Fienberg
Carnegie Mellon University

I. Multivariate Probability Theory
II. Inference for the Linear Model
III. Regression Analysis
IV. Experimental Design
V. Cluster Analysis
VI. Contingency Tables
VII. Discriminant Analysis
VIII. Time Series Analysis
IX. Statistical Software Packages

GLOSSARY

Cluster analysis To find natural subgroups among measurements $X_1, \ldots, X_p$ on different objects. Example: The objects are stars, and the measurements are temperature, mass, redshift, bolometric reading, and distance. One can also cluster the measurements to find sets of variables that give similar information, such as redshift and distance.

Contingency table analysis To determine how membership in one feature category relates to membership in another, for objects that are classified with respect to several different features. Example: The objects are subatomic particles, and the features are charge, mass (lepton, boson, or baryon), conservation of parity and spin.

Discriminant analysis To use objects sampled from several different groups, and measurements $X_1, \ldots, X_p$ on each object, to build a classification rule that categorizes future objects into the most probable group. Example: A geologist has surface data on the mineralogy and seismology of previous drilling sites, and knows which sites produced oil. She wants to use surface data to decide whether a new site is likely to prove out.

Experimental design To determine whether different experimental treatments yield observations with different mean values. Example: One observes the lifespans of automobile tires, and the experimental treatments are manufacturer, radials versus nonradials, and weight class of the car.

Regression analysis To discover the functional relationship between a set of independent variables $X_1, \ldots, X_p$ and a dependent variable Y. Example: Y represents the breaking strength of a steel wire, and X_1 the wire's thickness, X_2 the percentage of carbon, X_3 the smelting temperature.

Time series analysis To characterize the dependence over time of sequentially observed data when neighboring values show serial correlation, and to predict future observations. Example: A meteorologist has 5 years of daily high temperature data for Gunnison, Colorado. He wants to predict tomorrow's high temperature, or model the seasonal variation in the temperature record.

MULTIVARIATE STATISTICS encompasses a diverse toolkit of methods for describing different kinds of relationships among multiple measurements. The ideas derive from the mathematical theory of probability and statistical inference. Typical multivariate procedures include cluster analysis, contingency table analysis, discriminant analysis, experimental design, regression analysis, and time series analysis. These examples illustrate some of the range of applications, but multivariate statistics includes many other topics. Even within each of these problem areas there are different approaches, depending upon the assumptions one can make about the underlying physical process.

The origins of many of the methods described in this article are rooted in work on the combination of astronomical observations and the development of the method of least squares by Pierre Simon Laplace, Adrien-Marie Legendre, and Carl Friedrich Gauss. This development occurred in the late 18th and early 19th centuries (for details, see Stigler, 1986). Systematic development of statistical theory began with work by Francis Galton, Francis Ysidro Edgeworth, Karl Pearson, and George Udny Yule during the period from 1885 to 1907. Major contributions to multivariate statistics were made in the 1920s, led by Sir Ronald Fisher, P. C. Mahalanobis, and H. Hotelling. The early applications were chiefly inspired by problems in agriculture, biology, and psychology, and methods depended strongly upon the assumption of the Gaussian (or normal) distribution for measurement noise and a linear model relating the variables of interest. As the subject has matured, the scope of application has broadened and these restrictive assumptions have been substantially weakened. Currently, a very active area is the development of computer-intensive procedures that substitute approximation for exact analytical calculation. This has enabled the extension of multivariate methods to more realistically complex problems.

I. MULTIVARIATE PROBABILITY THEORY

Before surveying the statistical theory of multivariate data, it may be useful to review some fundamental ideas from mathematical probability. These primary concepts are the distribution function, the marginal and conditional distributions, the mean and variance of a random vector, the independence of random variables, and four standard distributions: the normal (or Gaussian) distribution, the chi-squared distribution, and the multinomial distribution.

A multivariate random variable $\boldsymbol{X}$ in $\mathbb{R}^p$ is a vector whose components are random variables. The properties of that random vector are entirely determined by the corresponding multivariate distribution function $F(\boldsymbol{x})$. This distribution function is defined by

$$F(\boldsymbol{x}) = P[X_1 \leq x_1, \ldots, X_p \leq x_p],$$

so that $F(\boldsymbol{x})$ is the probability that each component of the random vector is less than or equal to the corresponding component of the argument $\boldsymbol{x}$, for all $\boldsymbol{x}$ in $\mathbb{R}^p$.

If $F(\boldsymbol{x})$ is a continuous function, then its derivative with respect to all of the components is referred to as the probability density function and is denoted by $f(\boldsymbol{x})$. If $F(\boldsymbol{x})$ is a discrete function, then its generalized derivative is referred to as the probability mass function and is also denoted by $f(\boldsymbol{x})$. The probability mass function is equal to 0 almost everywhere, but acquires increments of probability at a countable set of points in $\mathbb{R}^p$. Both the probability density function and the probability mass function are always nonnegative. Many multivariate distributions are neither completely continuous nor completely discrete, but the definition of the Riemann–Stieltjes integral handles all such cases on a common footing. For example, a basic property of any distribution function is that

$$\int_S dF(\boldsymbol{x}) = 1,$$

where $dF(\boldsymbol{x})$ is the appropriate differential form. For a continuous distribution function, this $dF(\boldsymbol{x})$ is the density; for a distribution with no continuous part, the integral is equivalent to a sum over the points where the probability mass function has increments.

In statistical applications most distributions belong to families that are indexed by a parameter vector $\boldsymbol{\theta}$ (these parameters are usually a primary object of the statistical inference). When the family is specified, one often writes $F(\boldsymbol{x}; \boldsymbol{\theta})$ to indicate the dependence upon the parameter vector.

The expected value of any function $g(\boldsymbol{X})$ of the random vector $\boldsymbol{X}$ is

$$E[g(\boldsymbol{X})] = \int_{\mathbb{R}^p} g(\boldsymbol{x})\, dF(\boldsymbol{x}).$$

Some specific functions $g(\cdot)$ are particularly important. For example, the mean $\boldsymbol{\mu}$ of any random vector is

$$\boldsymbol{\mu} = E[\boldsymbol{X}] = \int_{\mathbb{R}^p} \boldsymbol{x}\, dF(\boldsymbol{x}).$$

This quantity is the average value of the random vector. Also, the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$ is

$$\boldsymbol{\Sigma} = E[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^{\mathrm{T}}] = \int_{\mathbb{R}^p} (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\, dF(\boldsymbol{x})$$

where the superscript T denotes the transpose. The ith diagonal entry of $\boldsymbol{\Sigma}$ is

$$\sigma_i^2 = \mathrm{var}[X_i] = \int_{-\infty}^{\infty} (x_i-\mu_i)^2\, dF(x_i),$$

which is the variance of the ith component of $\boldsymbol{X}$ considered as a univariate random variable. The square root of this variance is referred to as the standard deviation of the random variable X_i. Also, the off-diagonal entries of $\boldsymbol{\Sigma}$,

$$\begin{aligned}\sigma_{ij} &= \mathrm{cov}[X_i, X_j] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i-\mu_i)(x_j-\mu_j)\, dF(x_i, x_j) \quad \text{for} \quad i \neq j,\end{aligned}$$

are the covariances between the ith and jth components of $\boldsymbol{X}$. Heuristically, large values of the diagonal entries of $\boldsymbol{\Sigma}$ indicate that the corresponding component varies greatly. Large positive values of the off-diagonal entries suggest that the two corresponding components tend to be large or small together, and extreme negative values suggest that one component is large when the other is small, and conversely.

The marginal distribution of a particular component X_r of a random vector $\boldsymbol{X}$ is just the univariate distribution of that random component, ignoring everything else in the vector. One can obtain the marginal distribution by integrating out all other components; i.e.,

$$\begin{aligned}F(x_r) = \int_{-\infty}^{x_r}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} dF(\boldsymbol{y})\, dy_1 \ldots dy_{r-1} \\ \times\, dy_{r+1}\ldots dy_p\, dy_r.\end{aligned}$$

Omitting the first integration gives the density function or the probability mass function, as appropriate. If one integrates out a smaller set of components, then one obtains the distribution of the remaining components.

From the distribution function of a random vector, one can find the conditional distribution of the first r components $X_1, \ldots, X_r$, given the values of the subsequent $p-r$ components, i.e., $X_{r+1}=x_{r+1}, \ldots, X_p = x_p$. Intuitively, one acts as though one has seen the last $p-r$ components of the random vector and wants to use these to infer the probability structure of the first r components. In order to reflect the additional information now available, one calculates the distribution obtained after conditioning upon the observed values, obtaining

$$\begin{aligned}&F(x_1, \ldots, x_r \mid x_{r+1}, \ldots, x_p) \\ &= \frac{\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_p} f(y_1, \ldots, y_r, x_{r+1}, \ldots, x_p)\, dy_1 \ldots dy_r}{g(x_{r+1}, \ldots, x_p)}.\end{aligned}$$

Here $g(x_{r+1}, \ldots, x_p)$ denotes the joint density or mass function evaluated at the fixed values of the conditioning variables.

If $F(\boldsymbol{x})$, the joint distribution function of $\boldsymbol{X}$, can be written as the product of all of the univariate marginal distributions, then the components of $\boldsymbol{X}$ are said to be independent. One implication of independence is that observing any subset of the components of $\boldsymbol{X}$ provides no information about the values of the unobserved components. In practical applications, it is usually desirable that one's sample data $\boldsymbol{X}$ consist of independent observations; in most cases this maximizes the amount of information gathered, and also enables simpler mathematical analysis.

The concept of independence generalizes to conditional independence. Suppose that the random vector $\boldsymbol{X}$ is partitioned into three disjoint subsets, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$. Then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are said to be conditionally independent given $\boldsymbol{X}_3 = \boldsymbol{x}_3$ if

$$F(\boldsymbol{x}_1, \boldsymbol{x}_2 \mid \boldsymbol{x}_3) = F(\boldsymbol{x}_1 \mid \boldsymbol{x}_3) F(\boldsymbol{x}_2 \mid \boldsymbol{x}_3).$$

Regarding $\boldsymbol{X}_1$, this implies that

$$F(\boldsymbol{x}_1 \mid \boldsymbol{x}_2, \boldsymbol{x}_3) = F(\boldsymbol{x}_1 \mid \boldsymbol{x}_3).$$

Heuristically, all of the conditional information about $\boldsymbol{X}_1$ is carried by $\boldsymbol{x}_3$, so that $\boldsymbol{X}_2$ is entirely redundant. An exactly symmetric relationship holds for $F(\boldsymbol{x}_1 \mid \boldsymbol{x}_2, \boldsymbol{x}_3)$.

The normal (or Gaussian, in honor of Carl Friedrich Gauss, who used it in developing the method of least squares) distribution has unique properties that make it the natural basis for much of multivariate statistical inference. For example, the Central Limit Theorem ensures that under very general conditions, as n increases, the average $\bar{\boldsymbol{X}}$ of a multivariate sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from an arbitrary distribution $F(\boldsymbol{x})$ tends toward a random variable that has multivariate normal distribution with mean $\boldsymbol{\mu} = E[\boldsymbol{X}_1]$ and variance $\boldsymbol{\Sigma} = (1/n)\mathrm{Var}[\boldsymbol{X}_1]$. Also, on empirical grounds, the multivariate normal distribution offers a very adequate description of many kinds of chance error, such as those that inevitably occur in a measurement process.

The normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The normal distribution function has no closed form; however, the probability density function of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{x})\right],$$

where the matrix $\boldsymbol{\Sigma}$ is positive definite (i.e., invertible, with all eigenvalues positive). Numerical integration is needed to determine the distribution function.

Suppose one partitions the random variable $\boldsymbol{X}$ into two subvectors, $\boldsymbol{X}_1$ of length r and $\boldsymbol{X}_2$ of length $p-r$. This induces a corresponding partition on the mean vector and the variance matrix; thus

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Here $E[\boldsymbol{X}_1] = \mu_1$ is an $r \times 1$ vector, $\text{var}[\boldsymbol{X}_1] = \boldsymbol{\Sigma}_{11}$ is a $r \times r$ matrix, and analogous results hold for $\boldsymbol{X}_2$. All information about the covariance between components in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ appears in $\boldsymbol{\Sigma}_{12}$, an $r \times (p-r)$ matrix with the property that $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^{\mathrm{T}}$.

If $\boldsymbol{X}$ is $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the marginal distribution of $\boldsymbol{X}_1$ is $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$. Similarly, the conditional distribution of the first r components of $\boldsymbol{X}$, or $\boldsymbol{X}_1$, given that the last $p-r$ components are equal to $\boldsymbol{x}_2$, is multivariate normal with mean $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$, where

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$$
$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

Also, $\boldsymbol{X}_1$ is independent of $\boldsymbol{X}_2$ if and only if $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$. Finally, if $\boldsymbol{X}$ is partitioned into three disjoint sets of components, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$, inducing a conformable partition of the inverse covariance matrix $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1}$, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are conditionally independent given $\boldsymbol{X}_3$ if and only if $\boldsymbol{K}_{12} = \boldsymbol{0}$. These four properties of the normal distribution are extremely powerful, and have enabled much of the development of the theory of multivariate analysis. For simplicity, this discussion has spoken as though the partitioning divides the vector into subvectors of consecutive components, so that, for example, $\boldsymbol{X}_1$ consists of the first r components of $\boldsymbol{X}$. Actually, with more formal expression, the sense of the results holds when $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$ are any collections of the components of $\boldsymbol{X}$.

The chi-squared distribution is also important to multivariate inference. This is derived from the normal distribution by letting $X_1, \ldots, X_p$ be a sample of independent random variables, each $\mathrm{N}(0, 1)$, and defining the new random variable $Y = \sum_{i=1}^{p} X_i^2$. This new random variable has the chi-squared distribution with p degrees of freedom; the degrees of freedom is the parameter that indexes the family of chi-squared distributions. One can show that the mean of Y is p and the variance of Y is $2p$. Unlike the normal distribution, the chi-squared distribution is not symmetric; in particular, $f(y) = 0$ for all $y < 0$.

Using chi-squared random variables, one can define the F-distribution, which is crucial to the analysis of experimental designs (see Section III). If Y_1 is a chi-squared random variable with p_1 degrees of freedom and if Y_2 is an independent chi-squared random variable with p_2 degrees of freedom, then

$$F = \frac{Y_1/p_1}{Y_2/p_2}$$

follows an F-distribution with p_1, p_2 degrees of freedom. The F-distribution is thus characterized by two parameters, just as the univariate normal distribution is characterized by the mean and the variance.

A multinomial random vector $\boldsymbol{X}$ has components that are constrained to be integers between 0 and n, and further constrained so that the sum of the components must equal n. If one randomly tosses n balls into p boxes so that the probability that a ball falls in box i is θ_i, and if X_i denotes the number of marbles in the ith box, then the vector $\boldsymbol{X}$ with these components has a multinomial distribution with parameters n and $\boldsymbol{\theta}$. This implies that the distribution is discrete, and the probability mass function is

$$f(\boldsymbol{x}) = \binom{n}{x_1, \ldots, x_p} \theta_1^{x_1} \ldots \theta_p^{x_p}$$

for $\boldsymbol{x}$ satisfying the constraints indicated above, and where $\sum \theta_i = 1$.

The mean of the multinomial distribution is $\boldsymbol{\mu} = n\boldsymbol{\theta}$, and the entries of the covariance matrix are $\sigma_i^2 = n\theta_i(1-\theta_i)$ and $\sigma_{ij} = n\theta_i\theta_j$. The marginal distribution of X_r reduces to a multinomial distribution with $p = 2$ and parameters n, θ_r, and $1 - \theta_r$; this is also known as the binomial distribution. Similarly, the marginal distribution of any subset of the components is also multinomial, with parameters that are simple functions of the parameters of the original multinomial. The conditional distribution of $X_1, \ldots, X_r$, given $X_{r+1} = x_{r+1}, \ldots, X_p = x_p$, is an r-variate multinomial with parameters n^* and $\boldsymbol{\theta}^*$, where

$$n^* = n - \sum_{j=r+1}^{p} x_i$$
$$\theta_i^* = \frac{\theta_i}{\sum_{j=r+1}^{p} \theta_j}.$$

The multinomial distribution is particularly important in analyzing categorical data, such as arise in the form of cross-classifications or contingency tables as described Section V.

II. INFERENCE FOR THE LINEAR MODEL

The classic linear model in statistics assumes that observations are linearly related to the explanatory variables and measured with independent, normally distributed error. Thus the ith observation is written in terms of the j explanatory variables as

$$y_i = \theta_0 + \sum_{j=1}^{p} \theta_j X_{ij} + \epsilon_i$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent random variables, with normal distribution having mean 0 and variance σ^2, written as $N(0, \sigma^2)$. Vector notation compactly expresses this relationship for a set of n observations:

$$\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of observations, $\boldsymbol{X}$ is $n \times (p+1)$ matrix of explanatory variables, $\boldsymbol{\theta}$ is a $(p+1) \times 1$ parameter vector, and $\boldsymbol{\epsilon}$ is a random vector with multivariate normal distribution $N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, where $\boldsymbol{0}$ indicates that the mean of each error is 0 and $\boldsymbol{I}$ is the identity matrix (a diagonal matrix with all diagonal entries equal to 1).

If the covariance matrix does not have the form $\sigma^2 \boldsymbol{I}$, but is known to equal $\boldsymbol{\Sigma}$, a specific positive-definite matrix, then one is working in the context of the general linear model. Whenever possible, the following discussion will include such extensions.

This framework enables the mathematical theory that underlies the statistical methods of estimation and hypothesis testing. Although real data rarely conform exactly to these stringent assumptions, the procedures developed from the linear model are usually quite robust. Even when the model assumptions fail entirely, the nature of the failure can often indicate how the experiment or the analysis must be modified to achieve the desired research goals.

A. Point Estimation

A fundamental estimation strategy is to minimize a measure of discrepancy between observed values and corresponding predicted values. If $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, then a tractable discrepancy is the sum of the squared differences, and the method is referred to as ordinary least squares estimation (the method of Laplace, Legendre, and Gauss). More broadly, weighted least squares extends the approach to arbitrary $\boldsymbol{\Sigma}$ by minimizing the sum of the squared generalized distances $(\boldsymbol{y} - \boldsymbol{X\theta})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X\theta})$. Mahalanobis (1936) used this metric to take account of the covariance structure of the data. Matrix calculus finds the solution to

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} (\boldsymbol{y} - \boldsymbol{X\theta})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X\theta}) \tag{2}$$

as

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}, \tag{3}$$

where $\hat{\boldsymbol{\theta}}$ is not unique unless $\boldsymbol{X}$ is of full rank and $\boldsymbol{\Sigma}$ is positive definite.

Another estimation strategy maximizes the likelihood of the observations. The likelihood function of the data is the joint probability density function, except that the observations are known quantities and parameters are treated as variables. For the general linear model, the likelihood function is

$$L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\Sigma}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-n/2} \times \exp\left[-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{X\theta})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X\theta}) \right].$$

For this model, the weighted least squares estimate and the maximum likelihood estimate agree. The likelihood function is maximized at $\hat{\boldsymbol{\theta}}$ of (3).

A third strategy is to find the best invariant (or equivariant) estimator. These estimators minimize the average error penalty within a class of estimators invariant under a specified group of transformations. A typical penalty function (or loss function) is the weighted squared error; i.e., for an arbitrary positive-definite matrix $\boldsymbol{W}$, the penalty for a numerical estimate $\boldsymbol{z}$ when the true value is $\boldsymbol{\theta}$ is

$$\mathrm{L}(z, \theta) = (z - \theta)^{\mathrm{T}} \boldsymbol{W} (z - \theta). \tag{4}$$

A typical group is the set of affine transformations of the data under the composition operation. For these choices, the estimate automatically adjusts to affine changes in the units of measurement and turns out to be $\hat{\boldsymbol{\theta}}$, as before.

For the linear model with normal error, all three estimation methods give the same result for $\boldsymbol{\theta}$, although they give somewhat different estimates for the error variances. For example, suppose $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{A}$, where $\boldsymbol{A}$ is known but σ^2 must be estimated. This simple change does not affect $\hat{\boldsymbol{\theta}}$ since σ^2 cancels in (3); however, the method of least squares finds

$$\hat{\sigma}^2_{\mathrm{LS}} = \frac{1}{n - p - 1} (\boldsymbol{y} - \hat{\boldsymbol{\theta}} \boldsymbol{X})^{\mathrm{T}} \boldsymbol{A}^{-1} (\boldsymbol{y} - \hat{\boldsymbol{\theta}} \boldsymbol{X}), \tag{5}$$

while the maximum likelihood procedure finds

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{n - p - 1}{n} \hat{\sigma}^2_{\mathrm{LS}},$$

and for the natural loss function for an estimate z of σ^2 given by

$$\mathrm{L}(z, \sigma^2) = \left(\frac{z}{\sigma^2} - 1 \right)^2$$

and the group of affine transformations, the best invariant estimate is

$$\hat{\sigma}^2_{\mathrm{BIE}} = \frac{n - p - 1}{n - p + 1} \hat{\sigma}^2_{\mathrm{LS}}.$$

Finally, suppose that $\boldsymbol{\Sigma}$ is entirely unknown; in this case, the maximum likelihood estimator is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^{\mathrm{T}}.$$

This estimator is biased; i.e., $E[\hat{\boldsymbol{\Sigma}}] \neq \boldsymbol{\Sigma}$. An unbiased estimator as

$$\boldsymbol{S} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}. \tag{6}$$

These examples show that as one relaxes the assumptions of the classical model, the estimation strategies produce substantially different estimators of the error variance. With greater departures, one can even find cases in which the estimates of the vector $\boldsymbol{\theta}$ will also differ according to the method employed.

A fourth criterion for estimation is admissibility. An estimator is inadmissible if there exists another estimator with smaller average penalty for all parameter values. Stein (1956) showed that when the parameter vector $\boldsymbol{\theta}$ contains more than two components, then $\hat{\boldsymbol{\theta}}$ is inadmissible under the loss function defined in expression (4); James and Stein (1961) found an estimator which is always slightly better than $\hat{\boldsymbol{\theta}}$. This remarkable result has had great theoretical impact, but has had little influence on applied multivariate statistics.

Lehmann and Casella (1998) give a technical survey of the various estimation and testing strategies in modern statistics and describe their merits and pitfalls from a frequentist inferential perspective.

B. Hypothesis Testing

The linear model provides a venue for exploring broad issues in the testing of parametric hypotheses. A test begins with the formulation of a null and an alternative hypothesis regarding the unknown parameter $\boldsymbol{\theta}$. Formally, one tests

$$\mathrm{H}_0\colon \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \qquad \text{versus} \qquad \mathrm{H}_\mathrm{a}\colon \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_\mathrm{a}$$

where $\boldsymbol{\Theta}_0$ is the region of the parameter space conforming to the null hypothesis and $\boldsymbol{\Theta}_\mathrm{a}$ is the region conforming to the alternative, where $\boldsymbol{\Theta}_0 \cap \boldsymbol{\Theta}_\mathrm{a} = \emptyset$. In some circumstances, $\boldsymbol{\Theta}_\mathrm{a}$ is chosen to include all other possible values of the parameter, i.e., $\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_\mathrm{a} = \boldsymbol{\Theta}$, for $\boldsymbol{\Theta}$ the entire set of allowable values. Informally, the alternative hypothesis is either the situation that the investigator wants to show, or the conclusion that leads to a change in the status quo.

The second step involves the choice of α, the probability of rejecting H_0 when H_0 is true, and either n, the sample size, or $\beta(\boldsymbol{\theta}_1)$, the probability of failing to reject H_0 for a specific value $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_\mathrm{a}$. For a particular statistical test, any two of α, n, and $\beta(\boldsymbol{\theta}_1)$ determine the third. Customary values for α are .05 and .01, whereas $\beta(\boldsymbol{\theta}_1)$ is often allowed to be much larger. This reflects a common view that Type I error (false rejection of H_0) is more serious than Type II error (failure to correctly reject H_0). The power of a statistical test is $1 - \beta(\boldsymbol{\theta}_1)$, the probability of rejecting the null hypothesis when the alternative is true. To obtain tests that are sufficiently powerful when $\alpha = .05$, a common rule of thumb is to take $n = 20d$, where d is the dimension of $\boldsymbol{\Theta}_\mathrm{a}$ (Tabachnick and Fidell, 1989).

The third step calculates the value of a test statistic, giving a single number summary of the information in the sample that bears on the hypotheses of interest. There are many strategies for selecting a test statistic; one of the most valuable is the likelihood ratio test statistic

$$\mathrm{L}(\boldsymbol{X}) = \frac{\sup\{f(\boldsymbol{X};\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}}{\sup\{f(\boldsymbol{X};\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}}, \tag{7}$$

where $f(\boldsymbol{X};\boldsymbol{\theta})$ is the joint probability density function evaluated at the data for a particular value of $\boldsymbol{\theta}$. For simple applications, such as when $f(\boldsymbol{X};\boldsymbol{\theta})$ is the density of the normal distribution, one can calculate exactly the distribution of the test statistic when H_0 is true. Otherwise, one needs to rely on the result that, under the null hypothesis, $2\log \mathrm{L}(\boldsymbol{X})$ is asymptotically (as the sample size n increases to ∞) distributed as a chi-squared random variable with $\dim\{\boldsymbol{\Theta}\} - \dim\{\boldsymbol{\Theta}_0\}$ degrees of freedom, provided that some reasonable conditions obtain.

The fourth step compares the observed value of the test statistic to the distribution of all possible values when H_0 is true. The test rejects H_0 if the statistic lies in a "rejection" region determined by α. Most software packages calculate the p value or significance probability of the observed value of the test statistic, which represents the smallest value of α for which H_0 would be rejected. One can interpret the significance probability as the probability of obtaining a value of the test statistic that is as extreme or more extreme than the sample value actually observed when the null hypothesis is true. Thus, if the p value were $p = .04$, then the test rejects H_0 at level $\alpha = .05$ but not at level $\alpha = .01$. A common error in working with significance probabilities is to regard p as the probability that the null hypothesis is correct; this is a major mistake that can distort the entire analysis and the inferences one draws from it.

C. Confidence Regions

The linear model also motivates procedures for the construction of confidence regions. A $100(1-\alpha)\%$ confidence region in $\boldsymbol{\Theta}$ is a random volume that has probability $1-\alpha$ of containing the true value $\boldsymbol{\theta}$. Customarily, one considers 90%, 95%, or 99% confidence regions.

Once the data have been gathered and the region calculated, it is incorrect to speak of coverage probability (since $\boldsymbol{\theta}$ is surely either inside or outside); instead, one reports that the region contains $\boldsymbol{\theta}$ with $100(1-\alpha)\%$ confidence. This means that if one used the same confidence region

procedure a very large number of times for multivariate normal situations, then the long-run proportion of times that the confidence regions actually contain the true parameter values approaches $1-\alpha$. This is a frequentist interpretation, and contrasts with the statistical philosophy of Bayesian inference, described in the following subsection. In Bayesian inference the analogue of the confidence region is the credible region, in which one appropriately refers to the probability that the region contains the parameter.

As the notation suggests, there is a duality between confidence regions and hypothesis tests. A $100(1-\alpha)$ confidence region contains a point $\boldsymbol{\theta}_0$ if and only if a test of

$$\mathrm{H}_0: \quad \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad \text{versus} \qquad \mathrm{H}_\mathrm{a}: \quad \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

fails to reject the null hypothesis at level α.

In this and most other multivariate normal applications, the confidence region is an ellipsoid with axes determined by the covariance matrix of the parameter estimates and center $\hat{\boldsymbol{\theta}}$. The formula for the confidence region on $\boldsymbol{\theta}$ is

$$\left\{\boldsymbol{\theta} \in \boldsymbol{\Theta}: (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{X}[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \leq F_{p+1,n-p-1,\alpha}\right\}, \tag{8}$$

where $F_{p+1,n-p-1,\alpha}$ is the $(1-\alpha)$-percentile point obtained from an F-distribution with $p+1$ and $n-p-1$ degrees of freedom (the appropriateness of the F-distribution for this application is a standard exercise in elementary probability theory). If the inverse matrix is not uniquely defined, then an entire subspace of $\boldsymbol{\Theta}$ satisfies the inequality. Often $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ with σ^2 unknown. Here one can estimate σ^2 from the data by $\hat{\sigma}^2$ as in (5), replace $\hat{\sigma}^2 \boldsymbol{I}$ for $\boldsymbol{\Sigma}$ in (8), and use the critical value $F_{p+1,n-p-2,\alpha}$.

D. Bayesian Inference

The inferential paradigm described above for point estimation, hypothesis testing, and confidence regions is usually referred to as the frequentist approach to inference, and it is not universally accepted. Many statisticians prefer using Bayesian inference to draw conclusions from data, expressed directly in terms of probabilities for the parameters. The frequentist view defines the probability of an event as the proportion of times that the event occurs in a sequence of possibly hypothetical trials; the Bayesian defines the probability of the same event in terms of the formalized uncertainty regarding its occurrence, based on an *a priori* assessment of θ (i.e., a prior distribution over Θ). This in some sense removes the distinction between parameters and random variables. One implication of this prior assessment is that different Bayesian statisticians, with different states of uncertainty regarding θ, may assign different probabilities to the same event. Frequentists can also disagree, but their differences are attributable to the selection of different probability models for describing the event. Perhaps the most striking aspect of this often subtle distinction is that Bayesians view unknown parameters as random variables, whereas frequentists regard them as fixed constants. This allows Bayesians to calculate probabilities associated with the parameters, given the data, whereas frequentists can only look at probabilities of the data, given values of the parameters.

The Bayesian method first requires one to formalize the uncertainty about the unknown parameters by placing a probability distribution $\pi(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}$. This prior distribution reflects beliefs about, among other things, the most probable and least probable values of $\boldsymbol{\theta}$; usually these beliefs derive from common sense and previous experience with similar data, but those who are unable to express such beliefs or are very cautious may resort to a "noninformative" prior. Bayesians often assess their own experience or they elicit probabilities from appropriate experts. The choice of a prior distribution in a particular application involves technical issues.

After selecting the prior distribution for $\boldsymbol{\theta}$, one gathers the data and uses Bayes' theorem to update the prior distribution in the light of the evidence:

$$\pi^*(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta}) f(\boldsymbol{X};\boldsymbol{\theta})}{\int \pi(\boldsymbol{\psi}) f(\boldsymbol{X};\boldsymbol{\psi})\, d\boldsymbol{\psi}}.$$

This posterior distribution for $\boldsymbol{\theta}$, $\pi^*(\boldsymbol{\theta})$, is the basis for subsequent inference.

The Bayesian method illustrates how prior beliefs should change as a result of observing the data. Even if two Bayesians disagree markedly upon their choice of prior distributions, observing the same set of data typically causes them to produce posterior distributions which are more similar, and continued observation eventually compels their conclusions to converge as long as they do not assign zero probability *a priori* to different parts of the parameter space $\boldsymbol{\Theta}$ (Blackwell and Dubins, 1962).

Bayesian inference is closely linked to decision theory, which is concerned with minimizing expected losses when one must make choices without certainty as to the consequences. Rather than detail the impact of this theory on Bayesian estimation and the Bayesian analogues of hypothesis testing and confidence regions, we simply note that common Bayes estimators for $\boldsymbol{\theta}$ are the mode and the mean of the posterior distribution. Similarly, the Bayesian analogue to hypothesis testing is performed by integrating over regions of $\boldsymbol{\Theta}$ corresponding to H_a or H_0. This enables the Bayesian to state the posterior probabilities of the null and alternative hypotheses. Finally, a Bayesian analogue to a confidence region is the credible

region, which is typically the smallest volume in Θ that contains a specified proportion of the posterior probability. Bernardo and Smith (1994) provide a detailed exposition of Bayesian theory.

E. Generalized Linear Model

The generalized (as distinct from general) linear model (GLM) was developed by Nelder and Wedderburn (1972) in an extension of the classical linear model. It unifies general linear model theory, log-linear models, logit and probit models, and more exotic models developed for specific applications. This enables application of standard procedures for estimation, model verification, hypothesis testing, and prediction in a variety of linear and nonlinear problems.

The form of the GLM is

$$\boldsymbol{Y} = g(\boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\epsilon} \tag{9}$$

where $\boldsymbol{Y}$ is the observed vector, $g(\cdot)$ is a known function, referred to as the link function, and the errors ϵ_i are (usually) independent, have mean 0, and belong to a known exponential family distribution. [Exponential family distributions include the normal, binomial, Poisson, inverse Gaussian, and many other distributions; Barndorff-Neilsen (1978) gives a detailed account of the central place of these in statistical inference.] The exponential family condition implies that the density of a single observation Y_i can be written in the form

$$f(y_i) = \exp\left\{\frac{y_i\theta_i - h(\theta_i)}{\psi} + a(y_i, \psi)\right\}, \tag{10}$$

where θ_i is referred to as the natural parameter and ψ is the dispersion parameter. One can show that

$$\frac{dh}{d\theta_i} = E\left[\beta_0 + \sum_{j=1}^{p} \beta_j X_i j\right]$$

and thus in most applications the natural choice for the link function (called the canonical link) takes $g = h'$. McCullagh and Nelder (1989) offer many examples, using both canonical and noncanonical link functions.

When the link function is the identity and one has independent and identically distributed normal errors with mean 0, then the model exactly agrees with the classical model given in (1). When the Y_i are proportions related to $\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ by the model given in (19), then the link function is $\log[p/(1-p)]$ and the errors are binomial; this corresponds to the logistic regression model described in the Section II. Finally, when the data are counts obtained from a Poisson distribution and the link function is $\log \mu_i$, one has the log-linear model described in Section V.

Maximum likelihood is usually the method of choice for estimating the parameters of a GLM. Hypothesis testing is typically done from likelihood ratio test principles, using a quantity known as the deviance, which can be partitioned into components and represented in an analysis of deviance table that is analogous to the analysis of variance table discussed in Section III. Most software packages and discussions of GLMs define the deviance as $-2\sum \log f(Y_i)$, where $f(Y_i)$ has the exponential family form given in (10). Distribution theory for the deviance is generally unavailable unless the dispersion parameter ψ in expression (10) is known or well estimated, as is the case for the linear model in (1). This makes formal inference difficult in many applications.

F. The Bootstrap

The inferential methods described thus far depend upon strong model assumptions and simple parameters for success. In practice, especially in multivariate problems, one rarely has accurate knowledge of the underlying distribution and fundamental parameters are complicated functionals of the distribution. Modern computer-intensive statistical methods can now substitute large-scale calculation for theoretically convenient assumptions, and this approach has greatly extended the range of multivariate applied statistics. One of the most important tools in this extension is the bootstrap, developed by Efron (1979).

The bootstrap method enables one to estimate standard errors and set confidence regions in very general situations. The key idea is that if the empirical cumulative distribution function (ECDF) of the data is approximately equal to the unknown true cumulative distribution function, then confidence regions that are constructed as if the ECDF were the true distribution are similar to the exact confidence regions that would have been set if the true distribution function were known. The same insight holds for estimating standard errors, which are the standard deviations of parameter estimates.

The empirical distribution function of a multivariate sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is

$$\hat{F}_n(\boldsymbol{x}) = \sum_{i=1}^{n} I_{A(\boldsymbol{x})}(\boldsymbol{X}_i), \tag{11}$$

where

$$A(\boldsymbol{x}) = \bigcup_{j=1}^{p} (-\infty, x_j]$$

and

$$I_{A(\boldsymbol{x})}(\boldsymbol{X}_i) = \begin{cases} 1 & \text{if} \quad \boldsymbol{X}_i \in A(\boldsymbol{x}) \\ 0 & \text{otherwise.} \end{cases}$$

The Glivenko–Cantelli theorem from probability theory ensures that $\lim_{n\to\infty} \hat{F}_n = F$ for univariate random variables. Special technical conditions are needed to guarantee

large sample convergence for multivariate ECDFs. Nonetheless, the cases in which convergence fails are somewhat pathological, and most applications may safely neglect this potential difficulty. Thus, if the sample size is sufficiently large, then $\hat{F}_n$ should offer a very good approximation to the unknown F.

If F were known, then for any parameter $\boldsymbol{\theta}$ one could estimate the accuracy of an estimator by drawing a very large number m of samples of size n from F, calculating the parameter estimate of interest for each, and comparing the m estimates to the known value $\boldsymbol{\theta}$. The distribution of these parameter estimates is called the sampling distribution, and knowing it and the true $\boldsymbol{\theta}$ entirely determines the reliability of any inference. The bootstrap method assumes that the known $\hat{F}_n$ is a good estimate of F, and draws m samples of size n from $\hat{F}_n$. One calculates the estimate for each sample and uses these estimates to define the bootstrap distribution. As m increases, the bootstrap distribution converges very generally to the sampling distribution; since the parameter of interest is exactly known (at least implicitly) for the ECDF, then one can use these as proxies for the sampling distribution and θ when assessing the reliability of an inference.

Operationally, the algorithm for setting the simplest bootstrap confidence region on a parameter $\boldsymbol{\theta}(F)$ is as follows:

1. Draw a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from the unknown distribution function $F(\boldsymbol{x})$.
2. Obtain the estimate $\hat{\boldsymbol{\theta}}$ from the sample as $\boldsymbol{\theta}(\hat{F}_n)$, where $\hat{F}_n$ is the ECDF defined in (11) and $\Theta(\cdot)$ is a functional mapping distribution functions into $\boldsymbol{\Theta}$.
3. Draw a large number m of random samples of size n from $\hat{F}_n$; we call these resamples, to distinguish them from the original data. For each resample determine the corresponding secondary ECDF; denote these by $\hat{\hat{F}}_{n1}, \ldots, \hat{\hat{F}}_{nm}$.
4. Calculate $\boldsymbol{\theta}_j = \boldsymbol{\theta}(\hat{\hat{F}}_{nj})$. Let $G(\boldsymbol{\theta})$ be the ECDF of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$.
5. Use $G(\boldsymbol{\theta})$ as though it were the sampling distribution of $\boldsymbol{\theta}$ for all inferences. Thus the bootstrap estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}$ is just the covariance matrix of a random variable with distribution $G(\boldsymbol{\theta})$. Also, the bootstrap confidence region for $\boldsymbol{\theta}$ at level α is just the smallest set in $\mathbb{R}^p$ such that the probability mass of that region under $G(\boldsymbol{\theta})$ equals $1-\alpha$.

This algorithm may look complex, but it is actually very simple to implement in the form of a computer program. For estimating the standard error of a univariate parameter, it is usually sufficient to take $m = 200$; for setting a confidence region, one can take $m = 1000$. The accuracy of the bootstrap increases with the sample size n, but surprisingly good inference for a single parameter has been obtained with samples as small as 15.

More complicated bootstrap algorithms attain improved accuracy in many large-sample situations, often at the expense of increased computation. One of the easiest methods to improve bootstrap performance is to bootstrap a studentized pivot. For a univariate parameter, instead of calculating the bootstrap distribution of an estimator $\hat{\theta}$, one finds the bootstrap distribution that approximates the sampling distribution of $(\hat{\theta} - \theta)/\sqrt{\mathrm{Var}(\hat{\theta})}$. This approximation is asymptotically superior to the simple bootstrap distribution described above whenever the parameter of interest is asymptotically normally distributed. Confidence regions can be built from this pivot in a straightforward way, and the method immediately generalizes to multivariate parameters.

Other bootstrap refinements are detailed by Efron and Tibshirani (1986); Banks (1989) surveys bootstrap theory and applications. An explicit study of bootstrap performance for a large class of multivariate inference problems is described by Beran and Srivastava (1985).

G. Cross-Validation

Just as computer-intensive methods permit bootstrap confidence regions for very general problems, so does modern computing enable flexible assessment of model adequacy. A key tool for assessment is cross-validation; the implementation of this method must be tailored to the problem, but the principle applies to a tremendous range of statistical techniques. This generality is important, since in realistic applications one is usually fitting a model that only approximates the complexity of the situation. In particular, cross-validation has become the primary method for fitting spline models to complex data (Wahba, 1985) and for determining the smoothing parameters in nonparametric curve-fitting (Hastie and Tibshirani, 1990).

Before using the results of a statistical analysis, one must know how well the fitted model will describe future values of the dependent variables. This cannot be done with confidence by simply assessing the agreement between the original sample values and the model's predictions of those values; the sample was used to estimate the parameters of the model, which leads to spuriously good agreement. To adjust for this optimism, cross-validation sets aside a portion of the data, fits the hypothesized model to the remainder, and assesses the predictive accuracy against the unused data. If one then interchanges the roles of the two subsets of the data, one obtains a very efficient estimate of model adequacy. More elaborate methods for dividing the sample lead to additional gains in assessment accuracy.

Let $\mathcal{M}(\theta)$ denote the model that is assumed for the analysis. In many applications this will be the model given in (1), but in other cases it could be a much more complex form, such as (9) or (20), discussed in Section II. The cross-validation strategy proceeds as follows:

1. Partition the sample at random into k disjoint subsamples. Denote these sets by $S_1, \ldots, S_k$.
2. For $i = 1$, pool all subsamples with subscripts not equal to i. Use these pooled subsamples to estimate the parameters in the model $\mathcal{M}(\theta)$. Let $\mathcal{M}(\hat{\theta}_i)$ be the model fitted without use of subsample S_i.
3. Use the model $\mathcal{M}(\hat{\theta}_i)$ to estimate the dependent variable values in subsample S_i; record these estimates as $\hat{\boldsymbol{Y}}_i$.
4. If $i \neq k$, set $i = i + 1$ and return to step 2.
5. Link $\hat{\boldsymbol{Y}}_1, \ldots, \hat{\boldsymbol{Y}}_k$ into a vector of estimated values $\hat{\boldsymbol{Y}}$. Compare this to the observed dependent variable values $\boldsymbol{Y}$ from the sample, using whatever criterion for aggreement is most appropriate to the problem.

If one has a small dataset and is fitting a simple model, then one can partition the sample into as many subsamples as there are observations; if the dataset is large or computational time is expensive, then a coarser partition may be more useful.

In order to compare several competing models via cross-validation, one can examine the discrepancies between $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$ for each. The model that attains the lowest total discrepancy is generally to be preferred for predicting future dependent values. Currently, there is no general theory that enables one to assess the statistical significance of the magnitude of the difference in total discrepancy between two models except in specialized situations. Although the cross-validation criterion is not equivalent to a goodness-of-fit test for assessing a model, it often happens that similar conclusions are drawn by each procedure.

H. Missing and Censored Data

A common practical problem arises when some of the data are unavailable for analysis. In general, there are three cases: (i) some components of the $\boldsymbol{X}$ vector are missing for some of the observations, (ii) some of the observations are missing, or (iii) the response variable Y is censored. Methodologies exist to address all three cases, but the solutions that they offer are not entirely satisfactory.

Missing components commonly occur in industrial data, when certain on-line measuring devices fail, and in census or survey data, when people provide incomplete information or responses. The basic strategy for analysis is to use the complete data records to estimate or "impute" values for the records with missing data. One approach is the hot-deck method, in which an incomplete record is randomly matched with a complete record or records that show close agreement in the observed components. Alternatively, one can use regression methodology to estimate the missing values, or the EM algorithm, described by Little and Rubin (1987) and McLachlan and Krishnan (1997), iteratively to estimate both missing data and model parameters simultaneously. Kalton and Kasprzyk (1986) survey the data imputation literature for the missing components problem.

If some of the observations are missing and their absence is unrelated to the outcome, then the impact of the missing data on regression analysis is minimal. In experimental design, however, one has typically prespecified the $\boldsymbol{X}$ values so as to ensure that certain planned comparisons can be made or to guarantee that the variance of the estimates has some desirable property. Losing even a single observation can affect such goals. The strategy for salvaging as much as possible of the analysis depends upon the assumptions one is willing to make, the structure of the initial design, and the location of the missing observations. Milliken and Johnson (1984) describe the various approaches.

Censored data arise when for some of the observations, one only knows that a dependent or independent variable exceeded a certain value. The appropriate analysis depends sensitively upon the type of censoring that occurs. The three primary types are as follows.

1. Type I censoring. In this application there is a fixed threshold, and one cannot observe the precise values of data exceeding that threshold. The circumstance might arise if one's measuring instruments had a ceiling; for example, a temperature gauge in a chemical process may not be able to report values greater than 500°C, even though the process may become much hotter.

2. Type II censoring. This arises in life testing for very reliable units; here it is often infeasible to wait until all units fail before estimating the average time to failure. To avoid undue delay, one terminates the experiment after observing the rth failure. For example, one might start 100 experimental engines and terminate the experiment when the third engine fails. Thus for 97 of the cases, one only knows that their true lifetimes would have been greater than the time to third failure.

3. Random censoring. This occurs when for each observation, one observes either the true value or a random censoring variable, depending upon which is smaller. For example, one may want to estimate the time to failure of an implanted pacemaker, but some patients die of other causes, move to a new city, or cease to report for other

reasons. In these cases one only knows that the pacemaker was still working as of the time that the patient's participation stopped.

There is a large literature on censored data problems. Miller (1981) offers an excellent introduction.

References

Arnold (1981), Christensen (1996), and Graybill (1976) are standard texts on univariate linear models. Arnold and Christensen take a coordinate-free approach, whereas Graybill presents a more classical treatment in terms of explicit coordinate systems and their matrix formulation. Christensen (1991) extends these ideas to the multivariate linear model described in this section. Lehmann (1986) discusses the likelihood ratio test and other hypothesis testing strategies; Lehmann and Casella (1998) detail issues in estimation. Berger (1985) is a moderately technical treatment of Bayesian inference; DeGroot (1970) is an accessible mathematical introduction to the field. Broemeling (1985) details Bayesian methodology in the context of the linear model. McCullagh and Nelder (1989) offer the most detailed treatment of GLM procedures. For bootstrap methods, the fundamental reference is Efron (1982); a readable introduction to more recent work is Efron and Tibshirani (1986). Stone (1977) surveys the main ideas in cross-validation and Mosteller and Tukey (1968) give a detailed illustration of the approach using a nested version which both produces a cross-validation estimate and then assesses its variability. Little and Rubin (1987) and McLachlan and Krishnan (1997) discuss a large class of missing data problems; Lawless (1982) presents methodology for the analysis of censored data in the context of life testing.

III. REGRESSION ANALYSIS

Regression problems first received mature attention when Francis Galton (1888) attempted to assess the relationship, or correlation, between the heights of fathers and the heights of sons. The area developed quickly, and is now the most widely used branch of applied statistics.

Multiple linear regression is a special case of the general linear model given in (1). When $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})$ is invertible, the estimate of the parameter vector $\boldsymbol{\theta}$ is unique, and inferential procedures based upon the likelihood ratio test are straightforward. Specifically, all standard hypotheses can be written in the form

$$\mathrm{H}_0: \quad \boldsymbol{A\theta} - \boldsymbol{b} = \boldsymbol{0} \qquad \text{versus} \qquad \mathrm{H}_{\mathrm{a}}: \quad \boldsymbol{A\theta} - \boldsymbol{b} \neq \boldsymbol{0} \tag{12}$$

for a square matrix $\boldsymbol{A}$ of rank q and vector $\boldsymbol{b}$ appropriate to the problem. Then, using (7), the likelihood ratio test statistic is

$$\mathrm{L}(\boldsymbol{X}) = \frac{(\boldsymbol{A\hat{\theta}} - \boldsymbol{b})^{\mathrm{T}}[\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}]^{-1}(\boldsymbol{A\hat{\theta}} - \boldsymbol{b})}{\boldsymbol{Y}^{\mathrm{T}}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}]\boldsymbol{Y}} \times \frac{n-p-1}{q}, \tag{13}$$

where p is the number of explanatory variables and q is the rank of the matrix $\boldsymbol{A}$. In most applications $\boldsymbol{\Sigma}$ is reasonably taken to be $\sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. Very often, $\boldsymbol{b}$ is $\boldsymbol{0}$.

Under the null hypothesis, the test statistic (13) follows an F-distribution with $\nu_1 = q$ numerator degrees of freedom and $\nu_2 = n - p - 1$ denominator degrees of freedom. The degrees of freedom indicate the amount of information that remains for testing the hypothesis after some of the information in the sample has been used to estimate the parameters. Each parameter estimate in linear modeling consumes a quantity of information equivalent to one observation; for nonlinear modeling, a parameter estimate uses more, often the equivalent of two to four observations.

Suppose we have fit the linear model

$$Y_i = \theta_0 + \theta_1 X_{i1} + \cdots + \theta_p X_{ip} + \epsilon_i$$

to a sample of an observation and are willing to assume independent, identically distributed normal error. Our intention is to make a level $\alpha = .01$ test of $\mathrm{H}_0: \theta_{k+1} = \cdots = \theta_p = 0$. If the null hypothesis is not rejected, this is evidence that data conform to a simpler model. To implement the test, set $\boldsymbol{b} = \boldsymbol{0}$ in (12) and take $\boldsymbol{A}$ to have all entries zero except for ones in the $(k+2)$th through $(p+1)$th positions along the diagonal (recall that our initial component of $\boldsymbol{\theta}$ is denoted θ_0, so that θ_{k+1} is the $(k+2)$th component of a vector in $\mathbb{R}^{p+1}$). Then evaluate (13) and compare its value to the $100(1-\alpha) = 99$th percentile point of an F-distribution with $p-k$ and $n-p+k$ degrees of freedom. If the value of the common variance σ^2 is unknown, then one replaces it with the estimate given by (5) and compares the test statistic to the percentile point of an F-distribution with $n-p-2$ degrees of freedom for the denominator.

As one moves away from the classical linear model, regression becomes more complicated. Analysis is slightly more difficult when $\boldsymbol{\Sigma}$ is unknown, and much more complicated when the variance of ϵ_i depends on $\boldsymbol{X}_i$. Nonconstant variance is called heteroscedasticity, and typically one must either transform the data to stabilize the variance or else use a weighted least squares algorithm. Fortunately, most inferences are insensitive to reasonably nonnormal error distributions, and regression procedures are surprisingly robust to violations of model assumptions.

Moreover, many special tricks enable one to overcome specific deficiencies in the data (Madansky, 1988).

A. Correlation Analysis

The correlation ρ between two random variables X and Y,

$$\rho = \frac{\text{cov}[X, Y]}{\sqrt{\text{var}[X]\text{var}[Y]}},$$

is a measure of the strength of the linear relationship between them. In multivariate applications, a random vector $\boldsymbol{X}$ has correlation matrix $\boldsymbol{P} = \{\rho_{ij}\}$, where ρ_{ij} denotes the correlation between the ith and jth components of $\boldsymbol{X}$. If one has a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from a multivariate normal distribution, then the maximum likelihood estimate $\boldsymbol{R} = \{r_{ij}\}$ for $\boldsymbol{P}$ has entries

$$r_{ij} = \frac{\sum_{k=1}^{n}(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)}{\sqrt{\left[\sum_{k=1}^{n}(X_{ki} - \bar{X}_i)^2\right]\left[\sum_{k=1}^{n}(X_{kj} - \bar{X}_j)^2\right]}},$$

where $\bar{X}_i$ is the average of the ith components. The matrix $\boldsymbol{R}$ is called the sample correlation matrix. In regression, it usually happens that associations involving the response variable Y are of main interest, but the correlations among the explanatory variables can profoundly affect one's conclusions.

For practical applications, one often wants to do the following:

1. Test a hypothesis about the correlation between two variables.
2. Assess the strength of the residual association between two random variables after accounting for their mutual association with a set of k other random variables.
3. Assess the strength of the association between a single random variable and the best linear combination of a set of k random variables.

The precise interpretation of the results of such analyses is sensitive to the assumption that the data are normally distributed; in particular, $\rho = 0$ implies that X and Y are completely mutually uninformative if and only if the joint distribution of X and Y is bivariate normal. Reasonable deviations from normality, however, usually have negligible effect in applications.

Two procedures are available to address the first question. For the null hypothesis that correlation ρ is equal to zero, Fisher (1915) showed that, when the data come from a multivariate normal distribution, the test statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

follows the t distribution with $n-2$ degrees of freedom, where r is the sample correlation. The percentile points of these distributions, as needed for a test with a given α value, are ubiquitously tabulated.

If the null hypothesis specifies a value for ρ other than zero, say ρ_0, then no exact procedure is easily available. Fisher (1921) showed, however, that $z = \tanh^{-1} r$ is asymptotically normal with mean $\tanh^{-1}\rho$ and variance $(n-3)^{-1}$. This enables one to test H_0: $\rho = \rho_0$ versus H_a: $\rho \neq \rho_0$, by referring

$$z = \sqrt{n-3}\left(\tanh^{-1} r - \tanh^{-1}\rho_0\right)$$

to a value from a standard normal table.

The second question leads to the definition of partial correlation. This measures the association between two random variables after controlling for the effects of one or more other random variables. Here, the term "controlling for" indicates that the measure is calculated as if the other variables were all held constant; more precisely, it is the correlation of the conditional bivariate distribution of the random variables, given the values of the other random variables. Mathematically,

$$\rho_{ij.k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{\left(1-\rho_{ik}^2\right)\left(1-\rho_{jk}^2\right)}} \tag{14}$$

is the partial correlation between the ith and jth components of $\boldsymbol{X}$, controlling for the effect of the kth. To obtain the partial correlation controlling for additional variables, one uses the recursion formula

$$\rho_{ij.k_1,\ldots,k_{s+1}} = \frac{\rho_{ij.k_1,\ldots,k_s} - \rho_{ik_{s+1}.k_1,\ldots,k_s}\rho_{jk_{s+1}.k_1,\ldots,k_s}}{\sqrt{\left(1-\rho_{ik_{s+1}.k_1,\ldots,k_s}^2\right)\left(1-\rho_{jk_{s+1}.k_1,\ldots,k_s}^2\right)}}. \tag{15}$$

To obtain the sample estimate of the partial correlation, one replaces the population values in (14) and (15) with the sample estimates obtained from applying recursion formulas to the entries of $\boldsymbol{R}$. [For more information, see Anderson (1984) and Tong (1990). These recursion formulas go back to the original work of Yule at the turn of the century.]

The third question gives rise to the multiple correlation coefficient. This is the maximum possible correlation between a given random variable and the best linear combination of a given set of random variables:

$$\rho_{Y.X} = \frac{1}{\sigma}\sqrt{\boldsymbol{\tau}^{\text{T}}\text{var}[\boldsymbol{P}]^{-1}\boldsymbol{\tau}}, \tag{16}$$

where $\boldsymbol{\tau}$ is the vector whose jth component is the covariance between Y and the jth component of $\boldsymbol{X}$. This is clearly important in linear regression since the multiple

correlation measures how well the dependent variable Y is explained by the best linear combination of p explanatory variables. In applications, one estimates $\rho_{Y.X}$ from sample data by replacing the values in (16) by the corresponding maximum likelihood estimates, yielding $R_{y.x}$. Distribution theory for the sample estimate of the multiple correlation coefficient was developed by Fisher (1928).

The multiple correlation coefficient can be generalized to what is known as the canonical correlation coefficient, a quantity which maximizes the correlation between linear combinations of two disjoint sets of random variables subject to constraints upon the linear combinations. This extension is important in understanding structures in the covariance matrix, and is related to topics discussed in Section V.

The square of the multiple correlation coefficient $R^2_{Y.X}$ is referred to as the coefficient of determination. This index is frequently used to assess the success of the analysis; it can be interpreted as the proportion of the variance in Y explained by the value of the best linear combination of the explanatory variables. Since $R^2_{Y.X}$ approaches 1 as p increases, one must adjust $R^2_{Y.X}$ to take account of the number of components in $\boldsymbol{X}$. One such correction,

$$R^2_a = \frac{(n-1)R^2_{Y.X} - p}{n - p - 1}$$

was given by Ezekiel (1930). This method is similar in spirit to Mallows' C_p statistic, which is used in the context of variable selection (Mallows, 1973). A comparative discussion of both methods appears in Rawlings *et al.* (1998).

In the context of the multivariate linear regression model with normal errors of this section, (i) $\rho_{ij} = 0$ is equivalent to the marginal independence of X_i and X_j, (ii) $\rho_{ij.k_1,\ldots,k_{s+1}} = 0$ is equivalent to the conditional independence of X_i and X_j given $X_{k_1}, \ldots, X_{k_{s+1}}$, and (iii) $\rho_{Y.X} = 0$ is equivalent to the independence of Y and $\boldsymbol{X}$. In Section V, these types of independence will be an integral part of the specification of models for the case where both Y and $\boldsymbol{X}$ are categorical variables instead of continuous variables.

B. Multicollinearity

A potential difficulty in linear regression is that the rows of the data matrix $\boldsymbol{X}$ are sometimes highly correlated. This is called multicollinearity; it occurs when the explanatory variables lie close to a line or plane or other proper subspace of $\mathbb{R}^p$. If carried to the extreme of perfect correlation, this implies that the covariance matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ is not invertible, and thus the parameter estimates in (3) cannot be obtained. Multicollinearity is a common problem in observational studies, but can usually be entirely avoided in carefully designed experiments.

As an example, suppose the dependent variable is the compressive strength of an aluminium can and the explanatory variables include indices of the bowing in the can walls, obtained from both a dial gauge and a coordinate measuring machine. These two values are so highly correlated as to be almost perfectly redundant, causing multicollinearity. Similarly, multicollinearity will occur when the explanatory variables are near perfect linear functions of other explanatory variables. In chemical engineering, this happens when several explanatory variables are the proportions of additives in a mixture. If the sum of all proportions must equal one, then there is perfect correlation and the covariance matrix is singular. If one avoids this by excluding one of the proportions, then it can still happen that the excluded variable shows very little variation, inducing multicollinearity.

Conceptually, multicollinearity implies that there is less information in the sample than one would expect, given the number of measurements taken. The inferential impact of this is that the estimate $\hat{\boldsymbol{\theta}}$ is unstable, so that small perturbations in the data cause large changes in the inference. Consequently, predictions based on the fitted model can be very bad.

C. Residuals

Residuals are the differences between the observed response values and those predicted after estimating the model parameters. The residual corresponding to the ith observation Y_i with explanatory variable values $X_{i1}, \ldots, X_{ip}$ is

$$\hat{\epsilon}_i = Y_i - (\hat{\theta}_0 + \hat{\theta}_1 X_{i1} + \hat{\theta}_p X_{ip}).$$

If many residuals have large relative magnitudes, this suggests the model is wrong and a more complicated model should be fit. A single residual with very large magnitude indicates an outlier; this data point should be checked to be sure the record is accurate, and even then it is often wise to set it aside during the analysis. (Surveyors distinguish outliers from blunders; if the cause of the outlier is discovered, it is called a blunder.)

Plots of the residuals against the explanatory variables are an important tool in regression analysis. Qualitatively, a good plot of $\hat{\epsilon}_i$ versus the jth component of $\boldsymbol{X}_i$ shows an oval-shaped point cloud centered on the line $\hat{\epsilon} = 0$. We note that although in the standard model the true errors ϵ_i are independent with unknown constant variance σ^2, the corresponding estimated residuals have mean 0 but

$$\mathrm{var}[\hat{\boldsymbol{\epsilon}}] = \sigma^2(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}). \qquad (17)$$

This dependence complicates the problem of identifying outliers and data values that have unusual influence on the overall inference.

One consequence of the dependence structure in (17) is that residuals near the average $\bar{\boldsymbol{X}}$ of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ tend to be large, whereas residuals far from $\bar{\boldsymbol{X}}$ tend to be small. This happens because observations far from $\bar{\boldsymbol{X}}$ have disproportionate influence, and these are called "high leverage" cases. High leverage is undesirable since inadequacies in modeling are expected to be greatest at the extremes of the data. Values with high leverage are masked by the residuals. A partial correction for this effect is to plot the studentized residuals

$$e_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - \boldsymbol{X}_i(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_i^{\mathrm{T}}}}$$

against the $\boldsymbol{X}_{ij}$ in place of $\hat{\epsilon}_i$. One can show that this reweights the observed residuals so that large discrepancies are more likely to occur at the extreme range of the $\boldsymbol{X}$ values.

There are many techniques for assessing a regression from functions of the fitted residuals; these include Cook's D and PRESS statistics. Belsley *et al.* (1980) survey regression diagnostics with special emphasis upon leverage, outliers, masking and influence functions. Cross-validation and the bootstrap, discussed in Section I, also enable inference about the adequacy of the model without requiring too many formal assumptions about the joint distribution of the error terms.

D. Transformations

The scope of application of the classical model can be extended through transformations of the data. Often the data collected fail to satisfy the classical assumptions; for example, the variance of the errors may not be constant over the space of the independent variables (heteroscedasticity), or the error distribution may not be normal, or the relationship of the parameters to the data may be nonlinear. In these circumstances, one can sometimes find a transformation of $(Y, \boldsymbol{X})$ that stabilizes the variance, normalizes the error, and/or linearizes the problem.

For example, in chemical kinetics, the rate R of a reaction between Z_1 moles of compound A and Z_2 moles of compound B might have the known (nonlinear) functional form

$$R_i = cZ_{i,1}^{a} Z_{i,2}^{b} e^{\epsilon_i}.$$

By taking logarithms of both sides of the equation, one gets the linear model

$$\ln R_i = \ln c + a \ln Z_{i,1} + b \ln Z_{i,2} + \epsilon_i.$$

Setting $Y_i = \ln R_i$ and $X_{i,j} = \ln Z_{i,j}$, one can attempt a standard analysis.

As another example, heteroscedasticity frequently occurs when Y_i is a proportion or percentage. Here the error variance is largest when the true value p_i is near .5 and decreases to 0 when p_i approaches 0 or 1. However, one can show that the transformed value $Z_i = \arcsin\sqrt{Y_i}$ asymptotically has constant variance, and so can be regressed against the independent variables.

A very popular set of data transformations are the Box–Cox family. These are variance-stabilizing transformations, pertinent for fitting the model

$$Y_i^{\lambda} = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta} + \epsilon_i,$$

where Y_i is constrained to be positive, $\mathrm{var}[\boldsymbol{\epsilon}] = \sigma^2\boldsymbol{I}$, and λ is an unknown exponent that is to be estimated from the data. If the data contain zeros, then it is customary to replace Y_i by $Y_i + 1$ throughout, before estimating λ. The Box–Cox family is popular because it contains many standard transformations; when $\lambda = .5$, one has a square-root transformation that is theoretically appropriate for data that arise from counts distributed as Poisson random variables. Similarly, $\lambda = -1$ is useful when observations concentrate near zero but long tails occur; one distribution-theoretic grounds, this transformation is probably appropriate when the scatterplot shows $\mathrm{var}[\epsilon_i] \propto \mathrm{E}[Y_i]^4$. Also, if $\lambda = 0$, one can show by a continuity argument that the appropriate transformation is $\log Y_i$, which is appropriate when the range of the response variable is very broad and $\mathrm{var}[\epsilon_i] \propto \mathrm{E}[Y_i]^2$.

Box and Cox (1964) offer guidelines for estimating λ, and Carroll and Rupert (1988) give more recent suggestions. In practice, many prefer to examine the data in an exploratory fashion to find a simple transformation within the Box–Cox family rather than estimate an arbitrary fraction; this honors the fact that simple transformations often correspond to scalings for variables and thus are natural models for the data. Inference is generally insensitive to small changes in the transformation.

E. Stepwise Selection

A common problem in multiple regression is an excess of explanatory variables. If one fits a model with too many independent variables (either those originally measured, or powers, products, and transformations of the data), then one can achieve spuriously good fit. The model shows very good agreement with the data, but is nearly useless for describing the true relationship or predicting future responses at particular values of the independent variables. To avoid this problem of overfit, one should include only those independent variables that explain a significant proportion of the variation in $\boldsymbol{Y}$ that is not accounted for by

variables already used in the model. This goal entails a process called variable selection.

To perform a stepwise variable selection for univariate dependent variables, one can use routines in any major software package (BMDP, SAS, SPSS, Stata, SYSTAT, S-PLUS) that implement an iterative algorithm for including variables in the model. This algorithm begins with no variables in the model. At the kth step, it adds the most valuable explanatory variable from the set of currently unincluded variables, provided that some minimal level of explanatory value is attained. Then it reviews all of the variables now in the model, removing any whose explanatory value, given the most recently added variable, has dropped below a certain threshold. Other selection procedures exist, but none stands upon a firm mathematical justification. The methods are all intrinsically nonlinear, which makes theoretical analysis very difficult. Nonetheless, empirical work indicates that variable selection will substantially improve the accuracy of most multiple regression analyses.

F. Some Special Cases

Polynomial regression illustrates a general strategy for extending linear regression so as to fit curved lines to response data. For example, one can fit a cubic equation to the data using the model

$$Y_i = \theta_0 + \theta_1 X_i + \theta_2 X_i^2 + \theta_3 X_i^3 + \epsilon_i. \qquad (18)$$

Here the powers of X are entries in different columns of the $\boldsymbol{X}$ matrix, and the linear structure of (18) ensures this is formally equivalent to the classical model. Similarly, one can handle interactions between two different kinds of explanatory variables, say X_{i1} and X_{i2}, by including terms of the form $X_{i1}^r X_{i2}^s$, where r and s are chosen by the investigator to best capture the type of interaction expected.

Logistic regression is used when the response variables are binary but the explanatory variables are not. This would be the case if one were measuring whether a steel wire breaks under a 50-kg load, where the explanatory variables might be the cross-sectional radius and the percentage of chromium. One may then use the model

$$P[Y_i = 1] = \frac{\exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta})}{1 + \exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta})}, \qquad (19)$$

where $Y_i = 1$ if the ith wire snaps, and $\boldsymbol{X}_i$ is the vector of explanatory variables for the ith wire. To justify linear regression methods, one uses the transformation

$$p_i = \ln \frac{P[Y_i = 1]}{1 - P[Y_i = 1]}$$

to obtain the linear model $p_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}$. When fitting this model, it is convenient if there are multiple measurements on each combination of values for the explanatory variables; then one can use the transformed observed proportions of ones as the observations in (1). Otherwise, one can use transformed local averages of the Y_i values. Either approach gives a linear model, where the error variance depends upon $\boldsymbol{X}_i$. One way to estimate $\boldsymbol{\theta}$ in the logistic regression case is via weighted least squares. When the components of $\boldsymbol{X}$ are categorical, an alternative is provided by the maximum likelihood methods described in Section V.

The use of dummy variables greatly extends the scope of linear regression. Here one can model the effect of categorical explanatory variables in combination with continuous explanatory variables. For example, suppose one measures the temperature of heating elements from two different suppliers as a function of the electrical power supplied. The continuous explanatory variable is the power and the categorical variable is the supplier. If one defines the variable X_{1i} to be 1 if the ith element is from the first supplier and otherwise set this to 0, then one can use the linear regression form of the model in (1) to automatically fit the supplier effect. Judgment is needed in deciding to fit interaction terms with dummy variables and in handling situations with more than two categories per variable.

G. Generalized Additive Models

Recent research in computer-intensive statistics has developed methods for additive modeling that extend the linear regression model. These generalized additive models enable the practitioner to fit a flexible family of nonlinear functions. The simplest generalized additive model has the form

$$Y_i = \theta_0 + \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i, \qquad (20)$$

where the ϵ_i are taken to be independent with constant variance and mean zero, and θ_0 and the functions f_j are unknown. Since the f_j are estimated from the data, one can avoid the traditional assumption of local linearity which is often needed to justify the model in (1) for practical applications.

If one wanted to fit arbitrary functions in p-dimensional spaces, one would require enormous numbers of observations for even very moderate values of p. This reflects the "curse of dimensionality," which implies that the local density, and hence local information, of a fixed number of uniformly spaced points in a δ-neighborhood in $\mathbb{R}^p$ goes to zero exponentially fast as p increases. Conventional linear model theory avoids this problem by considering only classes of multivariate functions on $\mathbb{R}^p$ that are linear in all variables; hence, one need estimate only

$p+1$ coefficients. Generalized additive models are more broadly applicable and contain linear models as a special subfamily. They require substantially more data than linear regression since they estimate p different functions rather than $p+1$ coefficients; nonetheless, the requisite data increase only linearly in p.

The backfitting algorithm is the key procedure used to fit generalized additive models; operationally, the algorithm proceeds as follows:

1. At the initialization step, define functions $f_j^{(0)} \equiv 1$; also, set $\theta_0 = \bar{Y}$.
2. At the ith iteration, estimate $f_j^{(i+1)}$ by
$$f_j^{(i+1)} = \mathrm{Sm}\left(\boldsymbol{Y} - \theta_0 - \sum_{k \neq j} f_k^i \mid X_{1j}, \ldots, X_{nj}\right)$$
for $j = 1, \ldots, p$.
3. Check whether $f_j^{(i+1)} = f_j^{(i)}$ for all $j = 1, \ldots, p$. If not, return to step 2; otherwise, use the $f_j^{(i)}$ as the additive functions f_j in the model.

This algorithm requires a smoothing operation, indicated by $\mathrm{Sm}(\cdot \mid \cdot)$. For large classes of sensible smoothing functions, one can prove that the backfitting algorithm converges to a unique solution.

Conceptually, the smoothing operation takes a scatterplot of the response variable Y against the jth component of the explanatory variables and returns a somewhat smoother continuous function that runs through the center of the data. As an example of a smoother, consider the local averaging function; this has the form
$$f_j(x) = \mathrm{Sm}(Y_i \mid X_{1j}, \ldots, X_{nj}) = \sum_{i=1}^{n} K(x, X_{ij}) Y_i,$$
where $K(x, x^*)$, the kernel function, is typically maximized at $x = x^*$, decreases monotonically to zero as $|x - x^*|$ increases, and satisfies
$$\int_{-\infty}^{\infty} K(x, x^*)\, dx = 1.$$
Many more-complex smoothers are available; some of the most popular are based upon cubic smoothing splines. The optimal smoother depends upon the problem in hand, but inference is usually insensitive to the particular smoother employed.

These simple generalized additive models capture nonlinearities in the directions of the coordinate axes, but cannot describe local behavior in other directions. At need, one can build in dummy variables that are products of some of the explanatory variables or even modify the backfitting algorithm to fit bivariate functions, but this rapidly becomes data-intensive and the curse of dimensionality inevitably arises. Hastie and Tibshirani (1990) survey recent work in this area.

H. Projection Pursuit Regression, MARS, and ACE

The simple generalized additive models consider sums of functions taking arguments in the natural coordinates of the space of explanatory variables. In some circumstances (such as extreme multicollinearity), it may be more sensible to use projection pursuit regression (Friedman and Stuetzle, 1981). This procedure employs the backfitting algorithm to fit a model of the form
$$Y_i = \sum_{k=1}^{r} f_k\left(\boldsymbol{\alpha}_k^{\mathrm{T}} \boldsymbol{X}_i\right) + \epsilon_i,$$
where the ϵ_i are independent with mean zero and variance σ^2, and the $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r$ determine a set of r linear combinations of the explanatory variables. These linear combinations are similar in spirit to those used in principal components analysis (see Section IV). A salient difference is that these vectors need not be orthogonal, but in principle these are chosen to maximize the predictive accuracy of the model as assessed through cross-validation (see Section I).

Projection pursuit regression is very difficult to implement, and hard to interpret when $r > 1$. Unlike the simple generalized additive models, it is invariant to affine transformations of the data; this is an appealing property for many practical situations in which the measurements have no natural coordinate basis. In some applications one may want to include a *very* small number number of interaction terms between the explanatory variables; this can be accommodated by defining dummy variables that are products of other explanatory variables.

Friedman (1991) describes an enormous extension of the generalized additive model that depends upon multivariate adaptive regression splines (MARS). This procedure fits a model that is the weighted sum of multivariate spline basis functions, also known as tensor-spline basis functions, and takes the form
$$Y_i = \sum_{k=0}^{q} a_k B_k(X_1, \ldots, X_n) + \epsilon_i,$$
where the ϵ_i terms are, as usual, independent with mean zero and variance σ^2, and the coefficients a_k are determined in the course of cross-validation fitting. The constant term follows by setting $B_0(X_1, \ldots, X_n) \equiv 1$, and the other multivariate splines are products of univariate spline basis functions:
$$B_k(x_1, \ldots, x_n) = \prod_{s=1}^{r_k} b\left(x_{i(s,k)} \mid t_{s,k}\right), \qquad 1 \leq k \leq r.$$

Here the subscript $1 \le i(s,k) \le p$ indicates a particular explanatory variable, and the basis spline in that variable has a knot at $t_{s,k}$. The values of q, the $r_1, \ldots, r_q$, the knot sets, and the appropriate explanatory variables for inclusion are all determined adaptively from the data.

This procedure is closely related to the regression trees described by Breiman *et al.* (1984), and is a natural generalization of the recursive partitioning strategy described in Section VI. MARS admits an ANOVA-like decomposition that can be represented in a table and similarly interpreted (see Section III). Although no procedure can overcome the curse of dimensionality, MARS is designed to perform well whenever the true function has low local dimensionality. The procedure automatically accommodates interactions between variables and variable selection. Although it is very new and very computer-intensive, the methodology offers the only current approach for nonlinear multivariate modeling that begins to respond to the generality of problems encountered in practice.

A third extension of the generalized additive model permits functional transformation of the response variable as well as the explanatory variables. This depends on the ACE algorithm developed by Breiman and Friedman (1985), and it fits the model

$$g(Y_i) = \theta_0 + \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i,$$

where all conditions are as given for (20), except now the function g is also unspecified, satisfying only the technically necessary condition that $\mathrm{var}[g(Y)] = 1$.

ACE is an acronym for alternating conditional expectations, which is the mechanism for determining the fitted functions. Given variables Y_i and $\boldsymbol{X}_i$, one wants g and $f_1, \ldots, f_p$ such that $\mathrm{E}[g(Y_i) \mid \boldsymbol{X}_i] - \sum_{j=1}^{p} f_j(X_{ij})$ resembles independent error (without loss of generality, we can ignore the constant term θ_0). Heuristically, this is accomplished by finding

$$(g, f_1, \ldots, f_p) = \operatorname{argmin} \left\{ \sum_{i=1}^{n} \left[g(Y_i) - \sum_{j=1}^{p} f_j(X_{ij}) \right]^2 \right\},$$

where g satisfies the unit variance constraint. Operationally, one solves as this minimization problem as follows:

1. Estimate g by $g^{(0)}$, obtained by "smoothing" the Y_i values and standardizing the variance. Set $f_j^{(0)} \equiv 1$ for all $j = 1, \ldots, p$.
2. Conditional on $g^{(k-1)}(Y_i)$, apply the backfitting algorithm until convergence to find estimates $f_1^{(k)}, \ldots, f_p^{(k)}$.
3. Conditional on the sum of $f_1^{(k)}, \ldots, f_p^{(k)}$, obtain $g^{(k)}$ by applying the backfitting algorithm until convergence (this interchanges the role of the explanatory and response variables). Standardize the new function to have unit variance.
4. Test whether $e^{(k)} - e^{(k-1)} = 0$, where

$$e^{(k)} = n^{-1} \sum_{i=1}^{n} \left[g^{(k)}(Y_i) - \sum_{j=1}^{p} f_j^{(k)}(X_{ij}) \right]^2.$$

If it is zero, set $g = g^{(k)}$, $f_j = f_j^{(k)}$; otherwise go to step 2.

Although it may not obvious from this representation, steps 2 and 3 calculate smoothed expectations, each conditional upon functions of either the response or the explanatory variables; this alternation gives the method its name.

The ACE analysis finds sets of functions for which the linear correlation of the transformed response variable and the sum of the transformed explanatory variables is maximized. From this perspective, ACE methodology is close kin to correlation analysis, especially the calculation of the multiple correlation coefficient. ACE does not aim directly at regression applications and has some undesirable features; for example, it treats the response and explanatory variables entirely symmetrically and it does not reproduce model transformations. Tibshirani (1988) invented a variation of ACE called AVAS, which takes automatic account of variance stabilization issues. AVAS avoids many of the deficiencies of ACE in pure regression applications, but its theoretical properties have not yet been established.

References

Multiple regression has an enormous literature. Weisberg (1980) and Rawlings *et al.* (1998) cover basic applications of univariate linear regression, and Bates and Watts (1988) treat nonlinear regression. Hawkins (1980) examines strategies for outlier detection, and Madansky (1988) gives recipes for various departures from classical univariate linear regression. Tong (1990) summarizes results that enable the calculation of various distributions for the test statistics for correlational analyses.

IV. EXPERIMENTAL DESIGN

Experimental design is a methodology for assessing the effects of different treatments or manipulations when applied in combination. As an example, suppose one wants to compare the average effects of a different filaments and b different filling gases on the lifetime of incandescent light bulbs. Both gas and filament occur together within a light bulb; besides the direct effect of each on longevity, it may be that there are interactions between particular gases

and filaments that are of research interest. Often the primary concerns are to (i) determine whether there are differences in the means corresponding to different treatment combinations and (ii) distinguish treatments whose effects upon the mean are purely additive from those whose effects depend upon interactions with other treatments applied in combination. Additional analyses may be used to determine which treatment combination yields the largest average response, or to make estimates about the relative magnitudes of various effects.

There is no unique specification of the linear statistical model that goes with a designed experiment such as the one just described. For this example, a customary parametrization is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

where $i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; and $k = 1, 2, \ldots, n_{ij}$. Here Y_{ijk} is the observed lifetime of the kth light bulb receiving filament i and filling gas j, μ is the average lifespan of the population of all light bulbs, α_i is the change in lifespan caused by using filament i, β_j is the change in lifespan caused by using gas j, $(\alpha\beta)_{ij}$ is the change caused by the interaction between the ith filament and the jth gas, and the ϵ_{ijk} are independent normal errors with mean zero and variance σ^2. The α_i and β_j are called main effects due to filament and filling gas, respectively; the $(\alpha\beta)_{ij}$ terms are the interactions between level i of the filament factor and level j of the filling gas factor. In terms of the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, one has

$$\boldsymbol{\theta}^{\mathrm{T}} = (\mu, \alpha_1, \ldots, \alpha_a, \beta_1, \ldots, \beta_b, (\alpha\beta)_{11}, \ldots, (\alpha\beta)_{ab})$$

and the design matrix $\boldsymbol{X}$ consists of ones and zeros, determined by the particular levels (filament type and gas type) of the treatments applied to each of the light bulbs.

Notice that, unlike in the case of regression, the $\boldsymbol{X}$ matrix in the linear model for experimental design is not full rank. For example, the first column of $\boldsymbol{X}$ consists entirely of ones because all observations are modeled with the overall mean μ; similarly, the vector sum of the second through $(a+1)$th columns equals the first column since exactly one of $\alpha_1, \ldots, \alpha_a$ is applied to each observation. Since a $p \times p$ matrix has rank $p - k$, where k is the number of linear dependences among the columns, one can verify that this design matrix has rank $(1 + a + b + ab) - 3$. There is one linear dependence for each of the three sets of treatment parameters: the set of additive filament-type effects, indicated by α, the set of additive gas-type effects, indicated by β, and the interactions between the two sets, indicated by $(\alpha\beta)$.

This feature of experimental design models is called overparametrization, and is an inescapable property of the application. It implies that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ in (5) has no unique inverse, and thus certain functions of $\boldsymbol{\theta}$, such as α_1, cannot be estimated. However, in good designs, the $\boldsymbol{X}$ matrix possesses planned structure that enables the investigator to estimate the specific functions of major research interest; these functions are linear contrasts in the treatment effects, such as $\alpha_1 - \alpha_2$. (A linear contrast in $\boldsymbol{\theta}$ is any function of the form $\sum c_i\theta_i$ for which $\sum c_i = 0$.) Additionally, observations in a good design are allocated among the different treatment combinations so as to maximize the precision in estimating contrasts of research interest.

In classical experimental design, there is often a tight constraint on the number of observations that can be taken. This motivates the development of carefully planned experiments that maximize the information gleaned from each observation. Often the aim of a design is simply to determine which factors and interactions affect the responses, so that a second experiment can be focused more efficiently upon model building.

A. Univariate Response

The analysis of variance (ANOVA) table partitions the total sample variability in an experimental design into components attributable to specific independent variables and interactions between independent variables. Under the null hypothesis that a given variable (or interaction) does not affect values of the dependent variable, one can show that the ratios of certain quantities in the ANOVA table must follow a particular F-distribution. Also, under the alternative hypothesis that there is an effect on the response variable, then the ratio follows a distribution that tends to be larger than the F-distribution. This enables the use of standard hypothesis testing and interval estimation methodology.

The standard sequence of hypothesis tests in the ANOVA model is hierarchical; one tests first for interaction effects and then for main effects. This is done because sometimes an interaction effect can "mask" a main effect, making the test for main effects spuriously insignificant. By testing the most complex interactions first, one protects the inference from errors caused by masking. Also, the interpretation of interaction effects in the absence of main effects is problematic.

In the context of the example, one first looks for the presence of nonzero interactions $\{(\alpha\beta)_{ij}\}$. If the F-test for H_0: $(\alpha\beta)_{\mathrm{I}j} = 0$ for $i = 1, \ldots, a$, $j = 1, \ldots, b$, detects nonzero interaction effects, then both factors play a role and do so synergistically. This interaction effect means that some filling gases work better in combination with certain filaments than one would expect from their separate overall performances. If one concludes that interactions are not present, then one proceeds to test for the presence of main effects. For example, if all the α_i are zero, then the mean bulb lifespan is not affected by the

type of filament used. If some α_i are nonzero, then some filaments types are better than others. One can go beyond hypothesis testing to point and interval estimation; the former enables estimation of a specific contrast, such as the difference in average longevity between two types of filament, and the latter enables one to set a confidence interval for a contrast.

Statisticians have developed a large number of designs, each tailored to very specific applications. The following partial survey lists the major types of designs and describes their special properties.

Factorial design. There are p different factors; the kth factor has d_k levels. One takes n observations at each possible combination of factor levels, for a total of $n\Pi_{k=1}^{p} d_k$ measurements. Provided that $n > 1$, this design enables the researcher to examine all main effects, all two-way interactions between each pair of factors, all three-way interactions between each triplet of factors, and so forth down to the p-way interaction.

Fractional factorials. Factorial designs can require very large sample sizes, especially as the number of factors grows. To reduce required sample sizes, fractional factorial designs use a technique referred to as confounding. Confounding occurs when two main effects or interactions cannot be separated; if either is significant, then both will appear to be significant. Usually, a fractional factorial design attempts to confound high-order interactions with main effects; this conforms to the empirical wisdom that complex interactions are much less likely to occur than main effects. Thus the fractional designs give up the capability of estimating all possible interactions in order to gain a savings in experimental effort. A class of highly fractionated designs based upon a technique known as orthogonal arrays has become known as Taguchi designs, and these are useful when one is assessing a very large number of factors.

Block designs. Sometimes much of the observed variation is attributable to factors that are not of primary interest and are unlikely to interact with other independent variables. For example, some of the variation in time to first repair of an automobile may be explained by the day of the week on which it was manufactured. In such circumstances one can control for the day-to-day variation (referred to as the block effects) by ensuring that each of the levels of the other variables of interest occurs equally often on each manufacturing day (i.e., equally often within each block); this gives a randomized complete block design (RCBD). Blocking increases the ability of the ANOVA tests to detect and estimate with precision the effects of interest. Various refinements of this strategy are needed when one cannot expect to observe all combinations of all factor levels in a single block; these are referred to as balanced incomplete block designs (BIBs) and partially balanced incomplete block designs (PBIBs).

Analysis of covariance. This is a hybrid between experimental design and regression analysis. For example, suppose one compares the tensile strength of a steel alloy wire under two different pickling methods and two different die types, and there is covariate information on the percentage of chromium in each wire sample. In this case the analysis of covariance (ANACOVA) model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \theta Z_{ijk} + \epsilon_{ijk},$$

where Y_{ijk} is the observed tensile strength, α_i, β_j, and $(\alpha\beta)_{ij}$ represent the main effects and interaction of the pickling and die factors, and θ is a regression coefficient that measures the influence of chromium percentage Z_{ijk}. This enables one to assess the factor effects as well as the regression between the percentage of chromium and wire strength.

Nested designs. In factorial experiments, we say that the factors are crossed, meaning that each level of each factor occurs with each level of each of the other factors. This structure is not always possible. For example, one may want to compare color fastness in cloth swatches at two different plants, across the three different workshifts, using two measurements per bolt. In this case the factors are said to be nested, i.e., the observations are nested within level of bolt which is nested within level of shift which is nested within plant. It is impractical to move the workers on each shift to both locations and it is impossible to have the same bolt of cloth manufactured on two different shifts. Similarly, taking multiple measurements per bolt (referred to as subsampling) nests the observations within the bolt.

Split-plot designs. These designs are closely related to nested designs, in that certain factors of experimental interest cannot be crossed. Under traditional assumptions, this leads to two error terms, used for testing two different sets of effects. An example of a split-plot design is given in the following discussion of the analysis of variance table in the following subsection.

All of the designs listed above admit fixed, random, and mixtures of fixed and random effects models. Fixed-effects models correspond to the situation in which the levels of all factors are prespecified as part of the design. Random-effect models occur when the levels of all factors are random outcomes, and mixed models occur when some factors have fixed levels and other factors have random

levels. As an example of this distinction, suppose one has a factorial experiment that examines the nonconformance of manufactured yardsticks to target values as a function of the production machine and the operator. If the experimenter uses all production machines in the plant and chooses a set of operators at random to produce sample yardsticks on each of the machines, then the machine effect is fixed and the operator effect is random, so that one has a mixed factorial model. These different models affect the definition of θ in the linear model and the expected values of the terms in the ANOVA table.

B. Analysis of Variance Tables

The following example illustrates several of the key ideas of experimental design, including the generation and interpretation of an analysis of variance table corresponding to a reasonably representative model. It involves the analysis of data collected to assess radon concentrations in the basement and ground floors of 12 (nearly) randomly chosen Pittsburgh homes. The investigator was interested in determining the relationship between radon level of the basement and the ground floor level; the investigator also wanted to assess the effect of finished versus unfinished basements.

There were 12 homes, 6 with unfinished basements and 6 with finished basements. Levels of radon concentration were assessed for each floor of each house in three different locations, which may be viewed as subsampling. Since the finished versus unfinished effect can only be applied to an entire house, it is physically impossible to cross this with the house effect, and thus we have a split-plot design. The whole-plot treatment, which is applied to an entire block, is the finished/unfinished basement factor. The split-plot treatment, which can be crossed with the block effect, is the first floor versus basement factor. Note that the house effect is a random effect since the houses represent a random selection from a larger population; in contrast, the floor and finish effects are fixed effects.

The data are as described in Table I. A finished basement is recorded as a 1 and an unfinished basement is a 0; similarly, the basement readings are coded as 0 and the first floor readings are coded with a 1. Houses are numbered from 1 to 12. All readings are in picocuries/liter (pCi/liter).

A plot of the data within each home/floor level showed that the distribution of the readings is skewed to the right. This result, coupled with the fact that the readings depend upon counts of atomic events, suggested that a logarithmic transformation might make the data more normally distributed and within-group variances more nearly equal. This turned out to be the case, and so the following analysis uses the natural logarithms of the data in Table I, rather than the actual measurements.

TABLE I Radon Levels in 12 Pittsburgh Houses

House	Floor	Finish	Responses (pCi/liter)		
1	0	0	1.78	1.66	2.00
	1	0	0.65	0.57	0.66
2	0	1	10.91	11.46	10.49
	1	1	4.30	4.30	4.42
3	0	1	2.56	2.34	2.43
	1	1	1.62	1.72	1.49
4	0	1	5.33	5.31	5.00
	1	1	1.97	2.07	2.07
5	0	0	0.98	1.25	0.96
	1	0	.082	0.71	0.67
6	0	0	1.06	1.06	0.99
	1	0	0.58	0.83	0.87
7	0	1	1.70	1.53	1.77
	1	1	1.36	1.24	2.07
8	0	1	2.07	1.81	2.32
	1	1	0.65	0.78	0.89
9	0	0	2.03	1.86	2.35
	1	0	0.34	0.31	0.66
10	0	0	2.23	3.16	2.79
	1	0	0.29	0.48	0.57
11	0	0	1.75	1.29	1.44
	1	0	1.07	0.93	0.98
12	0	1	2.62	1.91	2.01
	1	1	1.90	1.33	2.26

Using standard statistical software in SAS, one obtains the ANOVA table shown in Table II. The first column indicates the components into which the variation in the sample is partitioned. The second column shows the degrees of freedom for each component, or, heuristically, the amount of sample information used in fitting each of the experimental factor effects of interest. The third column shows the sum of squares corresponding to each component of the partition; these are obtained as the solution of matrix expressions involving both the data and the design

TABLE II ANOVA Table for Radon Data

	Degrees of freedom	Sum of squares	Mean square	*F*-ratio
Finish	1	13.37	13.37	11.38
House(finish)	10	11.74	1.17	11.25
Floor	1	10.49	10.49	100.50
Floor × finish	1	0.53	0.53	5.08
Error	58	6.05	0.10	
Total	71	42.18		

matrix $\boldsymbol{X}$. The fourth column shows the mean squares, which are the sums of squares terms divided by the degrees of freedom within the same row. The fifth column shows the F-statistic for evaluating the significance of the effect of a particular component of the partition; this is the ratio of two terms in the mean square column. In order to determine which two mean square terms one should use in forming the F-statistic, one finds the two which have the same expectation under the null hypothesis of no component effect; the term that goes in the numerator is the term whose expectation increases under the alternative hypothesis. Much of the theory in experimental design is aimed at ensuring the existence of mean-squared terms with expectations appropriate for the formation of such test statistics.

Before reading this ANOVA table, notice that the source column identifies five components of the variation in the sample. The first four rows [finish, house(finish), floor, and floor × finish] represent effects that are of experimental interest. The error row enables an estimate of the intrinsic variance σ^2 in the observations that remains after fitting the best possible split-plot model to the data. If the experimental effects are large in comparison with this error variance, then the investigator has evidence that the effects are not simply due to random chance, but represent important features of the data. The sixth row (total) has no interpretative importance, but was originally used as an arithmetical check; modern software ensures accuracy, but it is customary to retain this row. Finally, there is an implicit sixth component of variation in addition to those listed in the first five rows, corresponding to the mean of all observations over all groups; it accounts for one degree of freedom, which is why the degrees of freedom in the last row is one less than the number of observations. Since researchers are rarely interested in hypotheses about the overall mean, it is usual to omit this component in developing the table.

When interpreting an ANOVA table, one first examines the highest order interaction. In this case, the highest order interaction is floor × finish; it corresponds to the research hypothesis that even after additive effects for floor, house, finish, and house(finish) have been accounted for, there are still significant effects that must be attributed to a synergy between the type of finish within a house's basement and the floor on which readings are taken. The reason for testing high-order interactions first is that a phenomenon called masking can occur. This is a moderately rare circumstance, in which the synergy is such that low-order effects appear insignificant, although high-order terms do not. In such cases one declares that the existence of an interaction between floor and finish effects ensures that floor and finish effects must be present even if the significance probabilities corresponding to the test statistics in the first, third, or both first and third rows are unpersuasive. For our example, the F-statistic for testing the presence of a floor × finish interaction is the ratio of the mean squared error term for the floor × finish row to the error row; this value is 5.08, and it is referred to an F-distribution with 1 and 58 degrees of freedom, using the degrees of freedom column for the corresponding rows. One finds that the significance probability is .028, which corresponds to rejecting the hypothesis of no interaction at the $\alpha = .05$ level, but not at the $\alpha = .01$ level.

Testing the floor effect is straightforward, but probably unnecessary since the interaction effect is so strong that one would attribute an insignificant result to masking and declare the floor effect to be automatically significant. Nonetheless, to illustrate the method, assume that the previous test had not been highly significant. Then one would form the F-statistic as the ratio of the mean squared term in the floor column to the mean-squared term in the error column and compare this to values from an F-distribution with 1 and 58 degrees of freedom. The test statistic is 100.5, which is highly significant; one rejects the null hypothesis that there is no floor effect at α levels less than .0001.

The house(finish) effect tests whether there are significant differences between the houses chosen in the sample. The parenthetical inclusion of finish indicates that the house effect is nested within finish; i.e., a single house can only have one kind of basement finish. Since the house effect is random, it might happen that one must take the denominator in the test statistic from some other row than the error row; however, this model is just sufficiently simple that such additional complication does not arise. So, proceeding as before, one finds the F-statistic to be 11.74, which is compared to values from an F-distribution with 10 and 58 degrees of freedom. The test rejects the hypothesis that all houses are equivalent at a level less than .0001.

An unusual feature of split-plot designs is that there are two error terms. Besides the term already discussed, one uses the house(finish) mean squared term as the denominator of the test statistic for evaluating the finish effect; thus it functions as an error term in assessing the finish effect. The justification for this depends upon the calculation of the expected mean squared terms, which shows that under the null hypothesis of no finish effect, the expected mean squared term for finish equals the expected term for house(finish). In this example, the early rejection of the hypothesis of no floor × finish effect precludes the need for further examination. However, for completeness, we note that the value of the test statistic is 11.38, which is compared to values from an F-distribution with 1 and 10 degrees of freedom. The hypothesis of no finish effect is rejected; the significance probability is .0071.

C. Simultaneous Inference

Simultaneous inference occurs when the practitioner must make judgments about more than one parameter value. A common situation is the derivation of a multivariate confidence region, as described in Section I. Using the duality between confidence regions and hypothesis tests, the confidence region is equivalent to a set of hypothesis tests about all possible linear combinations in the components of the parameter vector $\boldsymbol{\theta}$. These applications are intrinsic to multivariate analysis, and some of the methods described here apply to regression and other areas. The problem is particularly important in experimental design, where it has provoked a large literature on multiple comparisons. The driving motivation in experimental design is that one is not concerned with all possible linear combinations, but wants to maximize the power of tests that focus on a small set of linear combinations having primary research interest.

To illustrate the issue, suppose one makes k hypothesis tests about k different parameters and that each test has Type I error probability α. Then, if the null hypothesis is true for each, the probability of one or more Type I errors among the entire set of comparisons (the experimentwise Type I error rate) is $1-(1-\alpha)^k$; this goes to 1 exponentially in k. From the simultaneous inference standpoint, one is not making k decisions, but rather 2^k decisions, corresponding to each set of possible inferences about all k parameters.

As the simplest example in experimental design, suppose one has performed a one-way ANOVA comparing t different groups. If one rejects the null hypothesis that all groups have equal means, one must usually proceed to order the group means, or at least determine which group has the largest mean. In the former case, one must make $t(t-1)/2$ hypothesis tests, each having the form H_0: $\mu_i \le \mu_j$, and in the latter case, one must make $t-1$ tests, pairing the groups as if in a knockout tournament. This discussion focuses on the first case, but the ideas generalize to other applications.

The most direct attack on the multiple comparison problem depends upon the Bonferroni inequality. It asserts that

$$P\left[\bigcap_{i=1}^{k} E_i\right] \ge 1 - \sum_{i=1}^{k} P\left[E_i^c\right],$$

where E_i denotes any event and E_i^c is the complement of that event. Setting E_i to be the event that no Type I error occurs on the ith comparison, then this suggests taking the comparisonwise error rate to be $\alpha^* = \alpha/k$, where α is the desired experimentwise error rate. Unfortunately, this method becomes impractically conservative as k increases.

A second method is Fisher's Least Significant Difference (LSD) test. One begins by performing a standard ANOVA test; if this fails to reject the null hypothesis of no differences among the means, one stops. Otherwise, one does simple two-sample t-tests on all possible pairs of group means, declaring significant differences whenever these reject the null. The initial ANOVA test ensures that the overall experimentwise Type I error rate is α.

The third main approach is Tukey's Honestly Significant Difference (HSD) test, and this also applies specifically to experimental designs. Using the standard assumptions of independent sampling, normal distribution with unknown common variance σ^2, and equal sample sizes, Tukey worked out the distribution of the maximum studentized difference in sample means when, in fact, all population means are equal. Thus one performs $t(t-1)/2$ tests, declaring two groups to be significantly different when $\bar{Y}_{(j)} - \bar{Y}_{(i)} > q_{t,t(n-1),\alpha}$, where $q_{t,t(n-1),\alpha}$ is the upper $1-\alpha$ percentile point of the studentized range distribution with parameters t and $t(n-1)$. Tukey's HSD tends to be slightly more powerful than Fisher's LSD, although both maintain the same experimentwise error rate. Tukey's original work was unpublished; Miller (1981) gives a detailed account of the technique.

The fourth method is called F projection, and was developed by Scheffé (1953). It exploits the duality between confidence regions and hypothesis tests, as discussed in Section I, and its generality enables applications in many statistical contexts. From (8), we know that a simultaneous confidence region on the parameter θ is an ellipsoid in $\mathbb{R}^p$, where θ pertains to any linear model. Let $\mathcal{C}$ be a fixed q-dimensional subspace of $\mathbb{R}^p$, and let $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\theta}$ be a linear combination or contrast in the components of $\boldsymbol{\theta}$ obtained from $\boldsymbol{c} \in \mathcal{C}$. Then it can be shown that

$$P\left[\ \forall \boldsymbol{c} \in \mathcal{C},\ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{\theta} \in \boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\theta}} \pm s\sqrt{q F_{q,n-p,\alpha}}\,\boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{c}\right] = 1-\alpha, \tag{21}$$

where $F_{q,n-p,\alpha}$ is the usual critical point from an F-distribution with q and $n-p$ degrees of freedom. Thus, (21) defines a confidence region; by the duality between confidence regions and hypothesis tests, points within the region having components equal to zero imply that certain linear combinations correspond to null hypotheses which cannot be rejected. Since the probability content of the entire region is $1-\alpha$, then the experimentwise error rate for all tests derived from vectors in $\mathcal{C}$ is α. Appropriate choice of $\mathcal{C}$ enables one to make simultaneous test of unequal group means, nonzero regression coefficients, or many other standard applications. (The method is called F projection since the simultaneous confidence region for a particular linear combination $\boldsymbol{c}$ is the projection of the joint confidence region determined by the

F value onto the one-dimensional subspace determined by $\boldsymbol{c}$.)

Many other methods of simultaneous inference have been proposed, but these are the ones that remain most popular. No one method dominates the others in all applications. Miller (1981) gives the authoritative review of this area and provides methods for many nonstandard situations.

D. Response Surfaces and Evolutionary Operation

Response surface methodology is closely related to univariate experimental design and has broad application to industrial statistics. In this framework one has the model $Y_i = f(\boldsymbol{x}_i) + \epsilon_i$, where $f(\cdot)$ is an unknown function and the ϵ_i are independent errors having common normal distribution with 0 mean, and one wants to find the vector value $\boldsymbol{x}$ that maximizes (or minimizes) the average value of the response variable Y. The goal is to discover the maximizing vector as efficiently as possible, where efficiency implies minimizing the number of observations taken. There are many strategies for accomplishing this search; response surface designs usually use Taylor's theorem to justify a locally linear approximation to the unknown function $f(\cdot)$ and then draw observations and fit linear regression models to enable inference on the direction of steepest ascent. When the vicinity of the maximum has been reached, the linearized model is enlarged to include quadratic and cross-product terms, enabling estimation of the optimal location. Standard response surface designs include the central composite, Box–Behnken, and Plackett–Burman designs. Khuri and Cornell (1987) survey the variety of techniques used to tailor the principles of response surface analysis to particular applications.

Response surface ideas are linked to evolutionary operation (EVOP), but in the latter the dimension of the $\boldsymbol{x}$ space is usually larger and the goal is continually to learn incrementally more about optimal management of the process. The location of the maximum is allowed to change over time, so there is a sense in which these methods are structured to chase a moving target. The most common EVOP protocol is based upon a "flipped simplex" design; essentially, one always guides the process in the opposite direction (in $\boldsymbol{x}$ space) from the worst of the recent process response measurements. Box and Draper (1969) provide a pioneering discussion of EVOP strategies, based upon the principles of experimental design.

E. Multivariate Response

Experimental design methods generalize to applications in which p response variables are measured on each of n experimental objects. These are referred to as multivariate analysis of variance (MANOVA) designs, and build directly on ANOVA procedures. The extended multivariate linear response model is

$$\boldsymbol{Y} = \boldsymbol{\theta X} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}$ is a $n \times p$ matrix whose ith row represents a vector of observations on the ith experimental unit, $\boldsymbol{\theta}$ is a $q \times p$ matrix of model parameters, and $\boldsymbol{\epsilon}$ is an $n \times p$ error matrix, usually with mean $\boldsymbol{0}$ and correlated rows. When the values of $\boldsymbol{X}$ are 0's and 1's arranged so that the matrix is not of full rank but has planned structure that enables estimation of pertinent linear contrasts, then the model is usually referred to as a MANOVA model.

Parameters in the MANOVA model can pertain to between-unit effects, within-unit effects, or both simultaneously. Between-unit effects correspond to the univariate models described before; inferences are based upon comparisons among the sets of units that receive the same treatment levels. Within-unit effects refer to the interplay between the different responses of each unit; for example, if one measures the change in paint reflectivity, hardness, and color on panels exposed for 6 months to two different temperature levels and two different salinity levels, then one might test whether the factorial treatment structure affects the three dependent variables. The triplets of measurements upon each panel are not independent, however, and MANOVA is designed to account for this feature. In our example, a between-unit test might examine whether the temperature effect is significant, a within-unit test might examine whether color change is greater than reflectivity change, and a between- and within-unit test might examine the effect of temperature upon the difference between changes in color and reflectivity.

The general MANOVA hypothesis test takes the form

$$\mathrm{H}_0: \quad \boldsymbol{A\theta B} = \boldsymbol{0} \qquad \text{versus} \qquad \mathrm{H}_a: \quad \boldsymbol{A\theta B} \neq \boldsymbol{0}, \quad (22)$$

where the matrix $\boldsymbol{A}$ determines the contrasts used to assess between-unit effects and the matrix $\boldsymbol{B}$ defines the contrasts for the within-unit effects. Many test statistics are available; for most applications, the choice among them is somewhat arbitrary. All customary test statistics are some function of the eigenvalues $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ of $\boldsymbol{HE}^{-1}$, where

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{A}^{\mathrm{T}}[\boldsymbol{A}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{A}^{\mathrm{T}}]^{-1} \\ &\quad \times \boldsymbol{A}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{YB} \qquad (23) \\ \boldsymbol{E} &= \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}]\boldsymbol{YB}. \end{aligned}$$

Since the $\boldsymbol{X}$ matrix is not full rank, the Moore–Penrose generalized matrix inverse is used in (23).

A standard procedure for developing an appropriate test statistic is based upon Roy's union–intersection

principle, which relates the multivariate test statistic to a supremum over a set of univariate test statistics (Roy, 1953). Other testing criteria use test statistics based upon (i) $\Pi_{i=1}^{p}(1+\lambda_j)^{-1}$ (the likelihood ratio approach), (ii) $\sum_{i=1}^{p}\lambda_j$, or the trace of $\boldsymbol{HE}^{-1}$, and (iii) the largest eigenvalue of $\boldsymbol{HE}^{-1}$. For further details and references, see Press (1982).

Repeated-measures analysis is one of the most common applications of MANOVA. Here one has a regular univariate design, except that each experimental object is measured at several time intervals (the time steps are equal for all objects). Typically, one wants to test main effects and interactions among the factors in the univariate design and also examine time effects. One strategy for this is to treat the time variable as a separate factor in the experiment and proceed as if it were a univariate analysis. This ignores the covariance structure implicit in the correlated measurements upon the same object, but, surprisingly, the method is justified in some applications. Huynh and Feldt (1970) describe conditions on the covariance matrix that allow such tests. More generally, the MANOVA formulation of repeated measures honors the multivariate structure, and the implementation directly follows from (22) and (23) after specifying the appropriate matrices.

Profile analysis is a further refinement of repeated measures that focuses on specific questions regarding changes in measurement values over time. Assume that the between-group structure partitions the experimental objects into k groups, each corresponding to a different combination of factor levels. Plot the means of each group at each time against time. Then the three questions that profile analysis considers are:

1. Do all k plots have piecewise parallel line segments?
2. If they are all piecewise parallel, then do the line segments coincide?
3. If all line segments are coincident, then do all have zero slope?

These questions are addressed in the indicated order, since subsequent questions are statistically meaningless unless the previous null hypothesis is not rejected. Obviously, chance variation will ensure that the line segments are never perfectly parallel, coincident, or flat, and our object is to test whether the deviations from these conditions are sufficiently large as to constitute evidence that the conditions are unreasonable.

The first question, concerning piecewise parallelism, involves both within- and between-group effects. For simplicity, we assume that design structure for the between-group effects corresponds to a one-way ANOVA with k groups in a fixed-effects model. Then the appropriate matrices in (22) are

$$\boldsymbol{A}=\begin{bmatrix}1 & -1 & 0 & \dots & 0\\ 0 & 1 & -1 & \dots & 0\\ \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & \dots & 1\end{bmatrix},\qquad \boldsymbol{B}=\begin{bmatrix}1 & 0 & \dots & 0\\ -1 & 1 & \dots & 0\\ 0 & -1 & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & -1\end{bmatrix}. \tag{24}$$

Here $\boldsymbol{A}$ is $(k-1)\times k$ and $\boldsymbol{B}$ is $p\times(p-1)$. A common test statistic is the largest eigenvalue of the matrix $\boldsymbol{HE}^{-1}$, as defined in (23).

The second question, concerning coincidence of the line segments, corresponds to taking $\boldsymbol{A}$ as in (24) and $\boldsymbol{B}=(1,\dots,1)^{\mathrm{T}}$, containing p ones. If the previous test failed to reject the null hypothesis, then one proceeds as above.

Finally, the third question, concerning the flatness of all profile plots, may be tested by taking $\boldsymbol{B}$ as in (24) and setting $\boldsymbol{A}=(n_1/n,\dots,n_k/n)$, where n_i denotes the number of objects in the ith group. Once again, if the previous two tests both failed to reject their null hypotheses, then it is meaningful to apply the standard procedure from (23) to examine the flatness hypothesis.

References

Fisher (1990) is a reprint of a timeless and readable introduction to many of these topics. Kempthorne (1952) offers a very practical account of univariate statistical designs; Christensen (1996) gives a more modern treatment from the perspective of linear models. Box *et al.* (1978) give a Bayesian tone to topic. Scheffé (1959) treats this subject at a more formal level. Practical multivariate analysis is clearly presented by Morrison (1990) and Press (1982); more theoretical treatments are available in Arnold (1981) and Anderson (1984). Kres (1983) has tables for the various test criteria for multivariate statistical hypotheses. Khuri and Cornell (1987) survey modern response surface methodology; Miller (1981) reviews the literature on simultaneous inference. Speed (1987) describes a grand unification of much of design methodology at an advanced mathematical level.

V. CLUSTER ANALYSIS

There are two major kinds of cluster analysis. Case cluster analysis starts with a set of multivariate observations $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ and attempts to identify meaningful subgroups of similar data. Variable cluster analysis starts with

the same data and attempts to discover sets of variables that behave similarly. Less commonly, one encounters a third kind called block clustering, developed by Hartigan (1975); this procedure is available in BMDP and is designed to handle simultaneous case/variable clustering. However, this discussion will treat only the more customary cluster analyses.

To draw the distinction between case and variable clustering, suppose that one observes 500 different geological core samples. For each core sample, the vector $\boldsymbol{X}_i$ measures the percentage of different kinds of minerals found in different strata. Then a case cluster analysis would attempt to find clusters among the core samples; perhaps these would correspond to different regions in which the samples were taken. In contrast, a variable cluster analysis looks for strongly associated variables; if the percentage of pyroxenes in a given stratum shows large absolute correlation with the percentage of amphiboles, then together these form a cluster of variables. Cluster analysis of cases is often used to develop a taxonomy of the observations; cluster analysis of variables is used to identify major sources of variation within the sample or as a data reduction technique.

A. Case Cluster Analysis

The most common method for grouping cases is hierarchical agglomerative clustering. This yields a cluster tree; the top consists of each separate case, and these are joined together to form subclusters until, at the bottom, all cases are pooled in a common macrocluster. Since one clearly does not want (in general) for all cases to be joined in one cluster, one needs a rule for stopping the agglomeration algorithm before complete joining occurs. Also, one needs a rule to determine which two subclusters should be joined at the next stage in the tree building process.

There is no universally accepted stopping rule for the cluster algorithm, although many have been proposed. Milligan and Cooper (1985) report a large-scale simulation study under a range of cluster models. The cubic clustering criterion used in SAS appears to be the only widely available rule, and their study indicates that it works well. In practice, most users inspect the entire tree to find a scientifically interpretable level of case aggregation and report the corresponding clusters as meaningful conclusions.

Similar pluralism exists among subcluster joining rules. If, at a given stage in the tree-growing process, all of the subclusters appear ellipsoidal with common covariance matrix, then one should probably proceed to the next stage by combining the two subclusters whose centroids have smallest Mahalanobis distance, as calculated in (2). However, if the subclusters appear nonconvex, it is usually better to use a nearest neighbor joining rule. Many other rules exist and are available in standard packages; no theoretical consensus provides guidance for most practical situations. Jardine and Sibson (1971) prove that any hierarchical agglomerative method that enjoys certain properties must use a single-linkage rule for joining subclusters, but this is a technical result that seems largely irrelevant in applications.

Besides hierarchical agglomerative clustering, BMDP and S-PLUS enable k-means cluster analysis of cases, using a strategy pioneered by MacQueen (1967). Given a user-specified number k of clusters in the data, this analysis iteratively find k cluster centroids that minimize within-cluster distances and maximize between-cluster distances, where distance may be taken in terms of the Euclidean, Mahalanobis, or some other metric. The method is most appropriate when one has *a priori* knowledge of the number of clusters present in the data, but it can be a useful exploratory tool, especially when the observations arise from a mixture of several populations with distinct multivariate normal distributions.

From a theoretical standpoint, case cluster analysis is viewed as a finite mixture distribution problem. One has k populations, each with multivariate distribution function F_j having density f_j. The population to which a particular observation belongs is unknown; the *a priori* probability that a particular observation arises from the jth population is p_j. In this framework, the density function of an observed random vector $\boldsymbol{X}_i$ is the finite mixture

$$f(\boldsymbol{x}_i) = \sum_{j=1}^{k} p_j f_j(\boldsymbol{x}_i).$$

There are several strategies for finite mixture problems, depending on whether one knows k, the $p_1, \ldots, p_k$, and/or the distributions or family of distributions for $F_1, \ldots, F_k$. In general, these problems are extremely difficult; Titterington *et al.* (1985) indicate the obstacles and current strategies for each of the various situations that can arise.

B. Cluster Analysis of Variables

Cluster analysis of variables carries forward in the same spirit as hierarchical agglomerative clustering, except that the distances between variables are entirely determined by their absolute correlations. For example, suppose that at some stage of the tree-growing process one has formed m sets of variables, some of which may consist of only a single variable; denote these sets by $S_1, \ldots, S_m$. Then the single-linkage rule for joining subclusters combines the two sets S_i and S_j that attain

$$\max_{i \neq j} \max_{\boldsymbol{v} \in S_i, \boldsymbol{w} \in S_j} |\mathrm{corr}(\boldsymbol{v}, \boldsymbol{w})|,$$

which may be seen as a minimum distance rule. Alternatively, to maximize internal correlation among variables within common clusters, one might use the complete-linkage rule, combining subclusters that attain

$$\max_{i \neq j} \min_{v \in S_i, w \in S_j} |\text{corr}(v, w)|.$$

These two rules represent the extreme cases; intermediate joining rules also exist. Currently, there is no theoretically based stopping rule for variable clustering.

C. Principal Components Analysis

In many applications the aims of a cluster analysis of variables might also be served by factor analysis or a principal components analysis. These methods look for hidden structure in the data and can often discover meaningful simplifications in complex datasets. Unfortunately, both methods take a narrow view of the types of conclusions and data descriptions wanted in practice, and inferential statements make strong use of the assumption that the data arise from a multivariate normal distribution.

Principal components analysis was invented by Pearson (1901). Hotelling (1933) developed its applications and theoretical foundation. Heuristically, it makes an orthogonal transformation of the axes to an orientation that matches the shape of a p-dimensional plot of the data. This enables principal components analysis to do the following:

1. Find linear combinations of the vector components that have maximum variance, subject to certain orthogonality constraints
2. Reduce the dimensionality of the dataset
3. Transform sets of correlated random variables into uncorrelated ones.

A principal components analysis is most valuable when p, the dimensionality of the observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, is large, and when the sample covariance matrix shows strong correlations. Unlike experimental design and case cluster analysis, which focus on the mean structure of the sample, principal components analysis examines the covariance structure of the data.

In applications, it is usually wise to first standardize the data so that each component has mean 0 and unit variance. The reason for this is that principal components analysis is not scale invariant, and ideally one's inference should not depend upon the units of measurement. To achieve scale invariance one needs the unknown covariance matrix. Since generally one must use the sample covariance matrix as an approximation of the true covariance matrix, the required distribution theory is quite complex.

Assume the data vectors have sample covariance matrix $\boldsymbol{S}$, calculated according to (6). The first task is to find the linear combination

$$Y_1 = \sum_{i=1}^{p} a_1 X_i = \boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{X}_i \qquad \text{for} \qquad \mathbf{a}_1 \in \mathbb{R}^p,$$

which has the largest sample variance. The problem as posed is not well defined; it is necessary to add the constraint that $\boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{a}_1 = 1$, as otherwise the sample variance will generally be infinite. Using Lagrange multipliers, one can show that the sample variance

$$s_{Y_1}^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{i1} a_{j1} S_{ij} = \boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{S} \boldsymbol{a}_1$$

is maximized under the constraint when $\boldsymbol{a}_1$ is the normalized eigenvector of $\boldsymbol{S}$ corresponding to the largest eigenvalue of S. (For simplicity, we shall assume S has distinct, positive eigenvalues.) The corresponding linear combination is called the first principal component of the data. Moreover, the sample variance $s_{Y_1}^2$ equals the largest eigenvalue.

In immediate analogy, one can find the jth principal component of the data as the linear combination

$$Y_j = \boldsymbol{a}_j^{\mathrm{T}} X_j, \qquad \boldsymbol{a}_j \in \mathbb{R}^p$$

that maximizes the sample variance subject to the constraints that

$$\boldsymbol{a}_j^{\mathrm{T}} \boldsymbol{a}_j = 1, \qquad \boldsymbol{a}_j^{\mathrm{T}} \boldsymbol{a}_i = 0, \qquad i = 1, \ldots, j-1.$$

Under our simplifying assumptions, this is given by the eigenvector corresponding to the jth largest eigenvalue of $\boldsymbol{S}$. The sample variance of Y_j is given by the jth largest eigenvalue.

If one extracts the p largest principal components, then one has re-expressed the total variability in the dataset. Each observation can now be equivalently represented as a point in a p-dimensional space whose natural coordinates are given in terms of the p principal eigenvectors. Furthermore, from the construction, it is clear that the re-expressed random variables are now uncorrelated. This avoids difficulties arising from multicollinearity, and may enable better understanding of the data, provided that the linear combinations corresponding to the principal components are interpretable.

It often turns out that the information in a multidimensional dataset is actually captured in only a few dimensions, corresponding to the first q principal eigenvectors. When this occurs, one obtains valuable dimensionality reduction by ignoring the components of the re-expressed data that correspond to the least informative, or least variable, components. Asymptotic theory developed by Anderson (1984) enables statistical determination of

the significance of the variability in each component; this theory relies largely upon the assumption of a multivariate normal distribution. When the variance carried in a component ceases to be significant, one can truncate the re-expressed vectors and analyze a simpler dataset that contains essentially all of the information in the original data.

Principal components analysis has been generalized in several ways. One direction leads to factor analysis; other directions concern principal component inference for several related populations, where the nature of the relationship may take several forms. Also, modern computation has enabled substantial weakening of the distributional assumptions that underlie the inferential theory; these rely largely upon bootstrap methodology (Stauffer *et al.,* 1985) and robust estimation [see Hampel *et al.* (1986) for a survey of robust statistical methods].

D. Factor Analysis

Principal components analysis may not be the best way to extract latent structure from data. It only finds a linear transformation that corresponds to the data plot; in many applications, one wants a method that responds directly to the covariance structure of the data. In particular, one may want to describe only the common covariance of the observations instead of also including error variance and variability unique to a particular component of the measurement vectors. This situation arose, for example, in the measurement of human intelligence, and it prompted Spearman (1904) to develop the factor analytic model. This model enabled an analysis that described how different components of mental capability were associated, rather than clouding the inference with additional variability from other sources.

The factor analysis model assumes that each component of each observation $\boldsymbol{X}_i$ is a linear combination of q common factor values and a single latent value specific to the component. Formally, we write the model as

$$\begin{aligned} X_{i1} &= \lambda_{11} Z_{i1} + \cdots + \lambda_{1q} Z_{iq} + Y_{i1} \\ &\vdots \quad \vdots \\ X_{ip} &= \lambda_{p1} Z_{i1} + \cdots + \lambda_{pq} Z_{iq} + Y_{ip} \end{aligned} \tag{25}$$

where the λ_{jk} terms represent the importance (or the loading) of the kth common factor in determining the jth component value, Z_{ik} is the value of the kth component of the unmeasurable latent vector for the ith observation, and Y_{ij} is the contribution to the jth component of the ith observation that depends only upon the particular component. Research interest is focused on estimating the λ terms and the covariance matrix of the latent Z measurements.

Representing the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ as rows in a matrix, we can express (25) as

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Lambda} + \boldsymbol{Y},$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $n \times p$, $\boldsymbol{Z}$ is $n \times q$ and $\boldsymbol{\Lambda}$ is $p \times q$. The rows of $\boldsymbol{Y}$ are assumed to be independent multivariate normal random vectors with mean $\boldsymbol{0}$ and $\text{var}[Y_{ij}] = 1$. This enables the analyst to find maximum likelihood estimates of $\boldsymbol{\Lambda}$, the factor loadings, $\boldsymbol{\Sigma} = \text{var}[\boldsymbol{Y}]$, and $\boldsymbol{Z}$, the matrix containing each object's score on the q common factors. Also, it is usual to report the $p \times q$ factor structure matrix, which shows the estimated correlation between each original variable and each common factor.

Most factor analyses proceed in two stages. The first finds an initial estimate of the common factors; these are usually orthogonal and are often obtained by extracting the first q eigenvectors in a principal components analysis. The second stage rotates these factors to attain $\boldsymbol{L}$ from $\boldsymbol{\Lambda}$, where $\boldsymbol{L}$ is a simplified representation of the original factor loading relationships. Although there is no rule for defining this simple structure, Thurstone (1945) put forward some widely accepted desiderata:

1. Each column in $\boldsymbol{L}$ should contain at least q zeros.
2. Each row in $\boldsymbol{L}$ should contain at least one zero.
3. Any pair of columns should include multiple responses that have zero loadings in one of the columns.
4. If $q > 4$, then every pair of columns should have many zero entries in common rows.
5. If $q > 4$, every pair of columns should have only a few rows that are nonzero in both columns.

Heuristically, the simple structure segregates the responses into nearly mutually exclusive groups, and these groups have loadings that are high on a some factors, intermediate on a very small number of factors, and insignificant on the rest.

From a technical standpoint, the rotation in the second stage may be oblique rather than rigid; this leads to a nonorthogonal set of common factors. Also, the previous development assumed that q was known *a priori*. This is rarely the case, and various strategies exist for determining q from the data. These mostly rest on the idea of estimating the commonalities in the model, which are the proportions of variance in each of the components of the standardized observations that are explained by the posited q common factors. Operationally, the investigator uses the data to estimate the commonalities for a range of values of q. When there is strong factor analytic structure in the data, one typically sees that the estimated commonalities increase steadily in q until the best value of q occurs, and thereafter increasing values of q lead to

negligible improvements in the total proportion of variance explained.

Interpreting factor analytic studies is viewed by many as much more of an art than a science. If sensible understanding emerges from the factor loadings, then it is likely that actual structure has been discovered; however, it is almost always unwise to describe the results in terms of formal significance probabilities. Generally, factor analysis has value as a tool in exploratory data analysis, particularly in helping one to develop a taxonomy of the kinds of underlying factors that influence the observations.

References

Hartigan (1975) is the classic reference in cluster analysis of cases, and Kaufman and Rousseeuw (1990) is a modern treatment of very similar topics. Anderson (1984) gives much of the multivariate theory that underlies such variable clustering methods as principal components and factor analysis. Flury (1988) and Jolliffe (1986) provide good treatments of some of the extensions of principal component methodology, and Press (1982) gives a careful exposition of statistical methods for factor analysis and their relationship to principal component methods. Jöreskog (1970) presents these methods in the context of a more general approach to the analysis of covariance structures involving latent variables.

VI. CONTINGENCY TABLES

When the variables of interest in a study are categorical instead of continuous, then the data can be arrayed as a set of counts cross-classified according to both the explanatory and response variables in a form known as a contingency table. As is the case with the general multivariate linear model, one is interested in the relationships among the response variables, conditional on the explanatory ones, as well as the effects of the explanatory variables on the response. For example, suppose one takes a random sample of 100 automobiles, and classifies each as to (1) whether it required repair within the first year and (2) the day of the week on which it was manufactured. This yields a simple 2×7 two-way contingency table; a three-way table could be built if one also classified each car with respect to manufacturer.

The formal analysis of contingency table data goes back at least to the turn of the century and the pioneering work of Pearson (1900) and Yule (1900). A classic paper by Birch (1963) firmly established the likelihood approach to estimation for contingency table models that are linear in the logarithmic scale of the expectations of the cross-classified counts. Many of the ideas in the subsequent literature parallel those for the general linear model, although the interpretation of the loglinear model parameters has special meaning in the context of contingency tables.

A. Some Basic Notation and Results

In general one works with a vector of observed counts $\boldsymbol{x} = (x_1, x_2, \ldots, x_t)$ that are realizations of the vector of random variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_t)$. The models of interest are expressed in terms of the vector of expected counts $\boldsymbol{m} = (m_1, m_2, \ldots, m_t)$, where $m_i = \mathrm{E}(X_i)$ for $i = 1, 2, \ldots, t$, or more directly in terms of the logarithms of the expected counts $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_t)$, where $\mu_i = \log m_i$ for $i = 1, 2, \ldots, t$. When there are two binary variables of interest $t = 4$ and one can array the counts in the form of a 2×2 table:

x_1	x_2	$\leftrightarrow$	x_{11}	x_{12}
x_3	x_4		x_{21}	x_{22}

where the second tabular version uses multiple subscripts to denote the rows and columns of the table.

In a 2×2 table, the counts may be completely unconstrained, sum to a fixed total $n = \sum x_i$, or have fixed row or column totals. For the general unconstrained case one often assumes that the X_i have independent Poisson distributions with likelihood function

$$\prod_{i=1}^{t} \frac{m_i^{x_i} e^{-m_i}}{x_i!}.$$

When the counts sum to a fixed total n, then one can take $\boldsymbol{X}$ to have a multinomial distribution with likelihood function

$$\frac{n!}{x_1! x_2! \ldots x_t!} \prod_{i=1}^{t} \left(\frac{m_i}{n}\right)^{x_i}.$$

If one partitions the t cells into r sets of sizes $\{t_j\}$, with $t = \sum t_j$, and fixed sample sizes $\{n_j\}$, with $n = \sum n_j$, one has a separate multinomial distribution for each set of cells. The corresponding distribution for the table is said to be product multinomial with likelihood function

$$\prod_{j=1}^{r} \frac{n_j!}{x_{1j}! x_{2j}! \ldots x_{t_j j}!} \prod_{i=1}^{t_j} \left(\frac{m_{ij}}{n_j}\right)^{x_{ij}}.$$

Suppose that X follows the Poisson sampling model. Then the conditional distribution of $\boldsymbol{X}$ given n is multinomial and that of $\boldsymbol{X}$ given $\{n_j\}$ is product-multinomial.

In the context of the automobile manufacturing example given at the beginning of this section, the appropriate likelihood function is the multinomial, since the number

of cars was fixed at 100. However, if the cars selected were those whose owners obtained a driver's license upon a particular day, then the Poisson model would be appropriate; and, if one drew a balanced sample that took exactly 20 cars manufactured on each of the 7 days of the week, then one should use the product multinomial model.

By re-expressing $\boldsymbol{\mu} = \log \boldsymbol{m}$ using a linear model, $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$, one focuses on the estimation of the loglinear parameters in $\boldsymbol{\theta}$. Then $\boldsymbol{\mu}$ is said to lie in the constrained linear subspace M. Under the model M, the summary statistics used to describe the likelihood function is called the minimal sufficient statistics and is given by the projection of $\boldsymbol{x}$ onto M, i.e., $P_M \boldsymbol{x}$. For the models considered in this section these projections turn out to be sums of components of $\boldsymbol{x}$. For instance, in the 2×2 table case suppose M is the model that takes that the variable corresponding to the rows to be independent of the variable corresponding to the columns, i.e., $m_{ij} = m_{i+} m_{+j}$, where the "+" indicates summation over the corresponding subscript. Then $P_M \boldsymbol{x} = \{x_{1+}, x_{2+}, x_{+1}, x_{+2}\}$. For the automobile manufacturing problem, this model of independence corresponds to asserting that day of manufacture is entirely unrelated to whether the car requires repairs within the first year.

The maximum likelihood estimates (MLEs) $\hat{\boldsymbol{m}}$ of the expected cell counts $\boldsymbol{m} = \boldsymbol{m}(\boldsymbol{\theta})$ under Poisson sampling, when they exist, are given by setting the minimal sufficient statistics equal to their expected values, i.e.,

$$P_M \hat{\boldsymbol{m}} = P_M \boldsymbol{x}.$$

Furthermore, these estimates of the expected cell counts are the same as those for the other two sampling schemes for comparably defined models. Thus, suppose one lets M^* be the loglinear model for $\boldsymbol{m}$ under product-multinomial sampling corresponding to the model M under Poisson sampling such that the multinomial constraints fix a subset of the parameters θ used to specify M. Then the MLEs of $\boldsymbol{m}$ under product-multinomial sampling for the model M^* are the same as the MLEs of $\boldsymbol{m}$ under Poisson sampling for the model M. These results were well known for the 2×2 table, going back to the work of Fisher in the 1920s, and then were extended to the three-way contingency table case by Birch (1963) and later generalized by Haberman (1974) and others.

To decide whether a model fits a vector of observed counts, Pearson (1900) introduced the chi-squared goodness-of-fit statistic. If $\hat{\boldsymbol{m}}$ is the MLE of $\boldsymbol{m}$ under a loglinear model and if M is correct, then the statistic

$$X^2 = \sum_{i=1}^{t} \frac{(x_i - \hat{m}_i)^2}{\hat{m}_i}$$

has an asymptotic χ^2 distribution with $t - s$ degrees of freedom, where s is the total number of independent constraints implied by the model M (or by M^* under the product multinomial constraints). If M is not correct, then the value of X^2 tends to be larger than one expects from a random variable with χ^2_{t-s} distribution.

Both $X^2 = \sum_{i=1}^{t} (x_i - \hat{m}_i)^2 / \hat{m}_i$ and the likelihood ratio statistic $G^2 = 2 \sum_{i=1}^{t} x_i \log(x_i / \hat{m}_i)$ are examples of goodness-of-fit measures based on the power-divergence family of statistics,

$$I^{\lambda}(\boldsymbol{p} : \boldsymbol{q}) = \lambda^{-1}(1 + \lambda)^{-1} \sum_{i=1}^{t} p_i \left[\left(\frac{p_i}{q_i} \right)^{\lambda} - 1 \right], \quad -\infty < \lambda < \infty,$$

studied by Read and Cressie (1988). The quantity $I^{\lambda}(\boldsymbol{p} : \boldsymbol{q})$ measures the divergence of the probability distributions $\boldsymbol{p} = (p_1, p_2, \ldots, p_t)$ and $\boldsymbol{q} = (q_1, q_2, \ldots, q_t)$. When one lets $\boldsymbol{x}/n = \boldsymbol{p}$ and $\hat{\boldsymbol{m}}/n = \boldsymbol{q}$ and the model M is correct, $2n I^{\lambda}(\boldsymbol{x}/n : \hat{\boldsymbol{m}}/n)$ has an asymptotic χ^2 distribution with $t - s$ degrees of freedom. For each different value of λ, one gets a test statistic that is sensitive to different alternative models for the expected counts. Setting $\lambda = 1$ corresponds to the statistic X^2 and defining the quantity in the limit as $\lambda \to 0$ yields G^2.

B. Applying the Basic Theory to Log-Linear Models for Three-Way Tables

The three-dimensional case helps to make concrete these theoretical results. Suppose one has an $I \times J \times K$ table of counts $\{x_{ijk}\}$ with a corresponding table of expectations $\{m_{ijk}\}$. The most general model in the logarithmic scale is

$$\mu_{ijk} = \log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}, \quad (26)$$

where the parameters (u-terms) are subject to appropriate linear constraints similar to those for ANOVA models:

$$\begin{aligned} u_{1(+)} &= u_{2(+)} = u_{3(+)} = 0, \\ u_{12(+j)} &= u_{12(i+)} = u_{13(+k)} = u_{13(i+)} = u_{23(+k)} \\ &= u_{23(j+)} = 0, \\ u_{123(+jk)} &= u_{123(i+k)} = u_{123(ij+)} = 0. \end{aligned}$$

By setting various parameters in (26) equal to zero one can get different log-linear models M.

The simplest log-linear model tests

$$\mathrm{H}_0: \quad \mu_{ijk} = \log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}.$$

This corresponds to testing whether $u_{12(ij)} = u_{13(ik)} = u_{23(jk)} = u_{123(ijk)} = 0$, and implies the complete independence of variables 1, 2, and 3. Alternatively, one can include one or more two-factor terms yielding

$$\mathrm{H}_{01}: \quad \mu_{ijk} = \log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)},$$

which corresponds to the independence of variable 3 from variables 1 and 2 jointly, and is expressible in more standard multiplicative form as

$$m_{ijk} = m_{i++}m_{+j+}m_{++k}.$$

With two two-factor terms one has

$$\mathrm{H}_{02}: \quad \mu_{ijk} = \log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)},$$

which corresponds to the conditional independence of variables 2 and 3 given variable 1, and is expressible in multiplicative form as

$$m_{ijk} = m_{ij+}m_{i+k}/m_{i++}.$$

Finally, one has

$$\mathrm{H}_{03}: \quad \mu_{ijk} = \log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)},$$

which implies that there is no second-order interaction among variables 1, 2, and 3. Model H_{03} is not expressible in a multiplicative form for m_{ijk} in terms of the corresponding two-way totals m_{ij+}, m_{i+k}, and m_{+jk}.

Under Poisson sampling or multinomial, for the no-second-order-interaction model H_{03}, the minimal sufficient statistics are $P_M\boldsymbol{x} = [\{x_{ij+}\}, \{x_{i+k}\}, \{x_{+jk}\}]$. Also, the MLEs $\hat{\boldsymbol{m}}$ of the expected cell counts are the solution of equations

$$\begin{aligned} \hat{m}_{ij+} &= x_{ij+} \qquad \text{for all} \quad i, j, \\ \hat{m}_{i+k} &= x_{i+k} \qquad \text{for all} \quad i, k, \\ \hat{m}_{+jk} &= x_{+jk} \qquad \text{for all} \quad j, k. \end{aligned}$$

The solution of these equation requires some form of iteration; sometimes no unique solution exists. Alternatively, one could maximize the likelihood function directly, using the Newton–Raphson algorithm or some similar technique.

If one has generated the counts using product-multinomial sampling with the 1, 2-margin fixed, one has $x_{ij+} = n_{ij+}$ fixed for all i, j. For this situation one can convert model H_{03} into a linear logistic model. Suppose that variable 3 is binary. Then one can re-express the model in terms of one for the log odds (logit) of responses for variable 3 given variables 1 and 2:

$$\begin{aligned} \log(m_{ij1}/m_{ij2}) &= u + u_{1(i)} + u_{2(j)} + u_{3(1)} + u_{12(ij)} \\ &\quad + u_{13(i1)} + u_{23(j1)} - \left[u + u_{1(i)} + u_{2(j)} \right. \\ &\quad \left. + u_{3(2)} + u_{12(ij)} + u_{13(i2)} + u_{23(j2)}\right] \\ &= w + w_{1(i)} + w_{2(j)}, \end{aligned} \tag{27}$$

where $w = 2u_{3(1)}$, $w_{1(i)} = 2u_{13(i1)}$, and $w_{2(j)} = 2u_{23(j1)}$. This model has the same form as the logistic regression model of Section II [see expression (19)]. The results of the previous subsection imply that the maximum likelihood estimators for the expected counts under this model are the same as those under the no-second-order-interaction model and thus satisfy the equations in expression (27). For an $I \times J \times K$ table and model H_{03}, there is a total of $IJK - (I-1)(J-1)(K-1)$ parameters. Thus the degrees of freedom used to test the goodness of fit equals $(I-1)(J-1)(K-1)$.

TABLE III Observed Counts for $\{x_{ijk}\}$ for Industrial Study of the Rolling of Ingots for Different Soaking and Heating Times

Soaking time	Counts for given heating times				Total counts
	7	14	27	51	
1.0	10, 0	31, 0	55, 1	10, 3	106, 4
1.7	17, 0	43, 0	40, 4	1, 0	101, 4
2.2	7, 0	31, 2	21, 0	1, 0	60, 2
2.8	12, 0	31, 0	21, 1	0, 0	64, 1
4.0	9, 0	19, 0	15, 1	1, 0	44, 1
Totals	55, 0	155, 2	152, 7	13, 3	375,12
Grand total 387					

As an example, Table III contains the data from Cox and Snell (1989) for an industrial investigation designed to study effects of soaking time (variable 1) and heating time (variable 2) on whether on not an ingot is ready for rolling. The first entry in each cell is the number of ingots ready for rolling; the second entry is the number not ready.

The no-second-order-interaction model of H_{03} fits these data extremely well; $G^2 = 11.275$ with 11 degrees of freedom (the usual number 12 is reduced by 2 because the count of the number of ingots not ready for rolling in the first column of Table III is 0 and the number of ingots in the fourth row and column is 0). One can compare this value to the tabulated values of a χ^2 distribution with 11 degrees of freedom, and the observed values fall far from the right-hand tail of the distribution.

Because the two explanatory variables are ordered, a natural extension of model (27) that one might wish to explore expresses the log-odds as a simple linear function of the soaking time x_1 and the heating time x_2:

$$\mu = \theta + \theta_1 x_1 + \theta_2 x_2$$

[cf. expression (19)]. Now one no longer needs to worry about the linear constraints on the full version of model (27). The estimated equation using the method of maximum likelihood given by Cox and Snell (1989, p. 66) is

$$\mu = 5.559 - 0.0568x_1 - 0.0820x_2.$$

The coefficient of x_1 has a standard error of 0.0331, and thus the observed level of significance for testing $H_0: \theta_1 = 0$ versus $H_a: \theta_1 \neq 0$ is just under 0.10.

C. Log-Linear Models for Multi-Way Tables

The ANOVA-like log-linear models and theoretical results are useful for multi-way tables of counts. As in the three-way case, minimal sufficient statistics are marginal totals of the full table. All independence or conditional independence relationships are representable as log-linear models. These have estimated expected cell counts that can be computed directly. For all other log-linear models, we require iterative methods to compute the maximum likelihood estimates. Detail are available in Bishop *et al.* (1975).

D. Graphical Log-Linear Models

A special subset of log-linear models for multi-way tables has received considerable attention over the past decade. Darroch *et al.* (1980) define the class of graphical log-linear models as those satisfying a set of conditional independence relationships.

Suppose one considers a p-way array corresponding to the vector of p discrete random variables $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_p)$ and let $P = \{1, 2, \ldots, p\}$ be the corresponding set of vertices. Further, let $\boldsymbol{Y}_{P/\{i,j\}}$ denote the set of $p-2$ variables excluding Y_i and Y_j. A standard notation for representing the conditional independence of Y_i and Y_j given the remaining variables is

$$Y_i \perp Y_j \mid \boldsymbol{Y}_{P/\{i,j\}}. \tag{28}$$

This conditional independence model is also itself a log-linear model for the p-way array. Graphical log-linear models correspond to the simultaneous occurrence of several such conditional independence models. They have special attractive properties in addition to the interpretation associated with these conditional independence relationships (for details, see Whittaker, 1990) and are related to the spatial theory of Markov random fields.

The conditional independence graph of $\boldsymbol{Y}$ is the undirected graph $G = (P, E)$ such that the edge between i and j is not in the edge set E if and only if $Y_i \perp Y_j \mid \boldsymbol{Y}_{P/\{i,j\}}$. There is a one-to-one correspondence between independence graphs and graphical log-linear models. Given an independence graph G, the corresponding graphical log-linear model is the one in which all u-terms containing the pairs of coordinates corresponding to edges not in E are taken to be identically zero.

Example. Suppose one constructs an independence graph for an eight-way contingency table as in Fig. 1. The edge set for this graph is

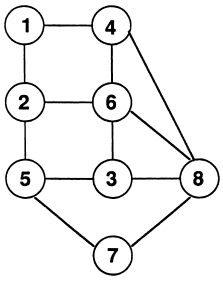


FIGURE 1 Example of an independence graph in an eight-way contingency table.

$$E = \{(1,2), (1,4), (2,6), (2,5), (3,5), (3,6), (3,8), (4,6), (4,8), (5,7), (6,8), (7,8)\}$$

and the corresponding graphical log-linear model satisfies 16 conditional independence relationships of the form given in (28). Alternatively, one can describe the loglinear model by setting the 12 two-factor u-terms corresponding to the edges not in E equal to zero, as well as all higher order terms containing them. The resulting model contains no four-factor u-terms.

References

Agresti (1990) and Fienberg (1980) give accessible introductions to the methodology for the analysis of multi-way contingency tables using log-linear models and estimation via the method of maximum likelihood. Bishop *et al.* (1975) and Haberman (1974) provide further technical details and methods for related structures and models. McCullagh and Nelder (1989) present log-linear models as a special case of GLIM techniques, and Santner and Duffy (1989) describe alternative methods of estimation; both consider the problem of over- and underdispersion in binary logistic response models. Douglas and Fienberg (1991) present a detailed review of log-linear and related models for cross-classified categorical data involving ordinal variables. Finally, Whittaker (1990) and Lauritzen (1996) present extensive treatments of graphical log-linear models and their continuous counterparts.

VII. DISCRIMINANT ANALYSIS

In discriminant analysis one has a preliminary learning sample that consists of n objects, each of which is measured with respect to p variables and each of which is known to belong to one of j categories. The goal is to use the learning sample to construct a classification function for future objects whose true category is unknown, but for which one can take measurements on each of the p variables.

Historically, discriminant analysis developed from a problem in botany. Fisher (1936) had measurements of sepal width, sepal length, petal width, and petal length for a sample of wild irises, each of which had been classified into one of three species by an expert botanist. Using these data, he calculated the equations for the two hyperplanes that best separated the vectors of observations in $\mathbb{R}^4$ into regions corresponding to different species. These hyperplanes were the classification (or discriminant) functions used to categorize future irises; specifically, a new iris would be measured on each of the four characteristics and then classified as belonging to the species within whose region the data vector lay.

A. Linear Discriminant Analysis

Fisher's original strategy can be formalized as follows: assume there are k populations, each having unknown mean measurement vector $\boldsymbol{\mu}_i$ and common unknown covariance matrix $\boldsymbol{\Sigma}$. The training sample consists of k independent samples, one from each population. The ith sample has n_i vector observations in $\mathbb{R}^p$, all independent; these are denoted by $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in_i}$ for $i = 1, \ldots, k$. The unknown parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ are estimated as

$$\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{X}}_i$$

$$\hat{\boldsymbol{S}} = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij} - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{X}_{ij} - \hat{\boldsymbol{\mu}}_i)^{\mathrm{T}}$$

where $n = n_1 + \cdots + n_k$. From these, the discriminant function corresponding to the hyperplane that best separates populations r and s in the training sample is

$$W_{rs}(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{S}^{-1} (\hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_s) - \tfrac{1}{2} (\hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_s)^{\mathrm{T}} \boldsymbol{S}^{-1} (\hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_s).$$

Thus for a new observation with measurement vector $\boldsymbol{X}$, the classification rule assigns it to population r if and only if $W_{rs}(\boldsymbol{X}) > 0$ for all $s \neq r$.

This classification rule is equivalent to assigning the observation to the group whose sample mean minimizes the estimated squared Mahalanobis distance between $\boldsymbol{X}$ and $\hat{\boldsymbol{\mu}}_r$. The procedure works well in practice, and is asymptotically optimal in a certain technical sense when the assumptions hold and the distribution for each population is normal. However, its properties are less certain when different groups have different covariance matrices or the underlying distributions are nonnormal.

A key issue in discriminant analysis is the estimation of the misclassification probabilities. Ideally, this would be done by applying the classification rule to a large number of new observations and then determining how often the group assignment was incorrect. In practice, it is often not feasible to determine the true memberships of new observations. Therefore, most computer software packages use cross-validation to enable the practitioner to estimate the misclassification rates from the training sample (Section II contains additional discussion of cross-validation methodology). In this application, cross-validation separates each observation in turn from the training sample, recomputes the discriminant functions from the remaining observations, and then uses the recomputed rule to classify the separated observations. This is done for all observations, and one counts n_{rs}, the number of times an observation from group r is misclassified as belonging to group s. From this, the estimate of the probability that linear discriminant analysis assigns a member of group r to group s is n_{rs}/n_r. Cross-validation is more accurate and less optimistic than the naive procedure, which estimates error rates from misclassifications obtained from the original assignment rule. Alternatives to cross-validation include bootstrap estimation, proposed by Efron (1983), and smoothing techniques described by Snapinn and Knoke (1985).

Another issue in implementing a discriminant analysis involves the prior probabilities of the different groups. In some applications the proportions of the different groups among new objects to be classified are expected to be similar to the proportions found in the training sample. In other applications, the proportions will be different. The accuracy of the classification rule for future observations will improve if it is customized to the situation. Most software packages permit the user to indicate the expected proportions.

Fisher's discriminant analysis is rarely used in the form described; most applications are better served by a variation that builds the discrimination functions through stepwise selection from the set of recorded variables. Very often, the number of components p in the measurement vector is large relative to the sample size, and not all of the measurement variables make distinct contributions to the classification problem. In order to prevent the procedure from overfitting the data, it is desirable to ensure that each variable used in the discriminant function explains a separate and nonnegligible proportion of the variation in group membership. This yields simpler classification rules with smaller misclassification rates. However, the stepwise discriminant analysis procedure is nonlinear, and few of its theoretical properties are understood.

B. CART Analysis

Discriminant analyses work well when the underlying assumptions are reasonably satisfied. Even with gross violations, as when the covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k$ are markedly unequal, it is often possible to develop asymptotically accurate procedures that show good performance in small sample settings. In practice, the three

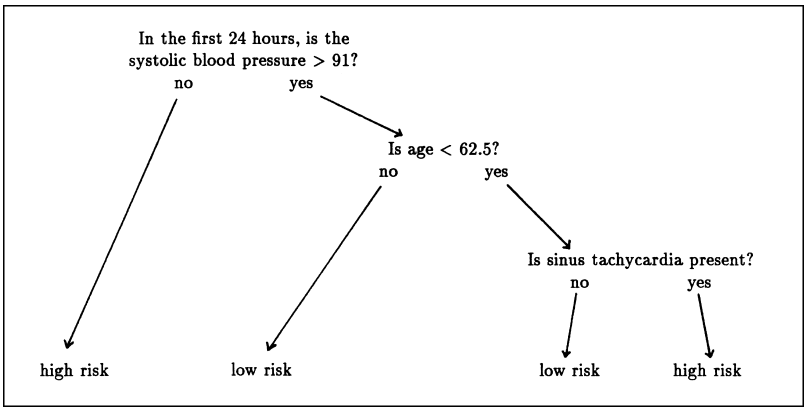


FIGURE 2 CART diagram for classifying emergency room patients into high- and low-risk groups with regard to myocardial infarction.

chief drawbacks of discriminant analysis are that it is unable to use categorical data, the discriminant functions are difficult to interpret in scientifically meaningful ways, and the theoretical framework relies on the assumption that the data arise from a mixture of normal distributions.

CART analysis is a modern, computer-intensive technique that attempts to deal with these deficiencies of discriminant analysis. CART is an acronym for "classification and regression trees"; in the context of building a classification rule, CART uses the training sample recursively to partition the space of possible measurements. The resulting rule can be represented as a decision tree, and this is generally viewed as more natural than classification according to discriminant functions. In this form, the method was developed by Breiman *et al.* (1984), and has been marketed in a commercial software package.

Operationally, CART partitions the training sample into increasingly homogeneous groups, thereby inducing a partition on the measurement space. At each stage, three criteria are considered in determining the next split for the training sample cases inside the current partition:

1. Is $X_i \leq d$ (univariate split)?
2. Is $\sum_{i=1}^{P} c_i x_i \leq d$ (linear combination split)?
3. Does $x_i \in A$ (categorical split, used when x_i is a categorical variable)?

CART searches all possible values of d, all coefficients $\{c_i\}$, and all subsets A of the possible category values to find the split which best separates the training sample cases under consideration into two groups of maximally improved homogeneity. Improvement is assessed using either Gini's index of diversity or a "twoing rule." Typically, CART splits until all cases in the training sample fall into separate partitions, and then uses a cross-validation assessment to eliminate previous partitions, thereby reducing the problem of overfit. This automatically includes estimation of misclassification probabilities and user-specified priors on category membership.

The result of a CART analysis is a classification tree. Figure 2, from Breiman *et al.* (1984), shows a tree built using a training sample from a Santa Barbara, California, emergency room. On admission, the patients complained of chest pain; doctors recorded their medical history, current profile, and eventual diagnosis. The goal of the study was to develop a classification rule that distinguished patients at high risk of myocardial infarction from those at low risk. Based on the data, CART generated the decision tree rule shown in Fig. 2. To use this tree, imagine that a new patient with chest pain enters the emergency room. First, sphygmomanometry determines whether the minimum systolic blood pressure over the initial 24-hr period is greater than 91. If not, then the patient is at high risk; otherwise, one checks the patient's age. If this is less than 62.5 years, then the patient is classified as low risk; otherwise, the doctor determines whether sinus tachycardia is present. If it is, then the patient is classified as high risk, and otherwise is declared to be at low risk. Cross-validation indicated the CART tree performed as well as discriminant analysis at distinguishing patients, and it also produced a diagnostic procedure which some view as more medically meaningful.

CART methods can also be applied to build regression models, although their popularity there has been less pronounced. Recently, CART methodology has been generalized, so that the entire strategy emerges as a special case of MARS, the procedure used for fitting locally low-dimensional structure via a generalized additive model (see Section I). However, this generalization forgoes the graphic summary illustrated in Fig. 2.

References

Lachenbruch (1975) and Flury (1997) offer good surveys of applications and topics in this area. Morrison (1990) addresses the problem from the perspective of multivariate normal data, and Dillon and Goldstein (1978) develop discriminant analysis methods that use discrete data. Anderson (1984) and Press (1982) detail the theory of discriminant analysis.

VIII. TIME SERIES ANALYSIS

Time series data arise when observations show serial correlation. Most commonly, such data occur when measurements are taken over time; for example, a control engineer might monitor the temperature of a blast furnace at 5-min intervals. However, this defining property of serial correlation would also be present in measurements of rail wear at 20-ft intervals along railroad track, or selenium concentrations assayed in mile-long steps on a river.

Theoretical analysis of time series data is computationally intensive and model sensitive. Recent research has attempted to weaken the modeling assumptions needed for forecasting and parameter estimation, but this work has not yet engendered a fundamental breakthrough that permits cookbook application of time series methodology. Nonetheless, some basic models are widely used; software that implements these models is available in all major statistical packages.

A. Detrending Data

The first step in a time series analysis is to achieve stationarity by detrending the data. Formally, a time series $\{Y(t)\}$ is strictly stationary if

$$P[Y(t_1) \le y_1, \ldots, Y(t_k) \le y_k] = P[Y(t_i + s) \le y_1, \ldots, Y(t_k + s) \le y_k]$$

for all k, s, and $\{t_i\}$. This is theoretically powerful, but rarely justified in practice. It is more realistic to hope that the data are weakly stationary, which occurs when

$$\begin{aligned} E[Y(t)] &= \mu \quad \forall t, \\ \text{var}[Y(t)] &= \sigma^2 \quad \forall t, \\ \text{cov}[Y(t), Y(t+s)] &= \gamma_s \quad \forall t, s. \end{aligned}$$

Both types of stationarity imply that the character of the series does not depend upon the time at which it is measured. Thus a stationary series shows no trend and the variance is constant throughout. Although one cannot prove stationarity, there are tests that detect specific kinds of nonstationarity (Brown *et al.,* 1975). Detrending techniques are used to adjust a dataset so that it conforms to a stationary model amenable to time series analysis.

There are three basic strategies for detrending a dataset:

1. Fit a model to the trend in the raw data and then subtract this from the time series.
2. Replace the raw values by the differences of adjacent values (this differencing operation may be repeated several times, until the new dataset passes some test of stationarity).
3. Transform the data, usually to stabilize the variance.

The first method implements a global detrending; the second performs a more flexible local correction. The third approach depends upon ideas in Section II, but will not be considered here.

Fitting a model is conceptually simple and often approximately correct in applications that involve a well-understood physical process. The usual model is

$$Y(t) = m(t) + \xi(t),$$

where $\{Y_t\}$ denotes the measurements, which we assume to be taken at equally spaced intervals indexed by t, $m(t)$ is a deterministic trend, and $\{\xi(t)\}$ is a stationary time series. As a simple example, one might analyze Newton's experiment on the rate at which a heated cannon ball cools. The deterministic idealization relates the surface temperature u to time t by the differential equation

$$\frac{du}{dt} = k(u - u_o), \tag{29}$$

where u_o is the ambient temperature and k depends upon the physical properties of the cannon ball. Newton recorded the observed temperatures $U(t_1), \ldots, U(t_n)$; these surely had correlated errors, so a plausible statistical model based on the solution to (29) is

$$U(t_i) = ce^{t_i k} + U_o + \xi(t_i). \tag{30}$$

To detrend this time series, one could estimate c, k, and u_o via nonlinear regression analysis and then subtract the estimated deterministic component from each observation. The resulting residuals $\{\hat{\xi}(t)\}$ may comprise an approximately stationary series.

Detrending methods must reflect the physical model. For example, in meteorology one can model diurnal high temperature $Y(t)$ as

$$Y(t) = p(t) + s(t) + \xi(t), \tag{31}$$

where $p(t)$ captures slow multiyear change as a random walk with polynomial drift, $s(t)$ is the seasonal component, equal to the previous year's value plus a disturbance term independent of both $p(t)$ and $\xi(t)$, and $\xi(t)$ is independent normal error with mean 0. One can remove the polynomial trend by taking d successive differences of adjacent values, where d is the degree of the polynomial. The

seasonal component is detrended by subtracting values separated by 1 year (the same strategy also removes daily, weekly, or monthly periodicities). What remains should be a stationary series whose analysis may enable accurate forecasts.

These two detrending examples correspond to two distinct approaches to time series analysis. Differencing takes the time-domain perspective, pioneered by Whittle (1951) and broadly elaborated upon by Box and Jenkins (1976). Regression modeling conforms to the spirit of the frequency- or harmonic-domain perspective, largely developed by Weiner (1949), in which one performs a Fourier decomposition of the data and discards terms that contribute meager explanation. This Fourier decomposition is comparable to regression in that finding the Fourier transform of the discrete observations is conceptually similar to fitting the model

$$Y(t) = \theta_o + \sum_{i=1}^{k} \psi_i \cos \lambda_i t + \phi_i \sin \lambda_i t + \xi(t).$$

There is no theoretical basis for preferring one style of time series analysis over the other.

B. Time-Domain Models

The time-domain methods attempt to express $Y(t)$ as a function of previous values and uncorrelated noise. The basis for this representation is the linear autoregressive moving average (ARMA) model. If differencing is used to achieve stationarity, this can be captured in the same framework as an autoregressive integrated moving average (ARIMA) model. This ARIMA generalization is very practical but not conceptually fundamental, so we shall avoid addressing it, by assuming that $\{Y(t)\}$ is weakly stationary.

There are three kinds of ARMA models: moving average models of order q, denoted by MA(q); autoregressive models of order p, denoted by AR(p); and autoregressive moving average models of orders p and q, denoted by ARMA(p, q). To discuss these, let $\{\xi(t)\}$ be a white noise process, defined as satisfying

$$\mathrm{E}[\xi(t)] = 0 \quad \forall t, \qquad \mathrm{cov}[\xi(t_i), \xi(t_j)] = 0 \qquad \forall t_i \neq t_j.$$

The time series $\{Y(t)\}$ is assumed to result from the application of a linear filter to this white noise process; i.e.,

$$Y(t) = \sum_{i=0}^{\infty} \psi_i \xi(t - i), \tag{32}$$

where $\psi_o = 1$ and the coefficients $\{\psi_j\}$ form a finite or infinite convergent sequence. Also, we define the backshift operator B, which steps back the time index of any process term one unit; e.g., $B^j Y(t) = Y(t - j)$.

The MA(q) model assumes that $Y(t)$ depends only on the recent past of the unobservable white noise process. Specifically,

$$Y(t) = \sum_{i=0}^{q} \psi_i \xi(t - i) = \Psi_q(B)\xi(t), \tag{33}$$

where $\Psi_q(B)$ is a polynomial operator in B of degree q; it is obtained from the symbolic representation of the backshift operator as the argument in the polynomial.

The AR(p) model assumes $Y(t)$ depends upon the infinite past of the white noise process through dependence on p previous terms of the time series. Thus

$$Y(t) = \xi(t) + \sum_{j=1}^{p} \phi_j Y(t - j) = \xi(t) + \Phi_p(B)Y(t), \tag{34}$$

where $\Phi_p(B)$ is a polynomial operator in B of degree p. (Differencing in this model corresponds to zeros of the polynomial Φ_p that lie on the unit circle; zeros outside the unit circle imply intrinsic nonstationarity, so it is of practical importance to check whether all zeros of Φ_p are inside the unit disk.)

The ARMA(p, q) model combines models (33) and (34), so that

$$Y(t) = \xi(t) + \sum_{j=1}^{p} \psi_j Y(t - j) + \sum_{i=1}^{q} \phi_i \xi(t - i), \tag{35}$$

or, in backshift notation, $\Phi_p(B)Y(t) = \Psi_q(B)\xi(t)$. This model is quite flexible and can fit many datasets, but one must be wary of fitting too many parameters, which leads to poor prediction.

The two basic tools for identifying a time-domain model are the estimated autocorrelation function (ACF) and partial autocorrelation function (PACF). The former represents the sample correlations r_s between all observations exactly s time units parts, so

$$r_s = \frac{\sum_{i=1}^{n-s} (Y(t_i) - \bar{\boldsymbol{Y}})(Y(t_{i+s}) - \bar{\boldsymbol{Y}})}{\sum_{i=1}^{n} (Y(t_i) - \bar{\boldsymbol{Y}})^2},$$

where $\boldsymbol{Y}$ is the mean of the series. Similarly, the estimated PACF gives the sample partial correlations (see Section II) between all observations exactly s time units apart, controlling for the intervening observations.

When the MA(q) model fits, then $r_1, \ldots, r_q$ are relatively large, but subsequent values are negligible. When the AR(p) model fits, one finds that just the first p partial correlations in the PACF are relatively large. Both the ACF and PACF are used to fit ARMA(p, q) models. Usually this model identification process is somewhat exploratory, and the choice may be revised late in the analysis. If the white noise process is Gaussian, then there is distribution theory that enables precise judgment about which terms

in the ACF and PACF are negligible; otherwise one must rely on empirical guidelines. More elaborate model identification procedures are available, such as Akaike's (1973) information criterion. These balance the explanatory value of a complex model against the loss of predictive power incurred by fitting too many parameters, using tools from information theory.

After identifying a model, one estimates the corresponding parameter values. There are several alternative parametrizations of a given model; for example, one can estimate the filter coefficients $\{\psi_i\}$ in (32) or the polynomial coefficients for the backshift operators in (35). Similarly, there are many estimation methods, and the choice of these is driven by the application and the parametrization. For the AR(p) model, the Yule–Walker equations enable linear estimation of $\Phi_1, \ldots, \Phi_p$; however, their tractability is undermined by their relatively poor performance for prediction. Modern computer-intensive calculation enables maximum likelihood, Bayesian, or iterative nonlinear least squares estimation, and the results are typically more stable.

C. Forecasting

The most common application of time series analysis is to forecast future observations. Depending upon the model and the parametrization, one may need to estimate not only the parameters, but also the terms in the white noise series. For example, if one is working with the representation in (32), then the value j steps ahead will be

$$Y(t+j) = \xi(t+j) + \sum_{i=0}^{\infty} \Psi_i \xi(j+t-i)$$

and this is estimated by

$$\hat{Y}(t+j) = \sum_{i=j}^{\infty} \hat{\Psi}_i \hat{\xi}(j+t-i)$$

since the expectations of all currently unrealized terms in the white noise process are 0. The estimated variance of the error in this forecast is

$$s_j^2 = \left(1 + \hat{\Psi}_1^2 + \cdots + \hat{\Psi}_{j-1}^2\right)\hat{\sigma}_\xi^2,$$

where $\hat{\sigma}_\xi^2$ is the sample variance of the white noise, which is calculated from the residuals $\hat{\xi}(t) = Y(t) - \hat{Y}(t)$ obtained by fitting the model to the observed series.

References

For time-domain models, Box and Jenkins (1976) is the classic reference. The frequency-domain models are described in Brockwell and Davis (1991) and Brillinger (1981). Brockwell and Davi (1996) provides an elementary introduction to time series methodology; Anderson (1984) develops the area from the perspective of multivariate analysis, and Christensen (1991) emphasizes connections with the general linear model and coordinate-free methodology.

IX. STATISTICAL SOFTWARE PACKAGES

There are many packages, libraries, and languages for computer implementation of multivariate analysis. This brief survey restricts attention to the most popular of these along with a few of the special-purpose packages that permit analyses not available from other vendors. We note that the vendors of many of these products are extremely market-conscious and therefore provide regular upgrades to incorporate improvements and extensions accomplished by benchmark competitors.

On workstations, for standard applications, the four most widely used and mature packages are BMDP, SAS, SPSS, and S-PLUS. On personal computers, DataDesk, Stata, and SYSTAT are widely available as well as versions of the workstation software. For those who plan extensive statistical programming, it is sensible to examine the software libraries IMSL and NAG, both of which offer a smorgasbord of mathematical routines useful in multivariate analysis. Finally, there are a number of commercial and noncommercial programs available that implement some of the more advanced techniques discussed previously.

BMDP, SAS, SPSS, and S-PLUS all have broadly comparable functionality and all are regularly corrected, updated, and extended. BMDP and SYSTAT have been acquired by the developer of SPSS, so analysts interested in those programs should contact SPSS, Inc. Also, there is some interplay between the major packages; SAS and BMDP enable users to call each others' routines quite flexibly. SPSS and BMDP have commands that permit users to read data that have been stored in one of the SAS formats.

The BMDP routines were originally written with biomedical and educational research applications in mind. This has led to particular emphasis on statistical techniques related to those fields, such as logistic regression, block clustering, and repeated measures designs. To get a detailed grasp of the package's capabilities, one should examine the "BMDP Statistical Software Manual" (1998).

In slight contrast, SAS has historically focused on the linear model, time series methods, graphics, and its user interface. It can link with IMSL routines, enabling mathematically sophisticated analyses. A less powerful version of SAS is available for personal computers. There are several SAS manuals that outline the available analyses in particular detail, but the core manual is the "*SAS/STAT User's Guide*" (SAS, 1999).

SPSS design was guided largely by social science applications. Originally, it concentrated upon the analysis of very large datasets, report writing, and specialized graphics, including graphic templates for the United States and Canada. The "*SPSS Base 10.0 User's Guide*" (SPSS, 1999) describes the available routines.

SYSTAT, originally developed for personal computers, includes most of the functionality available with BMDP, SAS, and SPSS. A particular strength of SYSTAT in the context of multivariate analysis is its broad capability with the multivariate general linear model, which supports analysis of very complicated experimental designs.

S-PLUS is an extension of S, a programming language developed at Bell Laboratories. It is structured so that computations treat data structures as entities, rather than lists of separate values, and it has extremely flexible graphic capabilities. Also, insofar as it is more of a language than a package, users typically develop a library of S-PLUS routines tailored to their particular applications. Venables and Ripley (1998) illustrate many multivariate analyses using S-PLUS.

DataDesk, for personal computers, implements a range of modern graphic tools for multivariate analysis. These include point-cloud rotation, which enables the user to visualize three-dimensional scatterplots of the data and methods of linking data across several graphic displays for interactive exploration. Similarly, Stata was also originally developed for personal computers, but is designed to enable easy integration with the Internet and to support database management as well as statistical analysis.

IMSL and the NAG Library of Algorithms are software libraries that enable Fortran users to call advanced routines that implement a range of mathematical and statistical functions, from sorting a vector to fitting an ARMA(p,q) model. GLIM, a specialized package for inference on the generalized linear model, is distributed by the Numerical Algorithms Group Ltd., the owners of the NAG library. CART, a program for classification and regression analysis via recursive partitioning, is marketed by Salford Systems, Inc. BUGS, for "Bayesian updating via Gibbs sampling," is noncommercial code that has revolutionized the practice of multivariate Bayesian analysis by enabling a much broader scope of practical calculation. Current details on all of the software discussed in this section may be easily found on the Internet.

SEE ALSO THE FOLLOWING ARTICLES

DATA MINING, STATISTICS • INFORMATION THEORY • PROBABILITY • STATISTICS, BAYESIAN • STATISTICS, FOUNDATIONS • STATISTICS, NON-PARAMETRIC

BIBLIOGRAPHY

Agresti, A. (1990). "Categorical Data Analysis," Wiley, New York.

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle." *In* "Second International Symposium on Information Theory" (B. N. Petrov and F. Csàki, eds.), pp. 267–281, Akademia Kiadó, Budapest.

Anderson, T. W. (1984). "An Introduction to Multivariate Statistical Analysis," 2nd ed., Wiley, New York.

Arnold, S. (1981). "The Theory of Linear Models and Multivariate Analysis," Wiley, New York.

Banks, D. L. (1989). "Bootstrapping—II." *In* "The Encyclopedia of Statistical Sciences," Suppl. vol. (N. Johnson, S. Kotz, and C. Read, eds.), pp. 17–22, Wiley, New York.

Barndorff-Neilsen, O. E. (1978). "Information and Exponential Families," Wiley, New York.

Bates, D. M., and Watts, D. G. (1988). "Nonlinear Regression Analysis and Its Applications," Wiley, New York.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). "Regression Diagnostics: Identifying Influential Data and Sources of Collinearity," Wiley, New York.

Beran, R., and Srivastava, M. S. (1985). "Bootstrap tests and confidence regions for functions of a covariance matrix," *Ann. Stat.* **13,** 95–115.

Bernardo, J. M., and Smith, A. M. F. (1994). "Bayesian Theory," Wiley, New York.

Berger, J. O. (1985). "Statistical Decision Theory and Bayesian Analysis," Springer-Verlag, New York.

Birch, M. W. (1963). "Maximum likelihood in three-way contingency tables," *J. R. Stat. Soc. B* **25,** 220–233.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). "Discrete Multivariate Analysis: Theory and Practice," MIT Press, Cambridge, MA.

Blackwell, D., and Dubins, L. (1962). "Merging of opinions with increasing information," *Ann. Math. Stat.* **33,** 882–886.

"BMDP Statistical Software Manual, Release 7." (1998). SPSS, Inc., Chicago, IL.

Box, G. E. P., and Cox, D. R. (1964). "An analysis of transformations," *J. R. Stat. Soc. B* **26,** 211–252.

Box, G. E. P., and Draper, **XX**. (1969).

Box, G. E. P., and Jenkins, G. M. (1976). "Time Series Analysis: Forecasting and Control," Holden Day, San Francisco.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). "Statistics for Experimenters," Wiley, New York.

Breiman, L., and Friedman, J. (1985). "Estimating optimal transformations for multiple regression and correlation" (with discussion). *J. Am. Stat. Assoc.* **80,** 580–619.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. (1984). "Classification and Regression Trees," Wadsworth, Belmont, CA.

Brillinger, **XX**. (1981).

Brockwell, P. J., and Davis, R. A. (1991). "Time Series: Theory and Methods," 2nd ed., Springer-Verlag, New York.

Brockwell, P. J. , and Davis, R. A. (1996). "Introduction to Time Series and Forecasting," Springer-Verlag, New York.

Broemeling, L. D. (1985). "Bayesian Analysis of Linear Models," Marcel Dekker, New York.

Brown, R. L., Durbin, J., and Evans, J. M. (1975). "Techniques for testing the constancy of regression relationships over time," *J. R. Stat. Soc. B* **37,** 149–192.

Carroll, R. J., and Ruppert, D. (1988). "Transformation and Weighting in Regression," Chapman and Hall, London.

Christensen, R. (1991). "Linear Models for Multivariate, Time Series, and Spatial Data," Springer-Verlag, New York.

Christensen, R. (1996). "Plane Answers to Complex Questions: The Theory of Linear Models," 2nd ed., Springer-Verlag, New York.
Cox, D. R., and Snell, J. (1989). "Analysis of Binary Data," 2nd ed., Chapman and Hall, New York.
Darroch, J. N., Lauritzen, S.L., and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Stat.* **8,** 522–539.
DeGroot, M. H. (1970). "Optimal Statistical Decisions," McGraw-Hill, New York.
Dillon, W., and Goldstein, M. (1978). "Discrete Discriminant Analysis," Wiley, New York.
Douglas, R., and Fienberg, S. E. (1991). "An overview of dependency models for cross-classified categorical data involving ordinal variables." *In* "Proceedings of the Conference on Positive Dependence in Statistics and Probability" (H. W. Block, A. R. Sampson, and T. H. Savits, eds.), IMS.
Efron, B. (1979). "Bootstrap methods: Another look at the jackknife," *Ann. Stat.* **7,** 1–26.
Efron, B. (1982). "The Jackknife, the Bootstrap and Other Resampling Plans," Society for Industrial and Applied Mathematics, Philadelphia.
Efron, B. (1983). "Estimating the error rate of a prediction rule: Improvements on cross-validation," *J. Am. Stat. Assoc*. **78,** 316–331.
Efron, B., and Tibshirani, R. (1986). "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Stat. Sci.* **1,** 54–75.
Ezekiel, M. (1930). "Methods of Correlation Analysis," Wiley, New York.
Fienberg, S. E. (1980). "The Analysis of Cross-Classified Categorical Data," 2nd ed., MIT Press, Cambridge, MA.
Fisher, R. A. (1915). "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, **10,** 507–521.
Fisher, R. A. (1921). "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron* **1,** 1–32.
Fisher, R. A. (1928). "The general sampling distribution of the multiple correlation coefficient," *Proc. R. Soc. Lond. A* **121,** 654–673.
Fisher, R. A. (1936). "The use of multiple measurements in taxonomic problems," *Ann. Eugen.* **7,** 179–188.
Fisher, R. A. (1969). "Statistical Methods for Research Workers," 14th ed., Oliver and Boyd, London.
Fisher , R. A. (1990).
Flury, B. (1988). "Common Principal Components and Related Multivariate Models," Wiley, New York.
Flury, B. (1997). "A First Course in Multivariate Statistics," Springer-Verlag, New York.
Friedman, J. H. (1991). "Multivariate additive regression splines" (with discussion). *Ann. Stat.* **19,** 1–141.
Friedman, J. H., and Stuetzle, W. (1981). "Projection pursuit regression," *J. Am. Stat. Assoc.* **76,** 817–823.
Galton, F. (1888). "Co-relations and their measurement, chiefly from anthropometric data," *Proc. R. Soc.* **45,** 135–140.
Graybill, F. A. (1976). "Theory and Applications of the Linear Model," Duxbury Press, North Scituate, MA.
Haberman, S. J. (1974). "The Analysis of Frequency Data," University of Chicago Press, Chicago.
Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). "Robust Statistics: The Approach Based on Influence Functions," Wiley, New York.
Hartigan, J. A. (1975). "Clustering Algorithms," Wiley, New York.
Hastie, T. J., and Tibshirani, R. J. (1990). "Generalized Additive Models," Chapman and Hall, New York.
Hawkins, D. M. (1980). "Identification of Outliers," Chapman and Hall, New York.
Hotelling, H. (1933). "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol*. **24,** 417–441.
Huynh, H., and Feldt, L. S. (1970). "Conditions under which mean square ratios in repeated measurements designs have exact F-distributions," *J. Am. Stat. Assoc.* **65,** 1582–1589.
James, W., and Stein C. (1961). "Estimation with quadratic loss." *In* "Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability" (J. Neyman, ed.), Vol. I, pp. 361–379, University of California Press, Berkeley, CA.
Jardine, N., and Sibson, R. (1971). "Mathematical Taxonomy," Wiley, New York.
Jolliffe, I. T. (1986). "Principal Component Analysis," Springer-Verlag, New York.
Jöreskog, **XX**. (1970).
Kalton, G., and Kasprzyk, D. (1986). "The treatment of missing survey data," *Survey Methodol.* **12,** 1–16.
Kauffman, L., and Rousseeuw, P. (1990). "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley, New York.
Kempthorne, O. (1952). "The Design and Analysis of Experiments," Wiley, New York.
Khuri, A. I., and Cornell, J. A. (1987). "Response Surfaces: Designs and Analyses," Marcel Dekker, New York.
Kres, H. (1983). "Statistical Tables for Multivariate Analysis: A Handbook with References to Applications," Springer-Verlag, New York.
Lachenbruch, P. A. (1975). "Discriminant Analysis," Griffin, London.
Lauritzen, S. L. (1996). "Graphical Models," Oxford University Press, New York.
Lawless, J. F. (1982). "Statistical Models and Methods for Lifetime Data," Wiley, New York.
Lehmann, E. L. (1986). "Testing Statistical Hypotheses," 2nd ed., Springer-Verlag, New York.
Lehmann, E. L., and Casella, G. (1998). "Theory of Point Estimation," 2nd ed., Springer-Verlag, New York.
Little, R. J. A., and Rubin, D. B. (1987). "Statistical Analysis with Missing Data," Wiley, New York.
MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations." *In* "Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability," pp. 281–297, University of California Press, Berkeley, CA.
Madansky, A. (1988). "Prescriptions for Working Statisticians," Springer-Verlag, New York.
Mahalanobis, P. C. (1936). "On the generalized distance in statistics," *Proc. Nat. Inst. Sci. India* **12,** 49–55.
Mallows, C. L. (1973). "Some comments on C_p," *Technometrics* **15,** 661–676.
McCullagh, P., and Nelder, J. (1989). "Generalized Linear Models," Chapman and Hall, New York.
McLachlan, G. J., and Krishnan, T. (1997). "The EM Algorithm and Extensions," Wiley, New York.
Miller, R. G. (1981). "Simultaneous Statistical Inference," 2nd ed., Springer-Verlag, New York.
Miller, R. G. (1981). "Survival Analysis," Wiley, New York.
Milligan, G. W., and Cooper, M. C. (1985). "An examination of procedures for determining the number of clusters in a dataset," *Psychometrika* **50,** 159–179.
Milliken, G. A., and Johnson, D. E. (1984). "Analysis of Messy Data," Vol. I, "Designed Experiments," Van Nostrand, New York.
Morrison, D. F. (1990). "Multivariate Statistical Methods," McGraw-Hill, New York.
Mosteller, F., and Tukey, J. W. (1968). "Data analysis, including statistics." *In* "The Handbook of Social Psychology," 2nd ed. (G. Lindzey and E. Aronson eds.), pp. 80–203, Addison-Wesley, Reading, MA.

Nelder, J. A., and Wedderburn, R. W. M. (1972). "Generalised linear models," *J. R. Stat. Soc. A* **135,** 370–384.

Pearson, K. (1900). "On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Phil. Mag.* **50,** 157–175.

Pearson, K. (1901). "On lines and planes of closest fit to systems of points in space," *Phil. Mag. Ser. 6* **2,** 559–572.

Press, S. J. (1982). "Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference," 2nd ed., Krieger, Huntington, NY.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (1998). "Applied Regression Analysis: A Research Tool," Springer-Verlag, New York.

Read, T. R. C., and Cressie, N. A. C. (1988). "Goodness-of-Fit Statistics for Discrete Multivariate Data," Springer-Verlag, New York.

Roy, S. N. (1953). "On a heuristic method of test construction and its use in multivariate analysis," *Ann. Math. Stat.* **24,** 220–238.

Santner, T. J., and Duffy, D. E. (1989). "The Statistical Analysis of Discrete Data," Springer-Verlag, New York.

SAS. (1999). "SAS/STAT User's Guide: Statistics, Version 8 Edition," SAS Institute, Inc., Cary, NC.

Scheffé, H. (1953). "A method for judging all contrasts in the analysis of variance," *Biometrika* **40,** 87–104.

Scheffé, H. (1959). "The Analysis of Variance," Wiley, New York.

Snapinn, S. M., and Knoke, J. D. (1985). "An evaluation of smoothed classification error-rate estimators," *Technometrics* **27,** 199–206.

Spearman, C. (1904). "General intelligence objectively determined and measured," *Am. J. Psychol.* **13,** 333–343.

Speed, T. P. (1987). "What is an analysis of variance?" *Ann. Stat.* **15,** 885–910.

SPSS. (1999). "SPSS Base 10.0 User's Guide," SPSS, Inc., Chicago, IL.

Stata (1999). "Stata Reference Manual," Stata Press, College Station, TX.

Stauffer, D. F., Garton, E. O., and Steinhorst, R. K. (1985). "A comparison of principal components from real and random data," *Ecology* **66,** 1693–1698.

Stein, C. (1956). "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution." *In* "Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability" (J. Neyman, ed.), Vol. I, pp. 197–206, University of California Press, Berkeley, CA.

Stigler, S. M. (1986). "The History of Statistics: The Measurement of Uncertainty before 1900," Harvard University Press, Cambridge, MA.

Stone, M. (1977). "Cross-validation: A review," *Math. Operationsforsch. Stat.* **9,** 127–139.

Tabachnick, B. G., and Fidell, L. S. (1989). "Using Multivariate Statistics," 2nd ed., Harper & Row, New York.

Thurstone, L. L. (1945). "Multiple Factor Analysis," University of Chicago Press, Chicago.

Tibshirani, R. (1988). "Estimating optimal transformations for regression via additivity and variance stabilization," *J. Am. Stat. Assoc.* **83,** 394–405.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). "Statistical Analysis of Finite Mixture Distributions," Wiley, New York.

Tong, Y. L. (1990). "The Multivariate Normal Distribution," Springer-Verlag, New York.

Venables, W. N., and Ripley, B. D. (1998). "Modern Applied Statistics with S-PLUS," 3rd ed., Springer-Verlag, New York.

Wahba, G. (1985). "A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem," *Ann. Stat.* **13,** 1378–1402.

Whittle, P. (1951). "Hypothesis Testing in Time Series Analysis," University of Uppsala, Uppsala, Sweden.

Wiener, N. (1949). "Extrapolation, Interpolation and Smoothing of Stationary Time Series," Wiley, New York.

Weisberg, S. (1980). "Applied Linear Regression," Wiley, New York.

Whittaker, J. (1990). "Graphical Models in Applied Multivariate Statistics," Wiley, New York.

Yule, G. U. (1900). "On the association of attributes in statistics: With illustration from the material of the childhood society, &c." *Phil. Trans. R. Soc. A* **194,** 257–319.



Statistics, Nonparametric

Joseph W. McKean
Western Michigan University

Simon J. Sheather
University of New South Wales

I. Location Models
II. Linear Models
III. High Breakdown Robust Estimates
IV. Optimal Rank-Based Analyses
V. Rank-Based Analyses of Experimental Designs
VI. Measures of Association

GLOSSARY

Asymptotic Relative Efficiency (ARE) A measure of the performance of one procedure relative to another. For two asymptotically unbiased estimates of a parameter the ARE is the reciprocal of the ratio of their asymptotic variances.

Breakdown point The smallest proportion of corrupt data that renders an estimator meaningless.

Diagnostics Plots and statistics which are used to determine the quality of fit.

Distribution free Statistical tests whose null distributions do not depend on the underlying distribution of the data.

HBR High breakdown R estimates of the regression coefficients of a linear model.

Influence function A measure of the sensitivity of an estimator to an additional point (outlier) to the sample.

Rank-based analyses Analyses (tests of hypotheses, confidence procedures, diagnostic analyses) of models based on the fit of the model using R or HBR estimates.

R estimates Estimates of the regression coefficients of a linear model based on minimizing a norm of residuals. This norm is defined in terms of (rank) score function.

Robust Procedures whose results are not sensitive (influenced) by outlying observations.

THE TERMS nonparametric (NP) statistics or distribution-free methods have historically referred to a collection of statistical tests whose null distributions do not depend on the underlying distribution of the data; hence the term nonparametric, i.e., no parameters. Many of the early nonparametric procedures were based on replacing the data by their ranks which, before the advent of the computer, facilitated hand calculation. Contributors to this early development include Hodges, Kendall, Krsukal, Lehmann, Mann, Pitman, Spearman, Whitney, and Wilcoxon, to name a few. It is impossible to mention all the contributors to this early development. These and many more are cited in the references of the monographs listed in the following. These procedures were thought to be quick and dirty and more easily computable than classical least squares (LS) methods. They were thought, however, to be much less efficient than LS methods. Through

the work of Lehmann and Hodges (and others) in the 1950s and 1960s, though, it was found that many of these NP procedures were quite efficient. For example, if the data follow a normal distribution then many of these procedures have efficiency .955 relative to LS methods; while, if the data follow distributions with thicker tails than a normal (allow outliers like most real data), then these NP procedures are much more efficient than LS methods. The early NP procedures were essentially test statistics, but theory for related estimates and confidence intervals of location parameters were developed in the 1950s and 1960s. Taken together these NP procedures for simple location problems offer the user highly efficient and robust methods which form an attractive alternative to traditional least squares (LS) procedures. The monographs of Hájek and Šidák (1967), Puri and Sen (1971), Lehmann (1975), and Randles and Wolfe (1979) contain many references to this work and they extended the work into various settings.

LS procedures, though, generalize easily to any linear model and to most nonlinear models. The LS procedures are not model dependent. In contrast, there are very few classical nonparametric procedures for designs other than the simple location designs and, furthermore, these procedures vary with the problem. For instance, the ranking procedure for the Kruskal Wallis test of treatment effect in a one-way layout is much different than the ranking procedure for the Friedman test of treatment effect in a two-way design. Further the efficiencies of these two procedures differ widely, although both are based on linear rankings.

In geometric terms, LS procedures for estimation and testing for any linear model are based on the Euclidean norm. Although, traditional distribution-free procedures do not generalize to any linear model, rank-based procedures based on robust estimates of regression coefficients do. These estimates are based on minimizing a non-Euclidean norm. Rank-based analyses for linear models are then formulated analogous to LS analyses except this norm is substituted for the Euclidean norm. This allows a complete inference (estimation, test of linear hypotheses, confidence procedures, and diagnostic analysis of fit) for any linear model similar to LS. While in general these rank-based procedures are not distribution free, they are asymptotically distribution free. The rank-based analysis generalizes simple location rank procedures and, further, it possesses the same efficiency properties as rank procedures for simple location models. This analysis is robust to outliers in response space and can be modified to achieve high breakdown over factor space. This work on rank-based procedures has developed over the last 25 years. Monographs include the books by Hettmansperger (1984), Puri and Sen (1985), Maritz (1995), and Koul (1992). The recent monograph by Hettmansperger and McKean (1998) develops rank-based methods for location through linear models, univariate and multivariate. Discussions of robustness can be found in Hampel *et al.* (1986), Huber (1981), and Staudte and Sheather (1990).

In this article, we discuss nonparametric procedures for location models and their extension to rank-based analyses for linear models. Hettmansperger and McKean (1998) is a main reference for the methodology. Their list of references include many of the contributors who have worked on the methods discussed in this article. A more applied reference for these procedures is Hollander and Wolfe (1999).

The statistical software package Minitab (1988) contains procedures for many of these methods including rank-based analyses of linear models. Also computation of these rank-based analyses of linear models can be performed at the web site of one of the authors (http://www.stat.wmich/mckean/rank-based) or a pc-executable module can be obtained from the authors.

I. LOCATION MODELS

In this section, we consider inference procedures for one- and two-sample problems. These procedures can be used for observational studies or experimental designs. Much of our attention will focus on location problems but we will discuss scale problems at the end of the section. We will consider two-sample problems first. In terms of an experimental design, one sample represents the control (standard or placebo) while the other represents the treatment.

A. Two-Sample Location Problems

Let X denote a measurement drawn from the first population. Assume X has distribution function F and density f. Let Y denote an observation from the second population with distribution function G and density g. The natural null hypothesis is H_0: $F(x) = G(x)$; i.e., the populations are the same. Let $X_1, \ldots, X_{n_1}$ be a random sample from the first population and let $Y_1, \ldots, Y_{n_2}$ be a random sample, independent of the first, from the second population. Let $n = n_1 + n_2$ denote the size of the combined samples.

In terms of an experimental design, note that the **completely randomized design** (CRD) results in a two-sample problem. Here, n experimental units are selected at random and n_1 are randomly assigned to the control (Population I) and n_2 assigned to the treatment (Population II). The CRD, at least at the onset of the study, assures the independence within and between the sample items.

While H_0 is a natural null hypothesis, there are several interesting alternative hypotheses. One is stochastic

dominance in which Y stochastically dominates X, i.e., $G(x) \le F(x)$, with strict inequality for at least one x. For this article, unless otherwise noted, we will consider a **location model**. In this case, $G(x) = F(x - \Delta)$ for some parameter Δ. Hence, Δ is the shift in locations between the populations; for example, it is the difference in means ($\Delta = \mu_2 - \mu_1$, provided the means exist) or the difference in the population medians. In experimental design terminology, Δ is the effect between the populations. For the location model, the null hypothesis is represented by $\Delta = 0$, whereas the alternatives are usually the one- or two-sided alternatives given by

$$H_0\colon \Delta = 0 \text{ versus one of } \begin{cases} H_1\colon & \Delta > 0 \\ H_2\colon & \Delta < 0\,. \\ H_3\colon & \Delta \neq 0 \end{cases} \tag{1}$$

1. Tests

For convenience, suppose we want to test H_0 versus H_1. The **Mann–Whitney–Wilcoxon** (MWW) test procedure is to combine the two samples and then rank the items from 1 (the smallest item) to n (the largest item). Letting $R(Y_i)$ denote rank of Y_i, the test statistic is

$$T = \sum_{j=1}^{n_2} R(Y_j), \tag{2}$$

and we reject H_0 in favor of H_1 if $T \ge c$ is too large. To determine c, we need the null distribution of T. Under H_0, the Ys and Xs are drawn from the same population. Hence, any subset of n_2 elements from the set $\{1, \ldots, n\}$ are equilikely to be the chosen for the ranks of the Ys, i.e., the probability that any subset is selected is $\binom{n}{n_2}^{-1}$. Thus the null distribution of T does not depend on F. We say that T is a **distribution-free** test statistic. Tables for null distribution of T are available (see Hollander and Wolfe, 1999).

The null distribution of T is symmetric with range $\{n_2(n_2+1)/2, \ldots, n_2(n+n_1+1)/2\}$, and mean and variance

$$E_0[T] = \frac{n_2(n+1)}{2} \quad \text{and} \quad V_0[T] = \frac{n_1 n_2 (n+1)}{12}. \tag{3}$$

It also follows that the distribution of T is asymptotically normal. An asymptotically level α test is to reject H_0 in favor of H_1 if $z_T > z_\alpha$, where

$$z_T = \frac{T - E_0[T]}{\sqrt{V_0[T]}}, \tag{4}$$

$z_\alpha = \Phi^{-1}(1-\alpha)$, and Φ is the distribution function of a standard normal random variable. The asymptotic distribution provides a reasonable approximation for sample sizes as low as $n = 10$. We would, however, recommend the usual continuity correction. The asymptotic p-value using this correction is given by

$$p = 1 - \Phi\left(\frac{T - (1/2) - E_0[T]}{\sqrt{V_0[T]}}\right). \tag{5}$$

Similarly, asymptotic level α tests for H_2 or H_3 would be to reject H_0 if $z_T < -z_\alpha$ or $|z_T| > z_{\alpha/2}$, respectively.

In real data observations are often tied. The usual method for handling ties is to use the average of the ranks that would have been assigned. For example, suppose four observations all had the same value and that they would have been assigned the ranks, say, 20–23. Then each of them would be assigned the average rank, 21.5. Ties do affect the null distribution of T. Unless the number of ties is considerable, the modification is slight. For severe cases of ties, there are corrections to the variance of T available (see Hollander and Wolfe, 1999).

2. Estimation

The associated estimate of the effect Δ is obtained by **inverting** the test statistic T. Let $T(\Delta)$ be defined by

$$T(\Delta) = \sum_{j=1}^{n_2} R(Y_j - \Delta), \tag{6}$$

where $R(Y_j - \Delta)$ denotes the rank of $Y_j - \Delta$ among $X_1, \ldots, X_{n_1}$ and $Y_1 - \Delta, \ldots, Y_{n_2} - \Delta$. The associated estimate of the effect Δ, is the estimator $\hat{\Delta}$ which solves this equation

$$T(\hat{\Delta}) = E_0[T] = n_2 \frac{n+1}{2}; \tag{7}$$

hence, $\hat{\Delta}$ is the most "acceptable" null hypothesis based on the data. To obtain this point estimate, note that the rank of $Y_j - \Delta$ can be expressed as, $R(Y_j) = \#_i\{X_i < Y_j - \Delta\} + \#_i\{Y_i \le Y_j\}$. Next, define $W(\Delta) = \#_{ij}\{Y_j - X_i > \Delta\}$. By summing over j, we get the identity

$$T(\Delta) = W(\Delta) + \frac{n_2(n_2+1)}{2}. \tag{8}$$

After algebraic simplification, $\hat{\Delta}$ solves the equation,

$$W(\hat{\Delta}) = \frac{n_1 n_2}{2}. \tag{9}$$

Hence $\hat{\Delta}$ exceeds and is exceeded by half of the differences $Y_j - X_i$. Thus the associated point estimator of Δ is the **median of the differences**, i.e.,

$$\hat{\Delta} = \text{med}\{Y_j - X_i\}. \tag{10}$$

In place of the test statistic T, the equivalent test statistic $W(0)$ is often used. This version is called the **Mann–Whitney** version.

The asymptotic distribution of the estimator $\hat{\Delta}$ is

$$\hat{\Delta} \text{ has an approximate } N\left(\Delta, \tau^2\left\{n_1^{-1}+n_2^{-1}\right\}\right) \text{ distribution,} \tag{11}$$

where τ is the scale parameter given by

$$\tau = \left(\sqrt{12}\int f^2(t)\,dt\right)^{-1}. \tag{12}$$

This is the same asymptotic distribution of the **least squares** (LS) estimator $\bar{Y}-\bar{X}$, except that the scale parameter τ replaces the LS scale parameter σ, the standard deviation of the density f.

3. Confidence Intervals

Besides a distribution-free test statistic, the Wilcoxon procedure produces a distribution-free **confidence interval**. Consider the set of n_1n_2 differences $\{Y_j - X_i\}$. Denote the ordered differences by $D_{(1)} \le \cdots \le D_{(n_1n_2)}$. Let α be given and choose $c_{\alpha/2}$ to be the lower $\alpha/2$ critical point of the MWW distribution. Then the interval $(D_{(c_{\alpha/2}+1)}, D_{(n_1n_2-c_{\alpha/2})})$ is a $(1-\alpha)100\%$ confidence interval for Δ. Based on the asymptotic null distribution theory for the MWW test statistic, $c_{\alpha/2}$ is approximately

$$c_{\alpha/2} \doteq \frac{n_1n_2}{2} - z_{\alpha/2}\sqrt{\frac{n_1n_2(n+1)}{12}} - .5. \tag{13}$$

Based on the asymptotic distribution of the estimate $\hat{\Delta}$, the interval $\hat{\Delta} \pm t_{\alpha/2,n-2}\hat{\tau}\sqrt{n_1^{-1}+n_2^{-1}}$ is an approximate $(1-\alpha)\,100\%$ symmetric confidence interval for Δ. It requires an estimate of τ which is discussed in Section II.B.

4. Efficiency

We define the **asymptotic relative efficiency** (ARE) between two asymptotically unbiased estimators to be the reciprocal of the ratio of their asymptotic variances. Hence, the ARE between the Wilcoxon and LS estimators is

$$e(\text{Wil, LS}) = \frac{\sigma^2}{\tau^2} = 12\sigma^2\left(\int f^2(t)\,dt\right)^2. \tag{14}$$

This varies with the underlying distribution. At the normal distribution, the ARE is .955; that is, the Wilcoxon procedure is 95% as efficient as the LS procedure if the true distribution is normal. If the true distribution has tails heavier than the normal then this efficiency is usually much larger than 1. An example is the contaminated normal distribution. Suppose $(1-\epsilon)100\%$ of the time the true distribution, F, is normal with standard deviation σ and $\epsilon 100\%$ of the time it is normal with standard deviation 3σ; that is, there is an $\epsilon 100\%$ probability of obtaining an outlier. For $\sigma = 3$, Table I displays the efficiencies between Wilcoxon and LS estimators for different values of ϵ. Even at 1% contamination, the Wilcoxon is more efficient.

TABLE I Efficiencies of the Wilcoxon and LS Methods for the Contaminated Normal Distribution

ϵ	.00	.01	.03	.05	.10	.15
e(Wil,LS)	.955	1.009	1.108	1.196	1.373	1.497

5. Sample Size Determination

A local approximation to the power function of the MWW test statistic can be determined similar to the asymptotic theory for the estimate. This can be used for **sample size determination** for a CRD. Consider the MWW test for the one-sided hypothesis (replace α by $\alpha/2$ for a two-sided alternative hypothesis). Suppose the level, α, and the power, γ, for a particular alternative Δ_A are specified. For convenience, assume equal sample sizes, i.e., $n_1 = n_2 = n^*$, where n^* denotes the common sample size. Under these conditions, the suggested sample size is given by

$$n^* = \left(\frac{z_\alpha - z_\gamma}{\Delta_A}\right)^2 2\tau^2. \tag{15}$$

Note that it does depend on τ which, in applications, would have to be guessed or estimated in a pilot study. For a specified distribution it can be evaluated; for instance, if the underlying density is assumed to be normal with standard deviation σ then $\tau = \sqrt{\pi/3}\sigma$. For LS, the formula for n^* would be the same, except τ would be replaced by σ. The reciprocal of the ratio of the sample sizes is the efficiency $e(\hat{\Delta}, \bar{Y}-\bar{X})$.

6. Example: Cholesterol Data

The data for this problem are drawn from a high-volume drug screen designed to find compounds which reduce low-density lipoproteins, LDL, cholesterol in quail. [See Hettmansperger and McKean (1998) for a discussion of this data set.] Here we consider the results for one drug compound. For the experiment, 30 quail were selected and 10 (Treated Group) of them (by random selection) were fed a special diet that was mixed with a drug compound while the remaining 20 (Control Group) were fed the same diet but without the drug compound. After a specified period of time, the LDL levels of the quail were measured. The data are displayed in Table II.

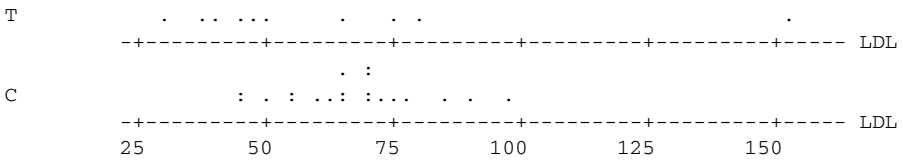
TABLE II Quail LDL Levels

Control	64	49	54	64	97	66	76	44	71	89
	70	72	71	55	60	62	46	77	86	71
Treated	40	31	50	48	152	44	74	38	81	64

Let θ_C and θ_T denote the true median levels of LDL for the control and treatment populations, respectively. The parameter of interest is $\Delta = \theta_C - \theta_T$. We are interested in the alternative hypothesis that the treatment has been effective; hence the hypotheses are

$$H_0: \Delta = 0 \text{ versus } H_A: \Delta > 0.$$

Comparison dotplots for the data are:



Note that there is one outlier, the fifth observation of the treated group, which has the value 152. For the data at hand, the treated group appears to have lower LDL levels.

The sum of the ranks of the control subjects is $T = 344.5$ (2), which leads to a standardized test value of $z_T = 1.52$ with an observed significance level of .064. The MWW indicates with marginal significance that the treatment performed better than the placebo. For comparison, the two-sample t statistic has the value $t = .56$ with a p-value of .29 which would result in the acceptance of H_0. The two-sample t analysis was impaired by the outlier.

The estimate of Δ (10) is $\hat{\Delta} = 14$ and a 90% confidence interval for Δ (13) is $(-2.01, 24.00)$. In contrast, the least squares estimate of shift is 5 and the corresponding 90% confidence interval is $(-10.25, 20.25)$.

B. Paired Designs

While the CRD is an often used design, it is usually not as efficient as the **randomized paired design** (RPD). In this design, an experimental unit is a pair or a block of length 2. For example, suppose we are testing the strength of a metal under two intensities of stress, T1 and T2. In a RPD, the piece of metal is divided in half and by random selection one of the halves is subjected to the stress level T1 while the second half is subjected to the stress level T2. At the end of the time period, let X and Y denote the strengths of the pieces under stress levels T1 and T2, respectively. Thus a pair of observations (X, Y) is obtained. In general, the data for a RPD experiment over n pairs (or blocks) consist of the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, where X_i and Y_i denote the responses under Treatments 1 and 2, respectively, for the ith block. The observations X_i and Y_i are not independent so the two-sample analysis of the last section is not appropriate. The sample of interest, though, consists of the differences $D_1 = Y_1 - X_1, \ldots, D_n = Y_n - X_n$. Note that the target parameter remains the same, i.e., $E(D_i) = E(Y_i) - E(X_i) = \Delta$. Hence our model is

$$D_i = \Delta + e_i, \tag{16}$$

where the errors are **independent and identically distributed** (iid) with common density $h(t)$. The hypotheses of interest are given by (1).

1. Tests

For the RPD, the treatments are randomly assigned within a block. Then under the null hypothesis of no treatment effect the differences D_i are symmetrically distributed about 0 and, hence, the observations $-D_i$ and D_i should receive the same rank. These are the rankings used in the **signed-rank Wilcoxon** (**SRW**) statistic. The test statistic is given by

$$T_S = \sum_{i=1}^{n} R|D_i| \operatorname{sgn}(D_i), \tag{17}$$

where the $\operatorname{sgn}(t)$ is $1, 0$, or -1 depending on whether $t > 0$, $t = 0$ or $t < 0$, respectively. The decision is to reject $H_0: \Delta = 0$ in favor of $H_1: \Delta > 0$ if T_S is too large. Under H_0, the ranks and signs are independent of one another and the $R|D_i|$ is equilikely to be any value 1 to n. Therefore, as with the MWW test statistic, the distribution of T_S under the null hypothesis does not depend on the distribution of errors. So T_S is distribution-free under H_0. Its null distribution is symmetric with range $\{-n(n+1)/2, \ldots, n(n+1)/2\}$ and mean and variance

$$E_0[T_S] = 0 \quad \text{and} \quad V_0[T_S] = \frac{n(n+1)(2n+1)}{24}. \tag{18}$$

Tables of the null distribution can be found in Hollander and Wolfe (1999). The distribution of T_S is asymptotically normal. An asymptotically level α test is to reject H_0 if $z > z_\alpha$, where

$$z = \frac{T_S - E_0[T_S]}{\sqrt{V_0[T_S]}}. \tag{19}$$

The asymptotic distribution provides a reasonable approximation for sample sizes as low as $n = 10$. We would, however, recommend the usual continuity correction. For testing $H_0: \Delta = 0$ versus $H_1: \Delta > 0$, the asymptotic p-value using this correction is given by

$$p = 1 - \Phi\left(\frac{T_S - (1/2)}{\sqrt{V_0[T_S]}}\right). \quad (20)$$

2. Estimation

Following Section I.A, the estimate of Δ solves the equation $T_S(\Delta) = 0$, where $T_S(\Delta) = \sum_{i=1}^{n} R|D_i - \Delta| \operatorname{sgn}(D_i - \Delta)$. We can obtain a closed form solution by using the identity $T_S(\Delta) = 2U(\Delta) - (n(n+1)/2)$, where

$$U(\Delta) = \#_{i \le j}\left\{\frac{D_i + D_j}{2} > \Delta\right\}. \quad (21)$$

Thus the estimate solves the equation $U(\Delta) = (n(n+1)/4)$. Because there are $n(n+1)/2$ pairwise averages, the solution is

$$\hat{\Delta} = \underset{i \le j}{\text{med}}\left\{\frac{D_i + D_j}{2}\right\}. \quad (22)$$

This is often called the **Hodges–Lehmann** estimate.

The asymptotic distribution of the estimator $\hat{\Delta}$ is

$$\hat{\Delta} \text{ has an approximate } N(\Delta, \tau^2 n^{-1}) \text{ distribution,} \quad (23)$$

where τ is the scale parameter given by (12).

3. Confidence Intervals

Besides a distribution-free test statistic, the SRW procedure produces a distribution-free **confidence interval**. Consider the set of $M = n(n+1)/2$ pairwise averages $\{(D_i + D_j)/2\}$. Denote these ordered averages by $V_{(1)} \le \cdots \le V_{(M)}$. Let α be given and choose $c_{\alpha/2}$ to be the lower $\alpha/2$ critical point of the SRW distribution. Then the interval $(V_{(c_{\alpha/2}+1)}, V_{(M-c_{\alpha/2})})$ is a $(1-\alpha)100\%$ confidence interval for Δ. Based on the asymptotic null distribution theory for the signed-rank test statistic, $c_{\alpha/2}$ is approximately

$$c_{\alpha/2} \doteq \frac{n(n+1)}{4} - z_{\alpha/2}\sqrt{\frac{n(n+1)(2n+1)}{24}} - .5. \quad (24)$$

The interval $\hat{\Delta} \pm t_{\alpha/2,n-2}\hat{\tau}/\sqrt{n}$ is an approximate $(1-\alpha)100\%$ symmetric confidence interval for Δ.

4. Efficiency

The ARE between the SRW and LS procedures is the same as that between the MWW and the LS procedures in the two-sample location problem. The formula is given by (14) except that $f(t)$ is replaced by $h(t)$ the density of the errors in the paired model (16) and σ^2 is the variance of this density. Hence, if the errors are normally distributed then SRW procedures are 95% as efficient as LS procedures.

5. Sample Size Determination

Consider the signed-rank Wilcoxon test for the one-sided hypothesis (replace α by $\alpha/2$ for a two-sided alternative hypothesis). Suppose the level, α, and the power, γ, for a particular alternative Δ_A are specified. Let n denote the number of pairs or blocks to be selected. Under these conditions, the recommended number of pairs is given by

$$n = \left(\frac{z_\alpha - z_\gamma}{\Delta_A}\right)^2 \tau^2. \quad (25)$$

Note that it does depend on τ which, in applications, would have to be guessed or estimated in a pilot study. As in the two-sample location problem, if the underlying density of the errors is assumed to be normal with standard deviation σ then $\tau = \sqrt{\pi/3}\sigma$. For LS, the formula for n would be the same, except τ would be replaced by σ.

6. Example: Darwin Data

The data in Table III are measurements recorded by Charles Darwin in 1878. They consist of 15 pairs of heights in inches of cross-fertilized plants and self-fertilized plants (*Zea mays*), each pair grown in the same pot.

Let D_i denote the difference between the heights of the cross-fertilized and self-fertilized plants of the ith pot and let Δ denote the median of the distribution of D_i. Suppose we are interested in testing for an effect; that is, the hypotheses are $H_0: \Delta = 0$ versus $H_A: \Delta \neq 0$. The value of the signed-rank Wilcoxon statistic, (17), for these data is $T_S = 36$ with the approximate p value of .044.

TABLE III Heights of Plants

Pot	1	2	3	4	5	6	7	8
Cross-	23.500	12.000	21.000	22.000	19.125	21.500	22.125	20.375
Self-	17.375	20.375	20.000	20.000	18.375	18.625	18.625	15.250
Pot	9	10	11	12	13	14	15	
Cross-	18.250	21.625	23.250	21.000	22.125	23.000	12.000	
Self-	16.500	18.000	16.250	18.000	12.750	15.500	18.000	

The corresponding estimate of Δ is 3.13 in. and the 95% confidence interval is (.50, 5.21). The paired t-test statistic has the value of 2.15 with p value .050. The difference in sample means is 2.62 in. and the corresponding 95% confidence interval is (0, 5.23). The difference between the analyses is due to the possible outlying differences in Pots 2 and 15.

7. Paired Designs Where Pairing Is Not Randomized

There are of course designs where the pairing not randomized. An example is before and after designs. For such designs, the differences are still modeled as in (16) except that the errors are not necessarily symmetrically distributed under the null hypothesis. If symmetry is violated then the SRW test may not be distribution free under H_0. As with any analysis, we would recommend diagnostic checks such as q–q plots and boxplots to examine violations of symmetry. The following procedure is distribution free regardless of symmetry but it is not as efficient as the SRW procedure.

Consider the model (16) for the paired differences where the errors e_i are iid with common density $h(t)$ which has median 0 but which is not necessarily symmetric. Thus Δ is the population median of the D_is. The hypotheses of interest remain the same; for instance, the one-sided hypothesis that we have been discussing is still given by H_0: $\Delta = 0$ versus H_1: $\Delta > 0$. The **sign** test statistic is given by

$$S = \#_i\{D_i > 0\}. \tag{26}$$

Because the D_is are independent and identically distributed (under both H_0 and H_A), S has a **binomial** distribution with n trials and probability of success, $p = P[D_i > 0]$. We will denote this binomial distribution by $b(n, p)$. Under the null hypothesis, the median of the differences is 0; hence, S has a $b(n, 1/2)$ distribution. Thus S is distribution free under H_0. Under alternatives, though, p, and hence the distribution of S, depend on the distribution function of the errors F. Tables of the $b(n, 1/2)$ distribution are available (see Hollander and Wolfe, 1999).

The null mean and variance of S are given by

$$E_0[S] = \tfrac{n}{2} \quad \text{and} \quad V_0[S] = \tfrac{n}{4}. \tag{27}$$

The distribution of S is asymptotically normal. The large sample test rejects H_0 in favor of H_1 if $z \geq z_\alpha$, where $z = (S - E_0[S])/\sqrt{V_0[S]}$. We do caution the reader somewhat on this large sample approximation because the sign test statistic is much more discrete than the signed-rank Wilcoxon test statistic. The range of the sign test statistic consists of $n + 1$ values while the range of the SRW statistic consists of $n(n + 1) + 1$ values. Asymptotic normality sets in much faster for the SRW statistic.

The estimate of Δ corresponding to the sign test is the **median** of the differences, i.e., $\tilde{D} = \text{med}_i\, D_i$. Its asymptotic distribution is given by

$$\tilde{D} \text{ has an approximate } N\left(\Delta, \tau_S^2 n_1^{-1}\right) \text{ distribution,} \tag{28}$$

where τ_S is the scale parameter given by

$$\tau_S = (2h(0))^{-1}. \tag{29}$$

Estimates of τ_S are discussed in Section II.F.

The corresponding distribution-free confidence interval for Δ is the interval $(D_{(c_{\alpha/2}+1)}, D_{(n-c_{\alpha/2})})$, where $c_{\alpha/2}$ is the lower $\alpha/2$ critical point of the $b(n, 1/2)$ distribution. Based on asymptotic distribution of the sign test, $c_{\alpha/2}$ can be approximated as

$$c_{\alpha/2} \doteq \frac{n}{2} - z_{\alpha/2}\sqrt{\frac{n}{4}} - .5. \tag{30}$$

The ARE between sign test procedures and LS procedures is given by $e(\tilde{D}, \bar{D}) = 4h^2(0)\sigma^2$. If $h(t)$ is a normal density then $e(\tilde{D}, \bar{D}) = .63$ which is much lower than the ARE of the Wilcoxon procedures. On the other hand, if the error density $h(t)$ has very thick tails then the sign procedures can be considerably more efficient then the LS procedures. For comparisons Table IV presents the AREs of between the sign and LS procedures and between the Wilcoxon and LS procedures over the same family of contaminated normals used in Table I.

For sample size determination, simply replace τ in expression (25) by τ_S.

8. Example: Darwin Data, Continued

To illustrate the sign test, we use the Darwin data disscussed earlier. Out of the 15 differences, 13 are positive; hence, the sign test has an approximate p value of .0074. The corresponding estimate of θ is 3 in. and the approximate 95% confidence interval is (1.000, 6.125).

C. One-Sample Location Models

Let $X_1, \ldots, X_n$ be a random sample from a population of interest. Consider the location model,

TABLE IV Efficiencies of the Wilcoxon and LS Methods for the Contaminated Normal Distribution

ϵ	.00	.01	.03	.05	.10	.15
e(Wil, LS)	0.955	1.009	1.108	1.196	1.373	1.497
e(Sign, LS)	0.637	0.678	0.758	0.833	1.000	1.134

$$X_i = \theta + e_i; \qquad 1 \le i \le n, \tag{31}$$

where the random errors e_i are iid with density $h(t)$ which has median 0. Hence, θ is the median of X_i. Of course, we have already discussed this problem in the last section. The differences D_i form a single sample. Therefore the inference concerning Δ of the last section forms a complete inference (testing, estimation, and confidence intervals) for the parameter θ of this section. In particular, if we further assume that the density $h(t)$ is symmetric about 0 then the inference based on the SRW procedures can be used for θ. If this assumption cannot be made then inference based on the sign test procedures can be employed.

D. Robustness Properties

The robustness concepts, influence and breakdown, will be useful in the sequel and are easiest to describe in the one-sample setting. Let $\mathbf{X} = (X_1, \ldots, X_n)'$ be a sample from a population which has density $h(t)$ and an unknown parameter θ. Let $\hat{\theta}_n$ be an estimator of θ. The influence function of $\hat{\theta}_n$ measures the local sensitivity of $\hat{\theta}_n$ to an outlier while the breakdown point of $\hat{\theta}_n$ measures global senstivity to outliers. We first proceed with the influence function.

Let X^* be an additional point, an outlier, to the sample. Let $\hat{\theta}_{n+1}$ be the estimate based on the combined sample $(\mathbf{X}', X^*)'$. Then the rate of change of the estimator due to this outlier is

$$\frac{\hat{\theta}_{n+1} - \hat{\theta}_n}{1/(n+1)} \approx \Lambda(X^*). \tag{32}$$

The expression on the right is called the **sensitivity curve** of the estimator $\hat{\theta}_n$. It is a rate of change, a type of derivative. Hampel made this precise at the population model and called it the **influence function** of $\hat{\theta}_n$, which we will label as $\Lambda(X^*; \hat{\theta}_n)$. By (32), we have approximately that

$$\hat{\theta}_{n+1} \approx \hat{\theta}_n + \frac{1}{n+1}\Lambda\big(X^*; \hat{\theta}_n\big). \tag{33}$$

Thus, the influence function measures local sensitivity to an outlier. Assuming the estimate on a "good" sample is unbiased, then $\Lambda(X^*: \hat{\theta}_n)$ is a measure of the bias due to the outlier. Clearly, we want the influence function to be bounded; otherwise, the biasness can grow without bound as the outlier increases. We say an estimator is **bias robust** if its influence function is bounded.

It is easy to see that the influence function for the sample mean is $\Lambda(X^*; \bar{X}) = X^*$; hence, the bias of the mean is directly proportional to the outlier. The sample mean is not robust. On the other hand, the influence function of the sample median, $\tilde{X}$ is given by

$$\Lambda(X^*; \tilde{X}) = \frac{\operatorname{sgn}(X^*)}{2h(0)},$$

which is bounded by $(2h(0))^{-1}$. Hence the sample median is robust. Next consider the Hodges–Lehmann estimate, $\hat{\theta}_{HL} = \operatorname{med}_{i \le j} \{(X_i + X_j)/2\}$. Let $H(t)$ denote the distribution function corresponding to the density $h(t)$. The influence function of $\hat{\theta}_{HL}$ is given by

$$\Lambda(X^*; \hat{\theta}_{HL}) = \sqrt{12}\tau(H(X^*) - 1/2),$$

which is bounded by $\sqrt{12}\tau$ and, hence, the Wilcoxon estimate is also robust.

Similar results occur in the two-sample problem. The Wilcoxon estimate (median of differences) has bounded influence while the LS estimate (difference in the sample means) is not robust.

Next consider breakdown. As cited here, consider the sample $\mathbf{X} = (X_1, \ldots, X_n)'$ and an estimator $\hat{\theta}$. Suppose we corrupt m of the data points. Write the corrupted sample as $\mathbf{X}^{(m)} = (X_1^*, \ldots, X_m^*, X_{m+1}, \ldots, X_n)'$. Now think of the corrupt points as getting arbitrarily large. The smallest proportion of corrupt data points that results in the complete breakdown of $\hat{\theta}$ (i.e., $|\hat{\theta}|$ gets arbitrarily large) is the **finite sample breakdown** of $\hat{\theta}$. If this proportion has a limit as n gets large, we call it the **breakdown point** of $\hat{\theta}$. For example, consider the sample mean. If we corrupt one data point and drive it to infinity then $\bar{X}$ goes to infinity; hence, the finite sample breakdown of $\bar{X}$ is $1/n$ and thus its breakdown point is 0. On the other hand, we would have to corrupt half the data to breakdown the sample median, $\tilde{X}$. So the sample median has 50% breakdown. This is the largest value for a breakdown point. The Hodges–Lehmann estimate, $\hat{\theta}_{HL}$, has a breakdown point of 29%.

This discussion on breakdown extends to the two-sample problem also. In this case, though, the issue of the two sample sizes must be addressed. If the sample sizes are the same, $n_1 = n_2$ then the maximal breakdown point of an estimator is .25. The difference in sample means has breakdown 0, while the median of the differences has the maximal breakdown value of .25.

E. Two-Sample Scale Problem

In the two-sample problem, suppose the distribution functions of X and Y are given by $F_X(x) = F(x - \theta_X)$ and $F_Y(y) = F((y - \theta_Y)/\eta)$, respectively. The null hypothesis of interest is that the scales are the same, i.e.,

$$H_0: \eta = 1 \text{ versus } H_3: \eta \ne 1, \tag{34}$$

or it could be an appropriate one-sided alternative. Under the null hypothesis, $\eta = 1$, the distributions of $Y - \theta_Y$ and $X - \theta_X$ are the same. The only truly distribution-free tests for this problem are under the assumption that the

locations of X and Y are the same. In practice, though, this is an unsavory assumption. Often this hypothesis is tested before the test on locations. We instead present an asymptotically distribution-free procedure which has been shown to perform well in a Monte Carlo study over a large variety of practical situations (see Conover *et al.*, 1981). There are problems with traditional statistical tests in this area. The traditional (LS) two-sample F test for scales is only truly valid if X and Y have normal distributions. In the Monte Carlo study cited, the traditional F test had very poor performance in most situations. Its empirical α levels were much more liberal than the nominal levels. [It does not posssess robustness of validity; see also, Arnold (1981), Box (1953), and Hettmansperger and McKean (1998).]

Our strategy is to align the observations first and then employ the rank statistics on these aligned samples. Let $X_i^* = X_i - \hat{\theta}_X$ and $Y_j^* = Y_j - \hat{\theta}_Y$ denote the aligned observations, where $\hat{\theta}_X$ and $\hat{\theta}_Y$ are the sample medians of the X and Y samples, respectively. Next fold the aligned samples; i.e., obtain their absolute values. Combine these folded–aligned samples and rank them from low to high. Let $R(|Y_j^*|)$'s denote the ranks of folded–aligned Y samples. The **Fligner–Killeen** test statistic is given by

$$S_{FK}^* = \sum_{j=1}^{n_2} \left(\Phi^{-1}\left(\frac{R|Y_j^*|}{2(n+1)} + \frac{1}{2} \right) \right)^2. \tag{35}$$

The statistic S_{FK}^* is not distribution free for finite samples. If we further assume, however, that the distributions of X and Y are symmetric, then the test statistic S_{FK}^* is asymptotically distribution free. The null mean μ_{FK} and null variance σ_{FK}^2 of the statistic are given by

$$\mu_{FK} = n_2\bar{a} \quad \text{and} \quad \sigma_{FK}^2 = \frac{n_1 n_2}{n(n-1)} \sum (a(i) - \bar{a})^2, \tag{36}$$

where $a(i) = (\Phi^{-1}((i/(2(n+1))) + (1/2))^2$. The asymptotic version of this test statistic rejects H_0 in favor of H_3 at the approximate level α if $|z_{FK}| \geq z_{\alpha/2}$, where

$$z_{FK} = \frac{S_{FK}^* - \mu_{FK}}{\sigma_{FK}}. \tag{37}$$

II. LINEAR MODELS

Suppose the responses $Y_1, \ldots, Y_n$ follow a linear model of the form

$$Y_i = \alpha + \mathbf{x}_i\boldsymbol{\beta} + e_i; \quad 1 \leq i \leq n, \tag{38}$$

where $\mathbf{x}_i$ is a $p \times 1$ vector of constants, the random errors e_i are iid with distribution function $F(t)$ and density $f(t)$, and α and the $p \times 1$ vector $\boldsymbol{\beta}$ are unknown parameters. The parameter α is the intercept parameter and the β_js are the slope parameters. We will often treat the intercept and slope parameters separately because we want to allow for skewed error distributions. (See, for example, Section IV on survival analysis.) Further, because we are including an intercept there is no loss in generality in assumimg that the predictors are centered, i.e., $\sum_{i=1}^n x_{ij} = 0$ for all j. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ be the vector of responses, let $\mathbf{X}_c$ be the $n \times p$ matrix of regression constants x_{ij} (subscript c denotes centered), and let $\mathbf{1}$ denote an $n \times 1$ vector of ones. Then we can write this model equivalently as

$$\mathbf{Y} = \alpha\mathbf{1} + \mathbf{X}_c\boldsymbol{\beta} + \mathbf{e}, \tag{39}$$

where $\mathbf{e}$ is the $n \times 1$ vector of errors. Let Ω denote the p-dimensional column space of $\mathbf{X}_c$. Then a third way to write the model is

$$\mathbf{Y} = \alpha\mathbf{1} + \boldsymbol{\eta} + \mathbf{e}; \boldsymbol{\eta} \in \Omega. \tag{40}$$

In this last formulation, estimation is easily described. We need to find a vector $\hat{\boldsymbol{\eta}} \in \Omega$ which lies "closest" to $\mathbf{Y}$, where "closest" is defined in terms of a distance function. In the case of LS, the Euclidean distance function is used, whereas for the Wilcoxon the distance function defined in the following is used.

We present a complete robust analysis of the linear model (39) including estimation, tests of linear hypotheses, confidence procedures, and diagnostics to check quality of fit. The estimation is based on the work of Jaeckel (1972) and the analysis is based on the work of McKean and Hettmansperger (1976, 1978). [See Hettmansperger and McKean (1998) for a complete discussion.] These are rank procedures based on a fit of a linear model and we will call it a **rank-based analysis**. If Wilcoxon scores are used we will label it as the **Wilcoxon analysis**.

A. Wilcoxon Regression Estimates

First we will briefly review the LS estimates. Since we have an intercept parameter, we can obtain the LS estimates in two parts. The estimate of $\boldsymbol{\eta}$ is given by

$$\hat{\boldsymbol{\eta}}_{LS} = \underset{\boldsymbol{\eta} \in \Omega}{\text{Argmin}} \|\mathbf{Y} - \boldsymbol{\eta}\|_{LS}, \tag{41}$$

where the norm is defined by

$$\|\mathbf{v}\|_{LS} = \sqrt{\sum_{i,j} |v_i - v_j|^2}; \quad \mathbf{v} \in R^n. \tag{42}$$

Argmin is an abreviation for the value of the argument $\boldsymbol{\eta}$ which minimizes the distance between $\mathbf{Y}$ and the subspace Ω. Using matrix algebra we get, $\hat{\boldsymbol{\eta}}_{LS} = \mathbf{H}_c\mathbf{Y}$, where $\mathbf{H}_c$ is the projection matrix $\mathbf{H}_c = \mathbf{X}_c(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'$. The LS estimate of $\boldsymbol{\beta}$ solves the equation

$$\mathbf{X}_c\boldsymbol{\beta} = \hat{\boldsymbol{\eta}}_{LS}, \tag{43}$$

i.e., $\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{Y}$. The LS estimate of the intercept is the average of the LS residuals

$$\hat{\alpha}_{LS} = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\eta}_{LS,i}). \tag{44}$$

For the Wilcoxon estimates, we simply change the norm from the Euclidean (42) norm to the norm,

$$\|\mathbf{v}\|_W = \kappa\sum_{i,j}|v_i - v_j|, \quad \mathbf{v} \in R^n, \tag{45}$$

where $\kappa = 2(n+1)/\sqrt{3}$. The constant κ is irrelevant for the minimization but it does simplify later expressions. Then the **Wilcoxon regression** estimate of $\boldsymbol{\eta}$ is

$$\hat{\boldsymbol{\eta}}_W = \underset{\boldsymbol{\eta}\in\Omega}{\text{Argmin}}\|\mathbf{Y} - \boldsymbol{\eta}\|_W. \tag{46}$$

The estimate cannot be obtained in closed form, but since it is defined in terms of a norm, its computation requires the minimization of a convex function for which there are several good algorithms available; for example, Minitab (1988) or at the web site of one of the authors (http://www.stat.wmich/mckean/rank-based) or a pc-executable module can be obtained from the authors.

Once $\hat{\boldsymbol{\eta}}_W$ is obtained the Wilcoxon estimate of $\boldsymbol{\beta}$ solves,

$$\mathbf{X}_c\boldsymbol{\beta} = \hat{\boldsymbol{\eta}}_W, \tag{47}$$

i.e., $\hat{\boldsymbol{\beta}}_W = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\hat{\boldsymbol{\eta}}_W$. There are several estimates available for the intercept, but since we want to allow for skewed error distributions, the median of the residuals seems appropriate; i.e.,

$$\hat{\alpha}_S = \underset{i}{\text{med}}\{Y_i - \hat{\eta}_{W,i}\}. \tag{48}$$

The asymptotic distribution of the Wilcoxon estimates is given by

$$\hat{\mathbf{b}}_W = \begin{pmatrix}\hat{\alpha}_S\\ \hat{\boldsymbol{\beta}}_W\end{pmatrix} \text{ has an approximate } N_{p+1} \times\left(\begin{pmatrix}\alpha\\ \boldsymbol{\beta}\end{pmatrix}, \begin{bmatrix} n^{-1}\tau_S^2 & \mathbf{0}'\\ \mathbf{0} & \tau^2(\mathbf{X}_c'\mathbf{X}_c)^{-1}\end{bmatrix}\right) \text{ distribution,} \tag{49}$$

where the scale parameters τ_S and τ are given, respectively, by (29) and (12). Note that the asymptotic variance of $\hat{\boldsymbol{\beta}}_W$ differs from the asymptotic variance of the LS only by the constant of proportionality, τ^2 for the Wilcoxon and σ^2 for LS. Thus the efficiency properties of the Wilcoxon regression estimates to the LS regression estimates are the same as those in the simple location problems. As we will discuss in Section III, these Wilcoxon regression estimates have bounded influence in the **Y**-space, but unbounded influence in the **x**-space.

For later use, define the $\|\cdot\|_W$-distance between $\mathbf{Y}$ and Ω by

$$d_W(\mathbf{Y}, \Omega) = \|\mathbf{Y} - \hat{\boldsymbol{\eta}}_W\|_W. \tag{50}$$

Also, we can write the Wilcoxon norm as

$$\|\mathbf{v}\|_W = \sum_{i}^{n} a(R(v_i))v_i, \tag{51}$$

where $R(v_i)$ denotes the rank of v_i among $v_1, \ldots, v_n$ and $a(i) = \varphi_W(i/(n+1))$ and $\varphi_W(u) = \sqrt{12}(u - (1/2))$. This linear function $\varphi_W(u)$ is called the **Wilcoxon score function**. Finally, consider the norm as a function of $\boldsymbol{\beta}$, i.e.,

$$D(\boldsymbol{\beta}) = \sum_{i}^{n} a(R(Y_i - \mathbf{x}_i'\boldsymbol{\beta}))(Y_i - \mathbf{x}_i'\boldsymbol{\beta}). \tag{52}$$

This is a convex, continuous function of $\boldsymbol{\beta}$. It is often called Jaeckel's (1972) **dispersion function**. Note that the ranks involved are ranks of residuals, not of observations. Upon fitting, the residuals are estimates of iid random variables, so the ranking is similar to use of the ranking in the location models, Section I. Since the estimates are rank-based they are often referred to as **R** estimates.

B. Confidence Intervals

Formulae for approximate confidence intervals for the regression parameters based on the Wilcoxon estimates follow from the asymptotic distribution. Let $\hat{e}_{Wi} = Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_W$ denote the Wilcoxon residuals. Let $\hat{\tau}_S$ and $\hat{\tau}$ be the estimates of τ_S and τ based on these residuals that are discussed Section II.F. Then, in particular, $(1-\alpha)100\%$ confidence intervals for β_j and for a linear combination $\mathbf{h}'\boldsymbol{\beta}$ are given by

$$\hat{\beta}_{Wj} \pm t_{\alpha/2,n-(p+1)}\hat{\tau}\sqrt{(\mathbf{X}_c'\mathbf{X}_c)_{jj}^{-1}} \tag{53}$$

$$\mathbf{h}'\hat{\boldsymbol{\beta}}_W \pm t_{\alpha/2,n-(p+1)}\hat{\tau}\sqrt{\mathbf{h}'(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{h}}, \tag{54}$$

where $t_{\alpha/2,n-(p+1)}$ is the upper $(1-\alpha/2)$ Student t-critical value with $n-(p+1)$ degrees of freedom and $(\mathbf{X}_c'\mathbf{X}_c)_{jj}^{-1}$ is the jth diagonal entry of the matrix $(\mathbf{X}_c'\mathbf{X}_c)^{-1}$. These are the same confidence intervals one would use for a LS analysis except the scale parameter τ would be replaced by σ.

C. Tests of Linear Hypotheses

Hypotheses of interest in linear models are usually expressed as a linear function of the vector of parameters $\boldsymbol{\beta}$, i.e.,

$$H_0: \mathbf{M}\boldsymbol{\beta} = \mathbf{0} \text{ versus } H_A: \mathbf{M}\boldsymbol{\beta} \neq \mathbf{0}, \tag{55}$$

where $\mathbf{M}$ is a $q \times p$ matrix of rank q. Let ω denote the $p-q$ dimensional subspace of Ω constrained by H_0; that is, $\omega = \{\boldsymbol{\eta} \in \Omega \mid \boldsymbol{\eta} = X_c\boldsymbol{\beta} \,\&\, \mathbf{M}\boldsymbol{\beta} = \mathbf{0}\}$. In this terminology, Ω is the full model subspace and ω is the reduced model subspace, i.e., the full model constrained by H_0. As shown, let $d(\mathbf{Y}, \Omega)$ and $d(\mathbf{Y}, \omega)$ be the "distances" between $\mathbf{Y}$ and each of the subspaces and let $RD = d(\mathbf{Y}, \omega) - d(\mathbf{Y}, \Omega)$ denote the reduction in distance as we pass from the full to the reduced model. Small values of RD indicate H_0 while large values indicate H_A; hence, we reject H_0 in favor of H_A, if RD is too large. If d is the squared Euclidean distance function then RD is the traditional reduction in sums of squares. If d_W is the Wilcoxon normed distance function, (50), then we will call RD the **reduction in dispersion** and label it as RD_W, i.e.,

$$RD_W = d_W(\mathbf{Y}, \omega) - d_W(\mathbf{Y}, \Omega). \tag{56}$$

A formal test statistic is given by

$$F_W = \frac{RD_W/q}{\hat{\tau}/2}. \tag{57}$$

The null asymptotic distribution of F_W is given by,

$$qF_W \text{ has an approximate } \chi^2(q) \text{ distribution under } H_0\colon \mathbf{M}\boldsymbol{\beta} = \mathbf{0}, \tag{58}$$

where $\chi^2(q)$ denotes a χ^2 distribution with q degrees of freedom. Empirical evidence has shown, though, that F-critical values yield better small-sample properties. An approximate level α test is

$$\text{Reject } H_0\colon \mathbf{M}\boldsymbol{\beta} = \mathbf{0} \text{ in favor of } H_A\colon \mathbf{M}\boldsymbol{\beta} \neq \mathbf{0} \text{ if } F_W \geq F(\alpha, q, n-(p+1)), \tag{59}$$

where $F(\alpha, q, n-(p+1))$ denotes the upper α F-critical value with q and $n-(p+1)$ degrees of freedom. The test can be described in an ANOVA table that is quite similar to the traditional ANOVA table (See example in Section V.)

Besides the null behavior, the asymptotic distribution of F_W can be obtained for local alternatives. For the local alternative $\mathbf{M}\boldsymbol{\beta}^* \neq \mathbf{0}$, the asymptotic distribution of F_W is given by

$$qF_W \text{ has an approximate } \chi^2(q, \delta) \text{ with noncentrality parameter: } \delta = \tau^{-2}(\mathbf{M}\boldsymbol{\beta}^*)'\left[\mathbf{M}(\mathbf{X}_c\mathbf{X}_c')^{-1}\mathbf{M}'\right]^{-1}(\mathbf{M}\boldsymbol{\beta}^*). \tag{60}$$

This is the same noncentral χ^2 distribution of the LS traditional F-test statistic, except τ^2 replaces σ^2. In particular, the Wilcoxon F-test has the same efficiency properties as the Wilcoxon tests and estimates in the simple location problems. Furthermore, in the design of experiments, sample size determination can be obtained similar to the determination based on the traditional F-test, except for the use of τ^2 in place of σ^2.

Consider the special case where we are testing that all the β's are 0. For this situation, the reduced subspace is $\omega = \{\mathbf{0}\}$. Hence $d(\omega, \mathbf{Y}) = D(\mathbf{0})$, where D is the dispersion function (52) (see Section II.E).

D. Example: Telephone Data

We will illustrate Wilcoxon estimation and testing for a simple linear regression example. The response for this data set is the number of telephone calls (Table V) (tens of millions) made in Belgium for the years 1950 through 1973 and time, the years, is the predictor variable. [See Hettmansperger and McKean (1998) for original references.]

The Wilcoxon estimates of the intercept and slope are -7.13 and .145, respectively, while the LS estimates are -26 and .504. The estimate of τ is 2.90. The reason for this disparity in fits is easily seen in Panel A of Fig. 1 which is a scatterplot of the data overlaid with the LS and Wilcoxon fits. Note that the years 1964 through 1969 had a profound effect on the LS fit, whereas the Wilcoxon fit was much less sensitive to these years. The recording system for the years 1964 through 1969 differed from the other years. Panels B and C of Fig. 1 are the Studentized (see Section II.G.2) residual plots of the fits. Studentized residuals which exceed 2 in absolute value are potential outliers. Note that the internal Wilcoxon Studentized residuals clearly show that the years 1964–1969 are outliers while the internal LS Studentized residuals only detect 1969.

Panel D of Fig. 1 depicts the Wilcoxon dispersion function (52) over the interval $(-.2, .6)$. Note that Wilcoxon estimate $\hat{\beta}_R = .145$ is the minimizing value and that in the previously cited notation the minimum value $D(.145)$ is $d_W(\mathbf{Y}, \Omega)$. Next consider the hypotheses $H_0\colon \beta = 0$ versus $H_A\colon \beta \neq 0$. Note from the plot that the reduction in dispersion is given by $RD_W = D(0) - D(.145)$.

TABLE V Telphone Data

Year	50	51	52	53	54	55	56	57	58	59	60	61
No. calls	0.44	0.47	0.47	0.59	0.66	0.73	0.81	0.88	1.06	1.20	1.35	1.49
Year	62	63	64	65	66	67	68	69	70	71	72	73
No. calls	1.61	2.12	11.90	12.40	14.20	15.90	18.20	21.20	4.30	2.40	2.70	2.90

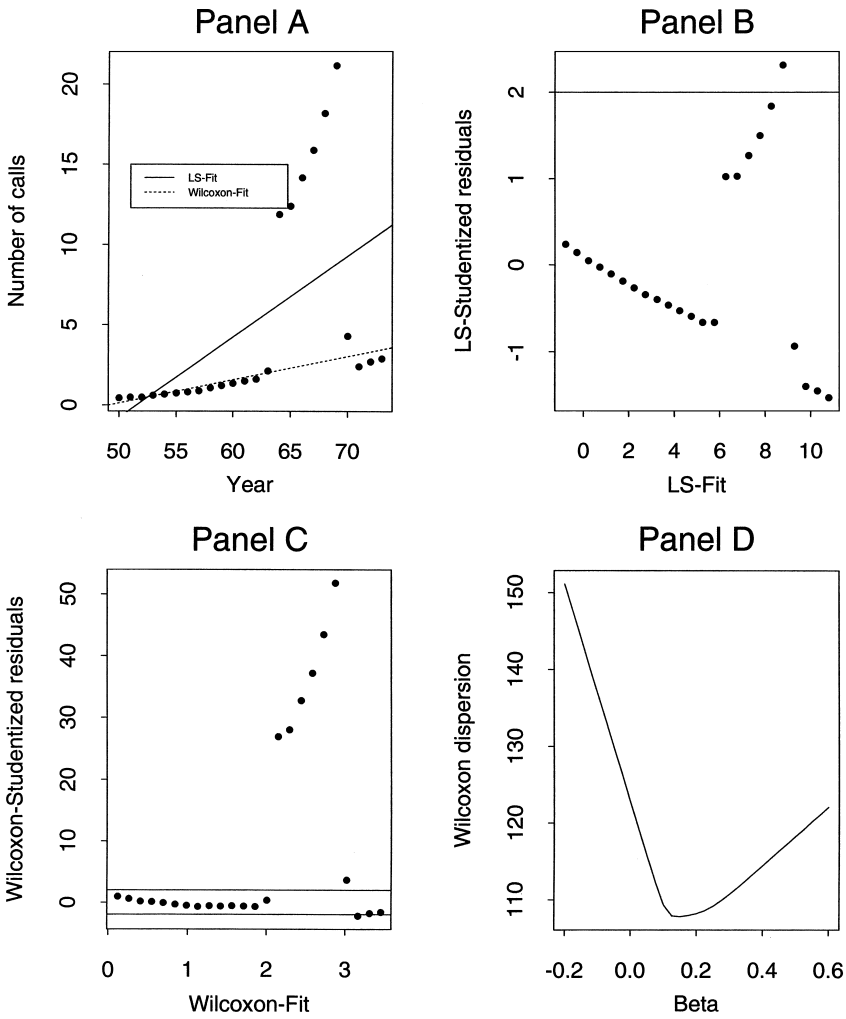


FIGURE 1 Panel A: Scatterplot of the Telephone Data, overlaid with the LS and Wilcoxon fits; Panel B: Internal LS Studentized residual plot; Panel C: Internal Wilcoxon Studentized residual plot; and Panel D: Wilcoxon dispersion function.

E. Coefficients of Multiple Determination

The LS coefficient of multiple determination R^2 is associated with the test of the hypothesis that all the slope parameters are 0, i.e.,

$$H_0: \boldsymbol{\beta} = \mathbf{0} \text{ versus } H_A: \text{at least one } \beta_j \neq 0. \tag{61}$$

The statistic R^2 is the ratio in reduction in sums of squares (due to H_0) to the total variation. An immediate Wilcoxon analog to this would be the ratio of the reduction in dispersion to the **total dispersion** ($\|\mathbf{Y}\|_W = \sum\sum |Y_i - Y_j|$). For R^2 there are two sources of nonrobustness: the reduction in sums of squares and the total variation. For this analog, while the reduction in dispersion is robust, the total dispersion is not. There is a simple robust analog, however. The statistic R^2 can be written also as $R^2 = (\kappa_{n,p} F_{LS}/(1 + \kappa_{n,p} F_{LS})$, where $\kappa_{n,p} = p/(n - (p+1))$, and F_{LS} is the LS F-test for the hypotheses (61). If we replace F_{LS} by F_W we get

$$R_W = \frac{\kappa_{n,p} F_W}{1 + \kappa_{n,p} F_W}. \tag{62}$$

Because the Wilcoxon test statistic F_W is robust, it follows that R_W is robust.

F. Estimates of the Scale Parameters τ and τ_S

The constant of proportionality τ plays a key role in the Wilcoxon analysis. In this section we define a consistent estimate of τ.

Assume we have obtained the Wilcoxon fit of the model (39). Let $\hat{\mathbf{e}}_W$ denote the vector of residuals and let $\varphi_W(u) = \sqrt{12}(u - (1/2))$ denote the Wilcoxon score function. Define the following empirical distribution function of these residuals,

$$\hat{H}_n(y) = \frac{1}{c_n n} \sum_{i=1}^{n} \sum_{j=1}^{n} \varphi_W\left(\frac{j}{n+1}\right) I\left(\left|\hat{e}_{W(i)} - \hat{e}_{W(j)}\right| \le y\right), \tag{63}$$

where $\hat{e}_{W(i)}$ denotes the ith ordered residual and $c_n = \sum_{j=1}^{n} \varphi_W(j/(n+1))$ is such that $\hat{H}_n$ is a distribution function; i.e., $\hat{H}_n(\infty) = 1$.

Let $\gamma = \tau^{-1}$. Since $\hat{H}_n$ is a distribution function, let $\hat{t}_{n,\delta}$ denote the δth quantile of $\hat{H}_n$; i.e., $\hat{t}_{n,\delta} = \hat{H}_n^{-1}(\delta)$. Then take $t_n = t_{n,\delta}/\sqrt{n}$. Our estimate of γ is given by

$$\hat{\gamma}_{n,\delta} = \frac{(\varphi_W(1) - \varphi_W(0))\hat{H}_n(t_{n,\delta}/\sqrt{n})}{2t_{n,\delta}/\sqrt{n}}. \tag{64}$$

The expression (63) for $\hat{H}_n$ contains a density estimate of f based on a rectangular kernel. Hence, in choosing δ we are really choosing a bandwidth for a density estimator. As most kernel type density estimates are sensitive to the bandwidth, $\gamma^*_{n,\delta}$ is sensitive to δ. Several small-sample studies have been done on this estimate to investigate the validity of procedures standardized by it. [See McKean and Shteather (1991) for a summary.] For moderate sample sizes where the ratio of n/p exceeds 5, the value of $\delta = .80$ yielded reasonable valid procedures. For ratios less than 5, larger values of δ, .95, yielded reasonable valid procedures. In all cases it was found that the following simple degrees of freedom correction benefited the analysis

$$\hat{\tau} = \sqrt{\frac{n}{n-p-1}}\hat{\gamma}^{-1}. \tag{65}$$

Note that this is similar to the least squares correction on the maximum likelihood estimate (under normality) of the variance. This last expression is the estimate of τ used in the sequel with $\delta = .80$.

The length of an asymptotically distribution-free 95% confidence interval for the intercept provides a simple, but reasonably valid estimate for τ_S. This is given by

$$\hat{\tau}_S = \sqrt{n}\left[\hat{e}_{W,(n/2+n^{1/2})} - \hat{e}_{W,(n/2-n^{1/2})}\right]/4. \tag{66}$$

G. Diagnostics

The assumption of a linear model is a major assumption, one that should be checked based on the data. There are many diagnostics available for Wilcoxon fits and in this section we look at several important checks.

1. Residual Plots

The major assumption of a linear model is that the errors are iid. In particular, they are to be independent of $\alpha + \mathbf{x}_i'\boldsymbol{\beta}$. A check of this assumption is the standard **residual plot** based on the fits, i.e., the plot of the residuals $\hat{e}_i$ versus $\hat{\alpha} + \mathbf{x}_i'\hat{\boldsymbol{\beta}}$. A random scatter in this plot is confirmatory of the model while patterns in the plot contradict the model. Consider fitting a misspecified model. Suppose the true model is

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{X}_c\boldsymbol{\beta} + \mathbf{Z}_c\boldsymbol{\gamma} + \mathbf{e}, \tag{67}$$

where $\mathbf{Z}_c$ is an $n \times q$ centered matrix of constants and $\boldsymbol{\gamma} = \boldsymbol{\theta}/\sqrt{n}$, for $\boldsymbol{\theta} \neq \mathbf{0}$. Suppose, though, we fit the model $\mathbf{Y} = \mathbf{1}\alpha + \mathbf{X}_c\boldsymbol{\beta} + \mathbf{e}$. Then (based on first-order asymptotics),

$$\hat{\mathbf{Y}}_W \doteq \alpha\mathbf{1} + \mathbf{X}_c\boldsymbol{\beta} + \tau\mathbf{H}_c\boldsymbol{\varphi}_W(F(\mathbf{e})) + \mathbf{H}_c\mathbf{Z}_c\boldsymbol{\gamma} \tag{68}$$

$$\hat{\mathbf{e}}_W \doteq \mathbf{e} - \tau\mathbf{H}_c\boldsymbol{\varphi}_W(F(\mathbf{e})) + (\mathbf{I} - \mathbf{H}_c)\mathbf{Z}_c\boldsymbol{\gamma}, \tag{69}$$

where $\mathbf{H}_c = \mathbf{X}_c(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'$ and $\varphi_W(u) = \sqrt{12}(u - (1/2))$ is the Wilcoxon score function. It also follows that the LS residuals and fitted values satisfy

$$\hat{\mathbf{Y}}_{LS} = \alpha\mathbf{1} + \mathbf{X}_c\boldsymbol{\beta} + \mathbf{H}_c\mathbf{e} + \mathbf{H}\mathbf{Z}\boldsymbol{\gamma} \tag{70}$$

$$\hat{\mathbf{e}}_{LS} = \mathbf{e} - \mathbf{H}_c\mathbf{e} + (\mathbf{I} - \mathbf{H}_c)\mathbf{Z}_c\boldsymbol{\gamma}. \tag{71}$$

Suppose that the linear model (39) is correct. Then $\boldsymbol{\gamma} = \mathbf{0}$ and $\hat{\mathbf{e}}_W$ is a function of the random errors similar to $\hat{\mathbf{e}}_{LS}$; hence, it follows that a plot of $\hat{\mathbf{e}}_W$ versus $\hat{\mathbf{Y}}_W$ should generally be a random scatter, similar to the least squares residual plot.

In the case of model misspecification, note that the Wilcoxon residuals and least squares residuals have the same bias, namely, $(\mathbf{I} - \mathbf{H})\mathbf{Z}\boldsymbol{\gamma}$, which depend on the misspecified part. Hence Wilcoxon residual plots, similar to those of least squares, are useful in identifying model misspecification.

Other model violations can be discovered by this residual plot. For example, the model assumption also implies that the errors are homoscedastic, i.e., have the same variance. **Flat** residual plots, in which the residuals are randomly distributed between two horizontal lines, are confirmatory to this. Fan-shaped residual plots in which the scale of the residuals varies with the fitted value are an indication of heteroscedasticity. Outlier detection is another prime reason to obtain a residual plot. Here scaled residuals as described in the following are useful in this detection. Sine wave patterns in the plot are one indication of a time series effect. Time series effects are frequently detected by **lag plots** of the residuals, i.e., plots of $\hat{e}_i$ versus $\hat{e}_{i-k}$ for various k. Besides the plot of residuals versus fitted values, histograms and q–q plots are helpful in assessing distributional properties of the errors.

2. Studentized Residuals

Let $\hat{e}_{W,i}$ denote the ith residual based on the Wilcoxon estimate of the regression coefficients. An estimate of $\mathrm{Var}(\hat{e}_{W,i})$ is

$$\tilde{s}^2_{W,i} = \text{MAD}^2\left(1 - \hat{K}_1\frac{1}{n} - \hat{K}_2 h_{c,i}\right), \tag{72}$$

where $h_{ci} = \mathbf{x}_i(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{x}_i$; $\text{MAD} = 1.483\text{med}_i\{|\hat{e}_{Wi} - \text{med}_j\hat{e}_{Wj}|\}$;

$$\hat{K}_1 = \frac{\hat{\tau}_S^2}{\hat{\sigma}^2}\left(\frac{2\hat{\delta}_S}{\hat{\tau}_S} - 1\right), \tag{73}$$

$$\hat{K}_2 = \frac{\hat{\tau}^2}{\hat{\sigma}^2}\left(\frac{2\hat{\delta}}{\hat{\tau}} - 1\right), \tag{74}$$

$$\hat{\delta}_S = \frac{1}{n-p}\sum |\hat{e}_{W,i}|, \tag{75}$$

and

$$\hat{\delta} = \frac{1}{n-p}D(\hat{\beta}_W).$$

We define the **internal R-Studentized residuals** as

$$r_{W,i} = \frac{\hat{e}_{W,i}}{\tilde{s}_{W,i}} \quad \text{for } i = 1, \ldots, n. \tag{76}$$

Notice that the Studentized residuals are adjusted for both underlying variation and location in factor space. The usual **benchmark** for Studentized residuals in ± 2; i.e., cases whose Studentized residuals exceed 2 in absolute value are deemed potential outliers which may require investigation. Studentized residuals are useful in plots featuring residuals.

H. Example: Cloud Data

We illustrate the previous discussion on a data set with one predictor. [See Hettmansperger and McKean (1998) for original references.] The dependent variable is the cloud point of a liquid, a measure of degree of crystallization in a stock. The independent variable is the percentage of I-8 in the base stock. The data are in Table VI.

Panel A of Fig. 2 displays the residual plot (R-residuals versus R-fitted values) of the R-fit of the simple linear model. The curvature in the plot indicates that this model is a poor choice and that a higher degree polynomial model would be more appropriate. Panel B of Fig. 2 displays the residual plot from the R-fit of a quadratic model. Some curvature is still present in the plot. A cubic polynomial was fitted next. Its R-residual plot, found in Panel C of Fig. 2, is much more of a random scatter than the first two plots. On the basis of residual plots the cubic polynomial is an adequate model. Panel D of Fig. 2 displays the q–q plot of the Studentized Wilcoxon residuals. Note that the points which appear to be outliers in the residual plots have Studentized absolute residuals over 2 in this plot.

TABLE VI Cloud Data

%I-8	0	1	2	3	4	5	6	7	8	0
Cloud point	22.1	24.5	26.0	26.8	28.2	28.9	30.0	30.4	31.4	21.9
%I-8	2	4	6	8	10	0	3	6	9	
Cloud point	26.1	28.5	30.3	31.5	33.1	22.8	27.3	29.8	31.8	

Table VII offers summary statistics of the fits. The estimates of the coefficients for the models are given and for the cubic model their standard errors. Based on the estimates of τ and the coefficients of determination, it appears that the cubic model is the best of these three models. Based on these statistics, we can test several hypotheses of interest. The test statistic to test that the quadratic and cubic terms are needed is $F_W = ((13.15 - 4.76)/2)/(.332/2) = 25.27$ which is significant beyond the .000 level (degrees of freedom are 2 and 15). The test statistic to test that the cubic term is needed is $F_W = ((6.97 - 4.76)/1)/(.332/2) = 13.31$ which is significant at the .002 level (degrees of freedom are 1 and 15).

III. HIGH BREAKDOWN ROBUST ESTIMATES

For the robustness properties of regression estimators, we need to consider influential points in both the Y and the $\mathbf{x}$-spaces. Let $(\mathbf{x}^*, Y^*)$ be a possible influential point. The influence function of the LS estimator is $\Lambda(\hat{\beta}_{LS}; \mathbf{x}^*, Y^*) = n\sigma(\mathbf{X}_c'\mathbf{X}_c)^{-1}Y^*\mathbf{x}^*$. This function is unbounded in both the Y and the $\mathbf{x}$-spaces; hence, an outlier in response space or an outlier in factor space can impair the LS fit. The influence function of the Wilcoxon estimate is

$$\Lambda\left(\hat{\beta}_W; \mathbf{x}^*, Y^*\right) = n\tau(\mathbf{X}_c'\mathbf{X})^{-1}\varphi_W(Y^*)\mathbf{x}^*, \tag{77}$$

where $\varphi_W(u)$ is the Wilcoxon score function. Because $\varphi_W(u)$ is a bounded function, the Wilcoxon estimate has bounded influence in response space, but note that it has unbounded influence in the $\mathbf{x}$-space. So while the Wilcoxon estimate is highly efficient and robust in the Y-space it is highly sensitive to outliers in factor space. For controlled experimental designs this does not pose a problem, but for observational studies it may. Certainly, as in all cases, a residual analysis should be performed but if influential points form a cluster in factor space they may not be discovered (the so-called **masking effect**) by standard diagnostic analyses.

The weighted Wilcoxon estimate, described next, does have bounded influence in both the Y and the $\mathbf{x}$-spaces. To define our class of weights, we first need a high breakdown estimate of location and scatter (variance) in factor space. There are several estimates available. We have chosen the

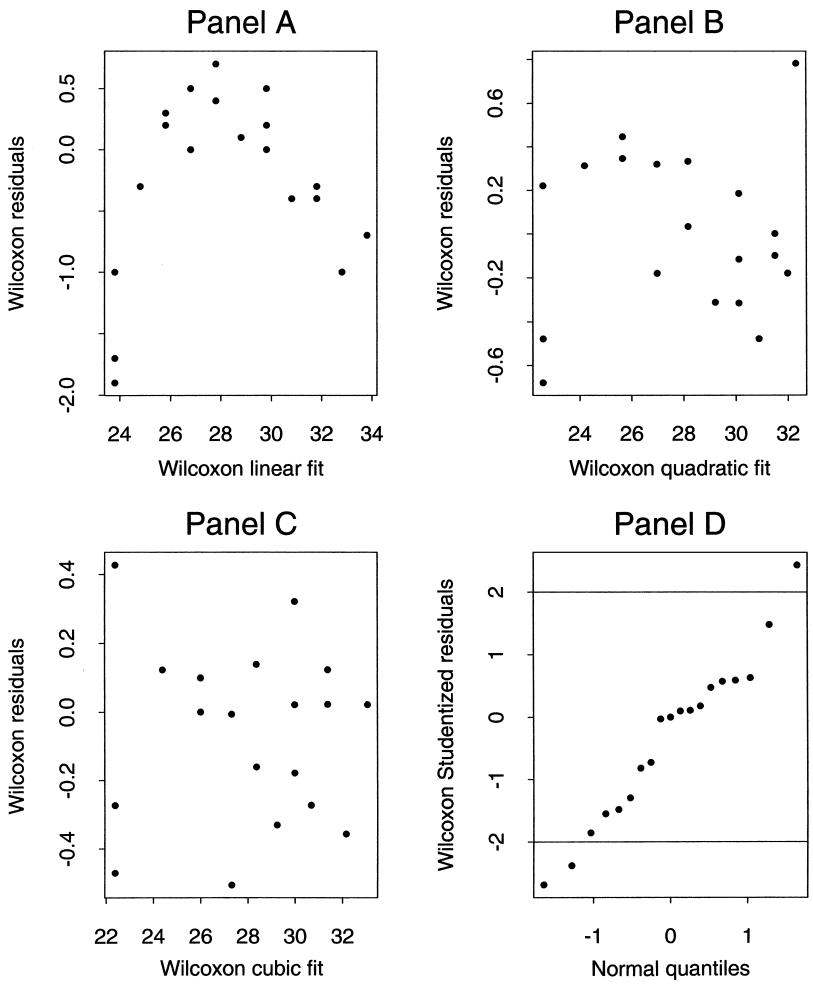


FIGURE 2 Panels A through C are the residual plots of the Wilcoxon fits of the linear, quadratic, and cubic models, respectively, for the Cloud Data. Panel D is the *q*–*q* plot of the Studentized residuals based on the Wilcoxon fit of the cubic model.

minimum covariance determinant (MCD) estimate of scatter. [See Rousseeuw and Van Driessen (1999) for a fast computational algorithm for the MCD.] The MCD is the ellipsoid which covers about half the data and yet has minimum determinant. This is the high breakdown estimate of scatter which we will use and which we label as **V**. Let **v** be the center of this ellipsoid. We next need a high breakdown initial estimate of the regression coefficients. We have chosen to use the **least trim squares** (LTS) estimate which is Argmin $\sum_{i=1}^{h}[Y-\alpha-\mathbf{x}'\beta]_{(i)}^2$, where $h=[n/2]+1$ and (i) denotes the ith ordered residual (see Rousseeuw and Van Driessen, 1999). Let $\hat{\mathbf{e}}_0$ denote the residuals from this initial fit.

TABLE VII Wilcoxon Summary Statistics for the Cloud Data

	Model			
	Linear	**Quadratic**	**Cubic**	**(SE)**
Statistic				
$\hat{\alpha}$	28.2	28.6	28.7	(.14)
$\hat{\beta}_1$	1.00	1.05	.878	(.06)
$\hat{\beta}_2$		−.075	−.079	(.01)
$\hat{\beta}_3$			.011	(.003)
Dispersion	13.15	6.97	4.26	
$\hat{\tau}$	.737	.493	.332	
R_W	.888	.934	.959	

Define the function $\psi(t)$ by $\psi(t)=1,\ t$, or -1 according as $t\geq 1,\ -1<t<1$, or $t\leq -1$. Let σ be estimated by the initial scaling estimate $\text{MAD}=1.483\ \text{med}_i|\hat{e}_i^{(0)}-\text{med}_j\{\hat{e}_j^{(0)}\}|$. Letting $Q_i=(\mathbf{x}_i-\mathbf{v})'\mathbf{V}^{-1}(\mathbf{x}_i-\mathbf{v})$, we can write

$$m_i=\psi\left(\frac{b}{Q_i}\right)=\min\left\{1,\frac{b}{Q_i}\right\}.$$

Hence the weights can be estimated by

$$\hat{b}_{ij} = \min\left\{1, \frac{c\hat{\sigma}}{|\hat{e}_i|}\frac{\hat{\sigma}}{|\hat{e}_j|}\min\left\{1, \frac{b}{\hat{Q}_i}\right\}\min\left\{1, \frac{b}{\hat{Q}_j}\right\}\right\}, \tag{78}$$

where the tuning constants b and c are both set at 4. From this point of view, it is clear that these weights downweight both outlying points in factor space and outlying responses.

Given these weights, the HBR Wilcoxon estimate is defined as,

$$\hat{\boldsymbol{\beta}}_{HBR} = \text{Argmin} \sum_{i,j} b_{ij}|Y_i - Y_j - (\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\beta}|. \tag{79}$$

Once the weights are determined, the estimates are obtained by minimizing a convex function; hence, a routine similar to that used to compute the Wilcoxon estimates can be used to obtain these estimates.

The HBR estimates have the following properties.

1. $\hat{\boldsymbol{\beta}}_{HBR}$ has a 50% breakdown point, provided the initial estimates used in forming the weights have 50% breakdown.
2. The influence function $\Lambda(\hat{\boldsymbol{\beta}}_{HBR}; \mathbf{x}^*, Y^*)$ is a bounded function in both the Y and the $\mathbf{x}$-spaces. Further, $\Lambda(\hat{\boldsymbol{\beta}}_{HBR}; \mathbf{x}^*, Y^*)$ is continuous everywhere and goes to zero as $(\mathbf{x}^*, Y^*)$ get large in any direction.
3. The asymptotic distribution $\hat{\boldsymbol{\beta}}_{HBR}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and standard errors as discussed in the following. As with all high breakdown estimates, $\hat{\boldsymbol{\beta}}_{HBR}$ is less efficient than the Wilcoxon estimates and this loss can be substantial on "good" data sets. The HBR estimates, however, are much more efficient than the initial LTS regression estimates.

As discussed in the following, the difference between high breakdown estimates and highly efficient estimates can be a powerful diagnostic in exploring messy data sets.

A. Standard Errors of the HBR Estimates

The asymptotic variance–covariance of $\hat{\boldsymbol{\beta}}_{HBR}$ is a function of the two matrices $\boldsymbol{\Sigma}$ and $\mathbf{C}$. The matrix $\boldsymbol{\Sigma}$ is the variance–covariance matrix of a random vector $\mathbf{U}_i$ which can be approximated by the expression,

$$\hat{\mathbf{U}}_i = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j - \mathbf{x}_i)\hat{b}_{ij}(1 - 2F_n(\hat{e}_i)), \tag{80}$$

where $\hat{b}_{ij}$ are the estimated weights, $\hat{e}_i$ are the HBR residuals, and F_n is the empirical distribution function of the residuals. Let $\hat{\boldsymbol{\Sigma}}$ be the sample variance–covariance matrix of $\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_n$. This serves as an estimate of $\boldsymbol{\Sigma}$.

The second matrix $\mathbf{C}$ can be approximated as $n^{-2}\mathbf{X}_c'\mathbf{A}\mathbf{X}_c$, where the (i, j) entry of $\mathbf{A}$ is $a_{ij} = -\hat{b}_{ij}/(\sqrt{12}\hat{\tau})$. Then the asymptotic variance–covariance matrix of $\hat{\boldsymbol{\beta}}_{HBR}$ can be approximated as $(4n)^{-1}\mathbf{C}^{-1}\hat{\boldsymbol{\Sigma}}\mathbf{C}^{-1}$. This can be uesd to determine asymptotic confidence intervals for linear functions of $\boldsymbol{\beta}$, for instance, the standard error of $\hat{\boldsymbol{\beta}}_{HBR,j}$ would be the square root of the jth diagonal entry of this matrix. Studentized residuals can be determined similarly.

B. Diagnostics To Differentiate between HBR and Wilcoxon Fits

For a given data set, highly efficient robust estimates and high breakdown estimates can produce very different fits. This can be due to influential points in factor space and/or curvature. High breakdown estimates tend to fit the center of the data so they may have trouble fitting areas of curvature away from the center. The diagnostics presented next indicate first whether or not the HBR and Wilcoxon fits differ and second, if they do differ, what cases are involved in the discrepancy.

First, as with the Wilcoxon estimates, estimate the intercept by the median of the HBR residuals, i.e., $\hat{\alpha}_{HBR} = \text{med}\{Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{HBR}\}$. Let $\hat{\mathbf{b}}_{HBR} = (\hat{\alpha}_{HBR}, \hat{\boldsymbol{\beta}}_{HBR}')'$. Then the difference in regression estimates between HBR and Wilcoxon estimates is the vector $\hat{\mathbf{b}}_W - \hat{\mathbf{b}}_{HBR}$. This difference needs to be standardized. An effective standardization is the estimate of the variance–covariance of $\hat{\mathbf{b}}_W$. This produces the following statistic, which measures the total difference in the fits of $\hat{\mathbf{b}}_W$ and $\hat{\mathbf{b}}_{HBR}$,

$$TDBETAS_R = (\hat{\mathbf{b}}_W - \hat{\mathbf{b}}_{HBR})'\hat{\mathbf{A}}_W^{-1}(\hat{\mathbf{b}}_W - \hat{\mathbf{b}}_{HBR}), \tag{81}$$

where

$$\mathbf{A}_W = \text{Cov}\begin{pmatrix}\hat{\alpha}_W \\ \hat{\boldsymbol{\beta}}_W\end{pmatrix} = \begin{bmatrix}\tau_S^2/n & 0 \\ 0 & \tau^2(\mathbf{X}'\mathbf{X})^{-1}\end{bmatrix}.$$

Large values of $TDBETAS_R$ indicate a discrepancy between the fits. A useful cutoff value, **benchmark for** $TDBETAS_R$, is $(4(p+1)^2)/n$.

The diagnostic $TDBETAS_R$ measures the overall difference in the estimates. If it exceeds its benchmark then usually we want to determine the individual cases causing this discrepancy in the fits. As we mentioned earlier, these would be the cases to investigate to see whether they are outliers in factor space or cases involved with curvature. Let $\hat{y}_{W,i} = \hat{\alpha}_W + \mathbf{x}'\hat{\boldsymbol{\beta}}_W$ and $\hat{y}_{HBR,i} = \hat{\alpha}_{HBR} + \mathbf{x}'\hat{\boldsymbol{\beta}}_{HBR}$ denote the respective fitted values for the ith case. A statistic which detects the observations that are fitted differently is

$$CFITS_{R,i} = \frac{\hat{y}_{R,i} - \hat{y}_{GR,i}}{\left(n^{-1}\hat{\tau}_S^2 + h_{c,i}\hat{\tau}^2\right)^{\frac{1}{2}}}. \tag{82}$$

An effective **benchmark for** $CFITS_{R,i}$ is $2\sqrt{(p+1)/n}$. We should note here that the objective of the diagnostic *CFITS* is *not* outlier deletion. Rather the intent is to identify the *critical few* data points for closer study, because these critical few points often largely determine the outcome of the analysis or the direction that further analyses should take. This closer study may involve subject matter expertise, rotation and spin plots, or data-collection-site investigation for accuracy of measurements, among other things. In this regard, the proposed benchmarks are meant as a heuristic aid, not a boundary to some formal critical region.

C. Example: Stars Data

This data set is drawn from an astronomy study on the star cluster CYG OB1 which contains 47 stars. [See Hettmansperger and McKean (1998) for the data.] The response is the logarithm of the light intensity of the star while the independent variable is the logarithm of the temperature of the star. The data are shown in Panel A of Fig. 3. Note that four of the stars, called giants, are outliers in factor space while the rest of the stars fall in a point cloud. The three fits, LS, Wilcoxon, and the HBR, are overlaid on this plot of the data. Note that the HBR fit falls through the point cloud, whereas the other two fits are drawn toward the outliers in factor space. This illustrates the HBR estimates' insensitivity to outliers in factor space. Panels B and C of Fig. 3 show the residual plots for the Wilcoxon and HBR fits. The HBR fit has fitted the point cloud well and the four outlying stars stand out in the plot.

The diagnostic $TDBETAS_R$ has the value 109.9 which greatly exceeds the benchmark and, hence, numerically indicates that the fits differ. Panel D of Fig. 3 shows the casewise diagnostic $CFITS_{R,i}$ versus case. In this plot, the four giant stars clearly stand out from the rest. The plot shows that the fits for two other stars also differ. These are

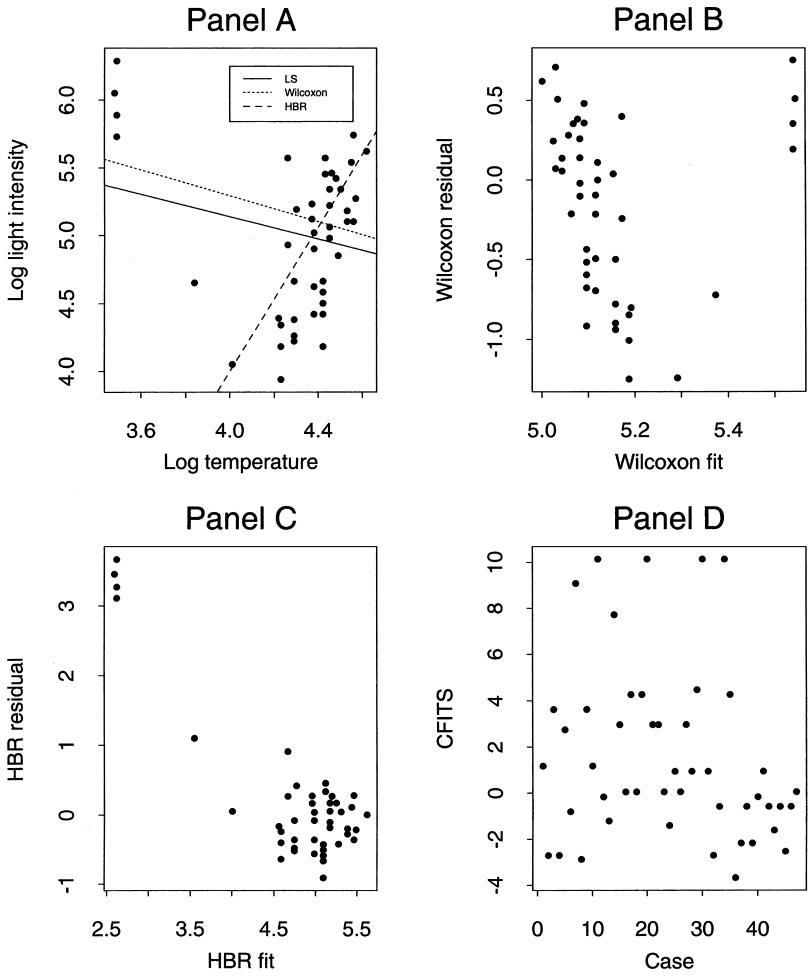


FIGURE 3 Panel A: Stars Data overlaid with LS, Wilcoxon, and HBR fits. Panel B: Wilcoxon residual plot. Panel C: HBR residual plot. Panel D: The diagnostic *CFITSR*, *i* versus case.

the stars between the giant stars and the rest of the stars as shown in Panel A.

As a final word on this example, suppose we only had the data and not the facts behind the data set. From Panel A, one might conclude that a quadratic model would be appropriate. Note that the Wilcoxon residual plot indicates a quadratic model here while the HBR residual plot does not. High breakdown estimates fit the center of the data and hence often do not detect curvature on the edge of factor space. But the diagnostics $TDBETAS_R$ and $CFITS_{R,i}$ would alert the user that the fits differ. Using further diagnostic plots, the user may be able to decide whether it is curvature or outliers in factor space which are causing the difference in fits.

IV. OPTIMAL RANK-BASED ANALYSES

Consider the linear model (38). Knowledge of the underlying density of the errors, $f(t)$, can be utilized to optimize the rank-based analysis. Recall that the Wilcoxon estimate of the regression coefficients minimizes the norm of the errors where the norm can be expressed as (51). Instead of using the Wilcoxon score function, other score functions can be selected. Suppose we select a function $\varphi(u)$ which is bounded and nondecreasing on (0, 1); (without loss of generality, $\int \varphi = 0$ and $\int \varphi^2 = 1$). Denote the resulting estimate which minimizes the norm of the errors by $\hat{\boldsymbol{\beta}}_\varphi$ and call it in general an **R** estimate. This estimate has the following properties:

1. The criterion function is convex, hence, $\hat{\boldsymbol{\beta}}_\varphi$ can be obtained using the same algorithm as is used by the Wilcoxon estimate (one simply replaces the Wilcoxon score function by $\varphi(u)$).
2. The influence function of $\hat{\boldsymbol{\beta}}_\varphi$ will be bounded in response space; i.e., the estimate is bias-robust.
3. $\hat{\boldsymbol{\beta}}_\varphi$ is asymptotically normal, centered at $\boldsymbol{\beta}$, and has variance–covariance matrix $\tau_\varphi^2(\mathbf{X}_c'\mathbf{X}_c)^{-1}$, where

$$\tau_\varphi^{-1} = \int_0^1 \varphi(u)\varphi_f(u)\,du \tag{83}$$

and

$$\varphi_f(u) = -\frac{f'}{f}(F^{-1}(u)). \tag{84}$$

The scale parameter τ_φ can be computed using the same algorithm as the Wilcoxon's τ.

How do we select φ to optimize the analysis? We want to select scores so that the constant of proportionality τ_φ^2 is as small as possible or its reciprocal is as large as possible, i.e., maximize

$$\begin{aligned}\tau_\varphi^{-1} &= \int \varphi(u)\varphi_f(u)\,du \\ &= \sqrt{\int \varphi_f^2(u)\,du}\,\frac{\int \varphi(u)\varphi_f(u)\,du}{\sqrt{\int \varphi_f^2(u)\,du}\sqrt{\int \varphi^2(u)\,du}} \\ &= \text{def}\sqrt{\int \varphi_f^2(u)\,du}\,\rho. \end{aligned} \tag{85}$$

The second equation is true since the scores were standardized as shown here. In the third equation ρ is a correlation coefficient and $\int \varphi_f^2(u)\,du$ is **Fisher location information**, which we denote by $I(f)$. By the Rao–Cramér lower bound, the smallest asymptotic constant of proportionality obtainable by an asymptotically unbiased estimate is $I(f)^{-1}$. Such an estimate is called **asymptotically efficient**. Choosing a score function to maximize (85) is equivalent to choosing a score function to make $\rho = 1$. This can be achieved by taking the score function to be $\varphi(u) = \varphi_f(u)$, (84). The resulting estimate, $\hat{\boldsymbol{\beta}}_\varphi$, is **asymptotically efficient**. The previous formulation is invariant to location and scale; hence, we need to know only the form of the density.

Here are several situations with their corresponding optimal score functions.

- If the form of the error density is the **logistic**, $f(x) = e^{-x^2}/(1 + e^{-x^2})$, then the Wilcoxon score function is optimal.
- If the form of the error density is the **normal**, $f(x) = (2\pi)^{-1/2}e^{-x^2/2}$, then the optimal score function is the **normal score** function given by $\varphi(u) = \Phi^{-1}(u)$, where $\Phi(t)$ denotes the distribution function of a standard normal random variable. Hence, under normal errors, R estimates obtained by using the normal scores have the same efficiency as the LS estimates.
- If the form of the error density is the Laplace density, $f(x) = (2)^{-1}e^{-|x|}$, then the optimal score function is the **sign score** function, given by $\varphi(u) = \text{sgn}(u - (1/2))$. Sign scores generalize the sign procedures discussed in the location models. While asymptotically efficient at the Laplace distribution, they only have efficiency .63 at the normal relative to LS. If the intercept is estimated as the median of residuals then these sign score estimates are the least absolute deviation L_1 regression estimates.

Asymptotically efficient rank-based estimates can be obtained if the form of f is known. Evidently, the closer the chosen score is to φ_f, the more powerful the rank analysis will be. There are adaptive analyses in which the score function is estimated based on an initial fit. For

certain areas such as survival analysis, interesting classes of score functions can readily be formulated.

A. Survival Analysis

In this section we discuss scores which are appropriate for lifetime distributions when the log of lifetime follows a linear model, i.e., **accelerated failure time models** (see Kalbfleisch and Prentice, 1980). Let T denote the lifetime of a subject and let $\mathbf{x}$ be a $p \times 1$ vector of covariates associated with T. Let $h(t;\mathbf{x})$ denote the **hazard function** of T at time t which is the rate of failure of an item at time t.

Suppose T follows a log linear model; that is, $Y = \log T$ follows the linear model

$$Y = \alpha + \mathbf{x}'\boldsymbol{\beta} + e, \tag{86}$$

where e is a random error with density f. Exponentiating both sides we get $T = \exp\{\alpha + \mathbf{x}'\boldsymbol{\beta}\}T_0$, where $T_0 = \exp\{e\}$. Let $h_0(t)$ denote the hazard function of T_0. This is called the baseline hazard function. Then the hazard function of T is given by

$$h(t;\mathbf{x}) = h_0(t\exp\{-(\alpha + \mathbf{x}'\boldsymbol{\beta})\})\exp\{-(\alpha + \mathbf{x}'\boldsymbol{\beta})\}. \tag{87}$$

Thus the covariate $\mathbf{x}$ **accelerates** or decelerates the failure time of T.

Many of the error distributions currently used for these models are contained in the **log**-F class. In this class, $e = \log T$ is distributed down to an unknown scale parameter, as the log of an F random variable with $2m_1$ and $2m_2$ degrees of freedom. We say that e has a $GF(2m_1, 2m_2)$ distribution. In general, this class contains a variety of shapes. The distributions are symmetric for $m_1 = m_2$, positively skewed for $m_1 > m_2$, and negatively skewed for $m_1 < m_2$. For values $0 \le m_1,\ m_2 \le 1$, the distributions tend to have thicker tails. For random errors with distribution $GF(2m_1,\ 2m_2)$, the optimal rank score function is given by

$$\varphi_{m_1,m_2}(u) = (m_1m_2(\exp\{F^{-1}(u)\} - 1))/ (m_2 + m_1\exp\{F^{-1}(u)\}), \tag{88}$$

where F is the cumulative distribution function of the $GF(2m_1,\ 2m_2)$ distribution. It follows that the scores are strictly increasing and bounded below by $-m_1$ and above by m_2. Hence an R-analysis based on these scores will have bounded influence in the Y-space.

Important subclasses of the accelerated failure time models are those where T_0 follows a Weibull distribution, i.e.,

$$f_{T_0}(t) = \lambda\gamma(\lambda t)^{\gamma-1}\exp\{-(\lambda t)^{\gamma}\}, \quad t > 0, \tag{89}$$

where λ and γ are unknown parameters. In this case it follows that the hazard function of T is proportional to the baseline hazard function with the covariate acting as the factor of proportionality, i.e.,

$$h(t;\mathbf{x}) = h_0(t)\exp\{-(\alpha + \mathbf{x}'\boldsymbol{\beta})\}. \tag{90}$$

Hence these models are called **proportional hazard models**. We can write the random error $e = \log T_0$ as $e = \xi + \gamma^{-1}W_0$, where $\xi = -\log\gamma$ and W_0 has the extreme value distribution. The optimal rank scores for these log-linear models are generated by the function

$$\varphi_{f_\epsilon}(u) = -1 - \log(1-u). \tag{91}$$

Procedures based on these scores are asymptotically equivalent to the procedures based on the **Savage** scores.

V. RANK-BASED ANALYSES OF EXPERIMENTAL DESIGNS

In this section we discuss the application of the rank-based analysis described in Section II to factorial designs and analysis of covariance designs.

A. One-Way Designs

For these designs we have a single factor, A, at k levels. The experimental design is the CRD design in which n experimental units are randomly assigned, n_i to level i, $n = \sum_{i=1}^{k} n_i$. At the end of the study, Y_{ij} denotes the response of the jth experimental unit assigned to level i. As a model, consider the **one-way** design,

$$Y_{ij} = \mu_i + e_{ij}, \quad j = 1,\ldots,n_i; \quad i = 1,\ldots,k, \tag{92}$$

where e_{ij} are iid with density $f(t)$ and μ_i is the location (e.g., mean or median) of the ith cell. Denote the vector of responses by $\mathbf{Y} = (Y_{11},\ldots,Y_{kn_k})'$, the vector of cell locations by $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_k)'$, and let $\mathbf{W}$ be the $n \times k$ incidence matrix. Then we can express this model as

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\mu} + \mathbf{e}. \tag{93}$$

Because a column of ones is in the column space of $\mathbf{W}$, we can always reparameterize this model so it is a linear model of the form (39). As in Section II, let Ω denote the column space of $\mathbf{W}$.

Suppose we fit this model using the Wilcoxon fit of Section II. Denote the fitted values by $\hat{\mathbf{Y}}_W$, the residuals by $\hat{\mathbf{e}}_W$, the estimate of τ by $\hat{\tau}$, and the minimum dispersion by $d_W(\mathbf{Y}, \Omega)$. Also denote the scored residuals by $\mathbf{a}(R(\hat{e}_W))$, where $a(i) = \varphi_W(i/(n+1))$ and $R(\hat{e}_i)$ denote the rank of the ith residual. Other score functions besides the Wilcoxon φ_W can be used but, here, we will only discuss the results for Wilcoxon scores. Note that the ranking here is on the residuals which are estimates of iid errors.

For easy reference, let us collectively label this Wilcoxon analysis fit as

$$\mathcal{F}_W = \mathcal{F}_W(\Omega) = \{\hat{\mathbf{Y}}_W, \hat{\mathbf{e}}_W, \mathbf{a}(R(\hat{e}_W)), \hat{\tau}, d_W(\mathbf{Y}, \Omega)\}. \tag{94}$$

We are interested in the overall null hypothesis that the levels are the same; i.e.,

$$H_0: \mu_1 = \cdots = \mu_k \text{ versus } H_A: \mu_i \neq \mu_{i'} \text{ for some } i \neq i'. \tag{95}$$

Upon rejecting this hypothesis, we are usually interested in the determination of significant differences between the levels, say, of the form $\mu_i - \mu_{i'}$. Note that these are **contrasts** in the μ_is; that is, linear functions of the form $\mathbf{h}'\boldsymbol{\mu} = \sum h_i \mu_i$, where $\sum h_i = 0$.

In general let $\mathbf{M}$ be a $q \times k$ matrix whose rows are contrasts. The corresponding hypotheses are

$$H_0: \mathbf{M}\boldsymbol{\mu} = \mathbf{0} \text{ versus } H_A: \mathbf{M}\boldsymbol{\mu} \neq \mathbf{0}. \tag{96}$$

We can write the hypotheses (95) in this form by taking $\mathbf{M}$ to be the $(k-1) \times k$ matrix whose ith row is $(1, 0, \ldots, 0, -1, 0, \ldots, 0)$ with the -1 appearing in the $(i+1)$ spot.

1. Tests of General Contrasts

Based on the results discussed in Section II, a robust test of the hypotheses (96), for a given contrast matrix $\mathbf{M}$, can be obtained. Let ω be the subspace of Ω subject to the constraints $\mathbf{M}\boldsymbol{\mu} = \mathbf{0}$ stipulated by H_0. Fit the full model and obtain the minimum dispersion $d_W(\mathbf{Y}, \Omega))$ and the estimate of scale $\hat{\tau}$. Next fit the reduced model and obtain the minimum dispersion $d_W(\mathbf{Y}, \omega))$. The reduction in dispersion is $RD = d_W(\mathbf{Y}, \omega)) - d_W(\mathbf{Y}, \Omega))$ and the test statistic is $F_W = (RD/q)/(\hat{\tau}/2)$. The tests reject H_0 at level α if $F_W \geq F(\alpha, q, n-k)$. Under local alternatives, the noncentrality parameters of these Wilcoxon tests are the same as their LS counterparts, except σ^2 is replaced by τ^2. Hence, designing experiments based on power can be carried out the same way for the Wilcoxon analyses as for LS analyses.

2. Test of $H_0: \mu_1 = \cdots = \mu_k$

Briefly to test the hypotheses (95), we obtain the fit $\mathcal{F}_W$ (94) of the full model (38). The reduced model is $\mathbf{Y} = \mathbf{1}_n \mu + \mathbf{e}$; hence, the reduced model dispersion is $\|\mathbf{Y}\|_W$. The reduction in dispersion is $RD_W = \|\mathbf{Y}\|_W - d_W(\mathbf{Y}, \mathbf{W})$ and the test statistic is F_W (57). We would reject H_0 at level α if $F_W \geq F(\alpha, k-1, n-k)$.

3. Confidence Intervals for Single Contrasts

Based on (54), the Wilcoxon confidence interval for any contrast can be obtained. In particular, this confidence interval will be the same as the one based on LS except $\hat{\tau}$ will replace $\hat{\sigma}$. For example, the Wilcoxon confidence interval for the difference in levels i and i' is,

$$\hat{\mu}_{Wi} - \hat{\mu}_{Wi'} \pm t_{(\alpha/2, n-k)} \hat{\tau} \sqrt{(1/n_i) + (1/n_{i'})}, \tag{97}$$

where μ_{Wi} are the full model Wilcoxon estimate of μ_i.

4. Extensions to Multiple Comparison Procedures

Often it is desired to look at many contrasts simultaneously; e.g., all $\binom{k}{2}$ pairwise differences. **Multiple comparison procedures** (MCP) of classical analysis give assurance to the loss of overall confidence for these simultaneous investigations. As demonstrated by the previously cited confidence interval for a single contrast, these traditional procedures are easy to robustify. For the most part, it is simply replacing $\hat{\sigma}$ by $\hat{\tau}$ and using the Wilcoxon estimates of location.

5. Kruskal–Wallis Tests of $H_0: \mu_1 = \cdots = \mu_k$

As stated earlier, the reduced model for this null hypothesis is $\mathbf{Y} = \mathbf{1}_n \mu + \mathbf{e}$. Under this reduced model the ranks of the Y_{ij}'s are uniformly distributed as they are for the null hypothesis in the two-sample location problem discussed in Section I. The MWW procedure discussed for the two-sample problem naturally extends to the the k-sample problem. The responses for all k samples are combined and ranked from 1 to n. Let $\bar{R}_i$ be the average of the ranks of the responses in the ith level. Then the **Kruskal–Wallis** test statistic is given by

$$H_W = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{n+1}{2} \right)^2. \tag{98}$$

This test statistic is distribution free under H_0 but it does depend on the configuration of sample sizes. Generally, the large-sample approximation, $\chi^2(k-1)$, is used. In this case, H_0 is rejected at level α provided $H_W \geq \chi^2(\alpha, k-1)$. As a final note, the Kruskal–Wallis test is asymptotically equivalent to the drop in dispersion test F_W for this hypothesis.

6. Pseudo-Observations

In this section, we discuss a convenient way to estimate and test contrasts based on the Wilcoxon fit $\mathcal{F}_W$ (94) of the full model (38). Given the fit we define the **pseudo-observations** by

$$\tilde{\mathbf{Y}} = \hat{\mathbf{Y}}_W + \hat{\tau} \zeta \mathbf{a}(R(\hat{\mathbf{e}}_W)), \tag{99}$$

where $\zeta^2 = (n-k)(n+1)/n(n-1)$.

Let $\hat{\tilde{\mathbf{Y}}}$ and $\hat{\tilde{\mathbf{e}}}$ denote the LS fit and residuals, respectively, of the pseudo-observations (99). Then

$$\hat{\hat{\mathbf{Y}}} = \hat{\mathbf{Y}}_W, \tag{100}$$

and

$$\hat{\hat{\mathbf{e}}} = \hat{\tau}\zeta \mathbf{a}(R(\hat{\mathbf{e}})). \tag{101}$$

From this last expression and the definition of ζ,

$$\frac{1}{n-k}\hat{\hat{\mathbf{e}}}'\hat{\hat{\mathbf{e}}} = \hat{\tau}^2. \tag{102}$$

Therefore the LS fit of the pseudo-observations results in the Wilcoxon fit of model (93) and, further, the LS estimator *MSE* is $\hat{\tau}^2$. This holds for estimation of contrasts also. Suppose we have a LS routine which tests the general hypotheses (96). Then a rank-based test will result if the pseudo-observations are inputted into the LS routine. This test will be asymptotically equivalent to the drop in dispersion test.

7. Tests for Ordered Alternatives

Ordered alternatives form another important class of alternative models. These are generalizations of the one-sided alternative hypotheses in the location problems. Furthermore, powerful yet highly efficient nonparametric procedures are readily formulated for this class of alternatives.

Consider the ordered alternative H_A given by

$$H_0: \mu_1 = \cdots = \mu_k \text{ versus } H_A: \mu_1 \leq \cdots \leq \mu_k,$$

with at least one strict inequality. The **Jonckheere–Terpstra** test statistic is defined by

$$J = \sum_{s<t} W_{st},$$

where $W_{st} = \#(Y_{tj} > Y_{si})$ for $i = 1, \ldots, n_s$ and $j = 1, \ldots, n_t$. The W_{st}'s are the pairwise Mann–Whitney statistics; see the discussion at expression (8). Under H_0,

$$(a) \quad E(J) = \frac{n^2 - \sum n_t^2}{4}.$$

$$(b) \quad V(J) = \frac{n^2(2n+3) - \sum n_t^2(2n_t+3)}{72}.$$

$$(c) \quad z = (J - E(J))/\sqrt{V(J)} \text{ is approximately } N(0,1).$$

Hence, based on (a)–(c) an asymptotic test for H_0 versus H_A, is to reject H_0 if $z \geq z_\alpha$.

B. Two-Way Designs

In this section, we consider the two-way crossed factorial design. Based on this discussion extensions to higher-order designs are straightforward. [See Chapter 4 of Hettmansperger and McKean (1998).] In a two-way crossed factorial design, we have two factors: A at a levels and B at b levels, that may have an effect on the response. Each combination of the $k = ab$ factor settings is a treatment. For a completely randomized designed experiment, n experimental units are selected at random from the reference population and then n_{ij} of these units are randomly assigned to the (i, j)th treatment combination; hence, $n = \sum\sum n_{ij}$. Let Y_{ijk} denote the response for the kth unit at the (i, j)th treatment combination.

If we impose no structure on the design, then this is a one-way design with k levels. Model (93) is the full model in this case. Let $\mathcal{F}_W(\Omega)$, (94), denote the Wilcoxon fit to this model. The submodels described in the following utilize the two-way structure of the design.

An interesting submodel is the **additive model** which is given by

$$\mu_{ij} = \bar{\mu} + (\bar{\mu}_{i\cdot} - \bar{\mu}) + (\bar{\mu}_{\cdot j} - \bar{\mu}). \tag{103}$$

For the additive model, the **profile plots** (μ_{ij} versus i or j) are parallel. A diagnostic check for the additive model is to plot the **sample profile plots**, ($\hat{\mu}_{ij}$ versus i or j) and see how close the profiles are to parallel. The null hypotheses of interest for this model are the **main effect hypotheses** given by

$$H_{0A}: \bar{\mu}_{i\cdot} = \bar{\mu}_{i'\cdot} \text{ for all } i, i' = 1, \ldots a \tag{104}$$

and

$$H_{0B}: \bar{\mu}_{\cdot j} = \bar{\mu}_{\cdot j'} \text{ for all } j, j' = 1, \ldots b. \tag{105}$$

There are $a-1$ and $b-1$ free constraints for H_{0A} and H_{0B}, respectively. Under H_{0A}, the levels of A have no effect on the response.

The **interaction parameters** are defined as the differences between the full model parameters and the additive model parameters, i.e.,

$$\begin{aligned}\gamma_{ij} &= \mu_{ij} - [\bar{\mu} + (\bar{\mu}_{i\cdot} - \bar{\mu}) + (\bar{\mu}_{\cdot j} - \bar{\mu})] \\ &= \mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}.\end{aligned} \tag{106}$$

The hypothesis of **no interaction** is given by

$$H_{0AB} = \gamma_{ij} = 0, \quad i = 1, \ldots, a, \quad j = 1, \ldots, b. \tag{107}$$

There are $(a-1)(b-1)$ free constraints for H_{0AB}. Under H_{0AB} the additive model holds.

These hypotheses are composed of contrasts of the μ_{ij}'s. Hence, the Wilcoxon rank-based test of these hypotheses can be obtained as discussed in Section V.A. The full model is the one way (93) with $k = ab$. Denote the Wilcoxon fit of this model by $\mathcal{F}_W(\Omega)$ as discussed in (94). For a given hypothesis, fit the reduced model, the full model constrained by the null hypothesis, and obtain the minimum dispersion. From this, the reduction in dispersion can be obtained and, hence, the test statistic F_W

as discussed in Section V.A.1. Alternatively, the pseudo-observations can be computed from $\mathcal{F}_W(\Omega)$ and a LS package can be used.

Usually the interaction hypothesis is tested first. If H_{0AB} is rejected then there is difficulty in interpretation of the main effect hypotheses, H_{0A} and H_{0B}. In the presence of interaction H_{0A} concerns the cell mean averaged over Factor B, which may have little practical significance. In this case multiple comparisons (see Section V.A.4) between cells may be of more practical significance. If H_{0AB} is not rejected then there are two schools of thought. The "pooling" school would take the additive model (103) as the new full model to test main effects. The "nonpoolers" would stick with the unstructured model (93) as the full model. In either case with little evidence of interaction present, the main effect hypotheses are much more interpretable.

1. Example: Lifetime of Motors

This data set is the result of an unbalanced two-way design. [See Hettmansperger and McKean (1998) for original references.] The responses are lifetimes of three motor insulations (1, 2, and 3), which were tested at three different temperatures (200, 225, and 250° F). The design is an unbalanced 3×3 factorial with five replicates in six of the cells and three replicates in the others. As discussed in Section IV, we considered the logs of the lifetimes as the responses. Let Y_{ijk} denote the log of the lifetime of the kth replicate at temperature level i and which used motor insulation j. These responses are displayed along with Wilcoxon estimates of the μ_{ij}'s in Table VIII.

The cell median profile plots based on the Wilcoxon estimates indicate that some interaction is present. The plot of the internal Wilcoxon Studentized residuals (not shown) versus fitted values indicates randomness but also shows several outlying data points: the fifth observation in cell (1, 1), the fifth observation in cell (2, 1), and the first observation in cell (2, 3).

Table IX is an ANOVA table for the R-analysis. Since $F(.05, 4, 30) = 2.69$, the test of interaction is significant at the .05 level. This confirms the profile plot. It is interesting to note that the least squares F-test statistic for interaction was 1.30 and, hence, was not significant. The LS analysis was impaired because of the outliers. Both main effects are significant. In the presence of interaction, though, we have interpretation difficulties with main effects.

TABLE VIII Data for Example, Lifetimes of Motors (hours)[a]

	Insulation		
Temperature	**1**	**2**	**3**
200°F	7.07	7.96	8.17
	7.32	8.07	8.17
	7.32	7.83	8.17
	7.32	8.07	
	8.17	8.17	
	(7.32)	(8.06)	(8.17)
225°F	6.44	6.70	6.58
	6.44	6.82	7.17
	6.44	7.17	7.31
	6.70	7.24	
	7.17	7.31	
	(6.58)	(7.11)	(7.15)
250°F	5.32	5.70	5.53
	5.43	5.78	5.70
	5.53	5.92	5.78
	5.70	5.92	
	5.78	6.10	
	(5.57)	(5.92)	(5.70)

[a] The last entry in each cell is the Wilcoxon estimate of the true cell median.

2. Friedman's Test

For the two-way and higher-order designs, there is no analog to the Kruskal–Wallis test. Recall that the reduced model in this case consisted of a column of ones; i.e., under H_0 all levels are the same and the responses are iid. This is not true of any of the usual hypotheses discussed for the two-way model.

For the randomized block design, however, we can change the way of ranking. This produces a distribution-free test but usually at a substantial loss of efficiency. Suppose we have k treatments of interest and we employ a block design consisting of a blocks, each of length k. Within each block, we randomly assign the k experimental units to the treatments. Suppose we model the responses Y_{ij} as

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}; \quad i = 1, \ldots, a, \quad j = 1, \ldots, k, \tag{108}$$

where e_{ij} are iid with density $f(t)$. We want to test

$$H_0: \beta_1 = \cdots = \beta_k \text{ versus } H_A: \beta_j \neq \beta_{j'} \text{ for some } j \neq j'.$$

TABLE IX Analysis of Dispersion Table for Lifetime of Motors Data

Source	***RD***	**df**	***MRD***	F_R
Temperature (T)	26.40	2	13.20	121.7
Motor insulation (I)	3.72	2	1.86	17.2
T × I	1.24	4	.310	2.86
Error		30	.108	

Considered as a two-way fixed effects design with no interaction, this test can be handled as discussed previously using the Wilcoxon rank-based analysis. The full model is (108).

Note that under H_0 there is no treatment effect, so the observations within a block are iid (but not between blocks). Thus we rank the items from 1 to k within each block. Let R_j be the sum of the ranks for the jth treatment. Under H_0 the expectation and variance of R_j are

$$\mathrm{E}(R_j) = \frac{a(k+1)}{2} \text{ and } \mathrm{Var}(R_j) = \frac{a(k+1)(k-1)}{12}.$$

Friedman's test statistic is given by

$$K = \sum_{j=1}^{k} \left(\frac{k-1}{k}\right)\left[\frac{R_j - E(R_j)}{\sqrt{\mathrm{Var}(R_j)}}\right]^2. \qquad (109)$$

This test statistic is distribution free under H_0 and is asymptotically a $\chi^2(k-1)$ random variable. Hence, the large sample test rejects H_0 at level α if $K \geq \chi^2(\alpha, k-1)$. The efficiency of the Friedman test depends on the number of treatments. At the normal distribution, this efficiency is $.955(k/(k+1))$. For $k=2$ this is .63, as it should be because at $k=2$ the Friedman test is the same as the paired sign test discussed in Section I. Even for 10 treatments, the efficiency is only .86.

C. Analysis of Covariance

In experimental design, we attempt to control all variables other than the factors and the response variable. Often this is impossible, so along with the response we record these extraneous variables which are called **covariates** or **concomitant variables**. Hopefully these variables help explain some of the noise in the data. The traditional analysis of such data is called **analysis of covariance**.

As an example, consider the one-way model (93) with k levels and suppose we have a single covariate, say, x_{ij}. A first-order model is $y_{ij} = \mu_i + \beta x_{ij} + e_{ij}$. This model, however, assumes that the covariate behaves the same within each treatment combination. A more general model is

$$y_{ij} = \mu_i + \beta x_{ij} + \gamma_i x_{ij} + e_{ij}$$
$$j = 1, \ldots, n_i,\ i = 1, \ldots, k. \qquad (110)$$

Hence the slope at the ith level is $\beta_i = \beta + \gamma_i$ and, thus, each treatment combination has its own linear model. There are two natural hypotheses for this model: $H_{0C}: \beta_1 = \cdots = \beta_k$ and $H_{0L}: \mu_1 = \cdots = \mu_k$. If H_{0C} is true then the difference between the levels of Factor A are just the differences in the location parameters μ_i for a given value of the covariate. In this case, contrasts in these parameters are often of interest as well as the hypothesis H_{0L}. If H_{0C} is not true then the covariate and the treatment combinations interact. For example, whether one treatment combination is better than another may depend on where in factor space the responses are measured. Thus as in crossed factorial designs, the interpretation of main effect hypotheses may not be clear. The previous example is easily generalized to more than one covariate.

The Wilcoxon fit of the full model (110) proceeds as described in Section II. Model estimates of the parameters and their standard errors can be used to form confidence intervals and regions and multiple comparison procedures can be used for simultaneous inference. Reduced models appropriate for the hypotheses of interest can be obtained and the values of the test statistic F_W can be used to test them.

VI. MEASURES OF ASSOCIATION

In this section, we discuss measures of association. Our discussion will be limited mostly to the bivariate model. Let (X, Y) be a pair of random variables with the joint density $f(x, y)$. Consider a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn on (X, Y). The hypotheses of interest are

$$H_0: X \text{ and } Y \text{ are independent versus}$$
$$H_A: X \text{ and } Y \text{ are dependent.} \qquad (111)$$

A. Kendall's τ

Monotonicity is an easily understood association between X and Y. Let (X_1, Y_1) and (X_2, Y_2) be independent pairs with density $f(x, y)$. We say these pairs are **concordant** if $\mathrm{sgn}\{(X_1 - X_2)(Y_1 - Y_2)\} > 0$ and are **discordant** if $\mathrm{sgn}\{(X_1 - X_2)(Y_1 - Y_2)\} < 0$. The variables X and Y have an increasing relationship if the pairs tend to be concordant and a decreasing relationship if the pairs tend to be discordant. A population measure of this is given by **Kendall's** τ,

$$\tau = P[\mathrm{sgn}\{(X_1 - X_2)(Y_1 - Y_2)\} > 0]$$
$$- P[\mathrm{sgn}\{(X_1 - X_2)(Y_1 - Y_2)\} < 0]. \qquad (112)$$

Positive values of τ indicate increasing monotonicity, negative values indicate decreasing monotonicity, and $\tau = 0$ reflects neither. Furthermore, if X and Y are independent then $\tau = 0$. The converse of this is not true; however, the contrapositive is true, i.e., $\tau \neq 0$ implies X and Y are dependent. Based on the sample, an unbiased estimate of τ is

$$K = \binom{n}{2}^{-1} \sum_{i<j} \mathrm{sgn}\{(X_i - X_j)(Y_i - Y_j)\};. \qquad (113)$$

If X and Y are independent then the statistic K is distribution free. Correct standardization under $\tau = 0$ leads to the test statistic,

$$z_K = \frac{K}{\sqrt{n(n-1)(2n+5)/18}}, \qquad (114)$$

which is asymptotically normal under $\tau = 0$. Large values of $|z_K|$ indicate $\tau \neq 0$. A large sample level α test is to reject X and Y are independent, if $|z_K| > z_{\alpha/2}$. Note by the contrapositive statement cited here, rejection of $\tau = 0$ leads to the rejection of independence between X and Y.

B. Spearman's ρ_S

The population correlation coefficient ρ is a measure of linearity between X and Y. The usual estimate is the sample correlation coefficient given by

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}. \qquad (115)$$

A simple rank analog is to replace X_i by $R(X_i)$, where $R(X_i)$ denotes the rank of X_i among $X_1, \ldots, X_n$, and likewise Y_i by $R(Y_i)$, where $R(Y_i)$ denotes the rank of Y_i among $Y_1, \ldots, Y_n$. Upon making this substitution, the denominator of the previous ratio is a constant. This results in the statistic,

$$r_S = \frac{\sum_{i=1}^{n}\left(R(X_i) - \frac{n+1}{2}\right)\left(R(Y_i) - \frac{n+1}{2}\right)}{n(n^2-1)/12}, \qquad (116)$$

which is called **Spearman's** ρ. Like Kendall's K statistic, r_S is an estimate of a population parameter, but it is a more complicated expression than τ. Under independence, this parameter is 0 and the statistic r_S is distribution free. Consider the standardized test statistic given by

$$z_S = \sqrt{n-1}\, r_s, \qquad (117)$$

which is asymptotically $N(0, 1)$ under independence. A large sample level α test is to reject independence between X and Y, if $|z_S| > z_{\alpha/2}$.

SEE ALSO THE FOLLOWING ARTICLES

STATISTICAL ROBUSTNESS • STATISTICS, BAYESIAN • STATISTICS, FOUNDATIONS • STATISTICS, MULTIVARIATE

BIBLIOGRAPHY

Arnold, S. F. (1980). "Asymptotic validity of F-tests for the ordinary linear model and the multiple correlation model," *J. Am. Stat. Assoc.* **75,** 890–894.

Box, G. E. P. (1953). "Non-normality and tests on variances," *Biometrika* **40,** 318–335.

Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981). "A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data," *Technometrics* **23,** 351–361.

Hájek, J., and Šidák, Z. (1967). "Theory of Rank Tests," Academic Press, New York.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. J. (1986). "Robust Statistics, the Approach Based on Influence Functions," Wiley, New York.

Hettmansperger, T. P. (1984). "Statistical Inference Based on Ranks," Wiley, New York.

Hettmansperger, T. P., and McKean, J. W. (1998). "Robust Nonparametric Statistical Methods," Arnold, London.

Hollander, M., and Wolfe, D. A. (1999). "Nonparametric Statistical Methods," Wiley, New York.

Huber, P. J. (1981). "Robust Statistics," Wiley, New York.

Jaeckel, L. A. (1972). "Estimating regression coefficients by minimizing the dispersion of the residuals," *Ann. Math. Stat.* **43,** 1449–1458.

Kalbfleisch, J. D., and Prentice, R. L. (1980). "The Statistical Analysis of Failure Time Data," Wiley, New York.

Koul, H. L. (1992). "Weighted Empiricals and Linear Models," Institute of Mathematical Statistics, Hayward, CA.

Lehmann, E. L. (1975). "Nonparametrics: Statistical Methods Based on Ranks," Holden-Day, San Francisco.

Maritz, J. S. (1995). "Distribution-Free Statistical Methods," 2nd edition, Chapman and Hall, London.

McKean, J. W., and Hettmansperger, T. P. (1976). "Tests of hypotheses of the general linear model based on ranks," *Commun. Stat. A Theory Methods* **5,** 693–709.

McKean, J. W., and Hettmansperger, T. P. (1978). "A robust analysis of the general linear model based on one step R-estimates," *Biometrika* **65,** 571–579.

McKean, J. W., and Sheather, S. J. (1991). Small sample properties of robust analyses of linear models based on R-estimates. *In* "A Survey. in Directions in Robust Statistics and Diagnostics," Part II (W. Stahel and S. Weisberg, eds.), pp. 1–20, New York: Springer-Verlag.

Minitab, Inc. (1988). "Minitab Reference Manual," Author, State College, PA.

Puri, M. L., and Sen, P. K. (1971). "Nonparametric Methods in Multivariate Analysis," Wiley, New York.

Puri, M. L., and Sen, P. K. (1985). "Nonparametric Methods in General Linear Models," Wiley, New York.

Randles, R. H., and Wolfe, D. A. (1979). "Introduction to the Theory of Nonparametric Statistics," Wiley, New York.

Rousseeuw, P., and Van Driessen, K. (1999). "A fast algorithm for the minimum covariance determinant estimator," *Technometrics* **41,** 212–223.

Staudte, R. G., and Sheather, S. J. (1990). "Robust Estimation and Testing," Wiley, New York.



Stefan Problems

A. D. Solomon
Negev Academic College of Engineering

I. A Classical Stefan Problem
II. History of the Problem
III. Recent Work
IV. Future Activities

GLOSSARY

Ablation Melting process in which liquid formed is driven away by flow conditions at the surface, a key example being an ablating surface of a rocket during its passage through air.
Interface Surface separating the solid and liquid phases of a material undergoing melting or freezing.
Interface condition Condition, usually representing energy conservation, holding at the interface between liquid and solid.
Latent heat of melting Difference of specific internal energies of solid and liquid at the melt temperature.
Marangoni convection Convection of a fluid driven by the surface tension effects of any interfaces in the fluid. Its effect is dominant in microgravity when gravity-driven natural convection becomes ignorably small.
Melt temperature Temperature for equilibrium between liquid and solid states.
Moving boundary problem Mathematical problem of finding a function and a domain in which the function satisfies a differential equation and the domain is partly described by interface conditions holding at its bounding surface.
Mushy region Region consisting of slush (liquid mixed with fine solid particles, uniformly at the melt temperature).
Supercooled liquid Liquid at a temperature below its melt temperature.

STEFAN PROBLEMS are a class of mathematical problems arising from macroscopic models of melting and freezing. They are the mathematical models for simple phase change processes. In this article, we will describe their form by means of a specific example, their history, recent developments, and anticipated future work.

I. A CLASSICAL STEFAN PROBLEM

A. Physical Setting

The Stefan problem arises from a mathematical model of melting and freezing of materials. In their solid phase, the materials are assumed to have an ordered crystalline structure maintained by intermolecular bonds. Liquid, on the other hand, is marked by the molecules essentially maintaining their close contact, but no longer having the crystalline structure of the solid. The energy associated with this structure is the latent heat of the material (in, for example, joules per gram), and it must be added to the solid in order to melt the solid and turn it into liquid.

Melting and freezing occur at a temperature, the melt temperature, characteristic of the material. At atmospheric

pressure, water, for example, melts and freezes at 0°C.

Consider a large block of material in its solid phase and at its melt temperature everywhere. Assume that all but one of the faces of the block are insulated. At this face let us suddenly impose a temperature above the melting point. Instantaneously, we will see liquid form on the heated face. If the block is contained in a thin-walled container, for example, then the liquid will remain in the container and it will constitute an ever-growing layer between the heated container wall and the solid beyond. If the block is not in any container, then the liquid will flow away from the solid, whose melting will be driven by the imposed hot temperature at its surface. This latter process is known as ablation. In the former case the solid and liquid phases meet at an interface. The interface will be parallel to the heated face, and as time progresses, the melted region will continue to grow at the expense of the frozen region. Eventually, all of the solid will have melted and the container will be filled only with liquid.

In our experiment, the temperature of the solid remains at the melt temperature for all time: indeed, it could not decrease, since temperature declines only when heat is extracted; similarly, it could not increase without the material having melted. Thus, ice above 0°C becomes water. On the other hand, the temperature in the liquid will always rise and lie between the melt temperature and the imposed face temperature.

For our experiment of melting a block of material, the Stefan problem can be stated as follows:

- At all times after the moment when the face temperature was imposed, find the temperature at all points in the block together with the solid and liquid regions.

In the melting process the interfacial surface separating liquid and solid regions is always moving. Problems in which the boundary of the region of study is moving with time and is an unknown of the problem are called "moving boundary problems." The Stefan problem is such a moving boundary problem.

B. Mathematical Formulation

The mathematical formulation of the Stefan problem in its classical form consists in relationships between the variables of the melting process. These are the temperature distribution at all times, the location of the boundary between solid and liquid regions, the thermophysical properties of the material (solid and liquid conductivities, specific heats, etc.), and the imposed conditions (for example, the face temperature) at the exterior fixed boundary of the block. These relationships take the form of partial differential equations in the solid and liquid regions and additional conditions at the moving boundary expressing local energy conservation and temperature behavior at the interface of solid and liquid material. The resulting problem is extremely difficult, since while the temperature is varying, so is the region in which it is to be found. Thus, for our case of melting a block of material, the temperature remains constant in the solid region, while in the liquid, it must be found by solving an equation within a time-varying region whose determination is part of the problem.

To illustrate the mathematical formulation, consider the process of melting a slab of material, initially solid, occupying the interval $0 < x < \delta$. Suppose that the initial temperature of the material is the melt temperature T_{m}, and at an initial time $t = 0$, the temperature at the face $x = 0$ of the material is set for all time as the constant value $T_{\mathrm{L}} > T_{\mathrm{m}}$. As a result a melting process takes place resulting in a boundary, located at a point $x = X(t) > 0$, separating liquid to its left from solid at the melt temperature to its right. With passing time, the moving boundary $X(t)$ moves to the right until all the material is melted. Within the liquid region, the temperature $T(x, t)$ at a point x and time t obeys the heat equation

$$T_t(x, t) = \alpha T_{xx}(x, t),$$

with the subscripts denoting the corresponding partial derivatives. At the front $x = X(t)$ the temperature is identically equal to the melt temperature:

$$T(X(t), t) \equiv T_m, \qquad t > 0,$$

while for $x = 0$,

$$T(0, t) \equiv T_{\mathrm{L}}.$$

Conservation of energy at the moving boundary states that the rate at which energy arrives at the front from the left (by conduction) is equal to the rate at which heat is absorbed by the material as its heat of melting,

$$\rho L X'(t) = -k T_x(X(t), t), \qquad t > 0.$$

This condition is referred to as the Stefan condition. Here α is the thermal diffusivity of the liquid, L is the latent heat of melting, k is the liquid thermal conductivity, and ρ is the liquid density.

II. HISTORY OF THE PROBLEM

The first serious discussion of a melting/freezing problem was by G. Lame and B. Clapeyron in 1831 and focused on the scenario of our example. No solution to the problem was found, but they did state that the front depth from the

wall was proportional to the square root of the elapsed time measured from the imposition of the wall temperature.

In 1889, J. Stefan examined the process of freezing of the ground. We note that this problem is still one of intensive study, with importance in such areas as storage of heat, proper design of roadways, and the performance of heat exchangers for heat pumps. Stefan studied the problem posed by Lame and Clapeyron as well as the question of what would happen if solid and liquid blocks of the same material, the first at a temperature below the melt temperature, the second above, were brought rapidly into contact: would the solid melt, or would the liquid freeze? Each was solved in the sense that formulas for the temperature distribution and for the location of the moving boundary between phases were found for all times. We note that credit for the solution of the first problem is given to F. Neumann, who is said to have done this already in the 1860s but did not publish his result.

There are two major difficulties that must be overcome in order to find the location of solid and liquid regions and the temperature distribution as time evolves. The first is the varying nature of the regions: the second is the nonlinearity of the problem. Because of this, the linear techniques of applied mathematics of the late 19th and early 20th centuries, such as Fourier and Laplace transforms, were essentially inapplicable to the problem studied by Stefan. For this reason, no significant advances in the study or solution to problems of this type were made over a 40-year period following Stefan's work.

From the time of Stefan's work until the 1930s two kinds of activities were taking place that would bring the Stefan problem into sharper focus and render it more tractable. The first was a growing understanding of so-called generalized functions, first treated rigorously by D. Hilbert in the early 1900s. Unlike the strictly continuous and differentiable functions to which mathematical analysis had been limited before, these functions could be undefined everywhere but still, in some integrated sense, would have mathematical and physical meaning. The most famous of such functions is, of course, the delta function $\delta(x - x_0)$ representing a source of finite strength located at the single point x_0. The second development relevant to the Stefan problem was a growing tendency toward placing problems in an algorithmic setting, with a view toward computing solutions that would become a reality with the advent of computing machines.

Beginning in the 1930s, modest advances in the understanding and theoretical development of the Stefan problem began to be made. Existence, uniqueness, and continuous dependence on input parameters were proved for a number of standard problems in one space dimension. This work reached its fruition in the 1970s when for the case of one space dimension it was proved that the Stefan problem was mathematically well posed. Similarly, beginning with the space program in the 1950s, significant work was done on analytical and computational methods for melting and freezing processes, aimed at such applications as the use of waxes as thermal buffers for electronic equipment and the treatment of ablation problems. This work, performed in the engineering community, produced approximation techniques such as the Goodman and Megerlin methods that could provide accurate estimates of front location and temperature for one space dimension.

The numerical solution of Stefan problems began to be of great interest to researchers in the oil industry in the mid-1950s. Initial efforts produced specific schemes relevant to methods for driving oil to desired areas by pumping other fluids into the ground. In the 1960s, a number of researchers developed so-called weak methods based on the use of the enthalpy as the key variable (in place of temperature) for the numerical solution of the Stefan problem in any number of space dimensions for arbitrary boundary conditions. These methods continue to form the general-purpose numerical methodology for phase change problems, while such approaches as finite element methods are commonly used for specific processes of interest.

The theoretical treatment of Stefan problems in two and three dimensions remains essentially undone. Unlike one space dimension, two or more dimensions make possible almost unlimited degrees of discontinuity. Thus, for example, a sculptured ice unicorn melting in a pool of heated water can be reduced to any number of small chunks of ice at a later time.

III. RECENT WORK

In recent years, fueled by physical problems of interest and an exponentially expanding computing capability, the community of interest in the Stefan problem has expanded to include such far-flung areas as geophysics, food processing, medicine, metallurgy, and economics. This in turn has added challenging complexities to the form of the original problem explored by Stefan. We now describe some of them.

A. Mushy Zones

Suppose that a block of ice and a pool of water are both at the uniform temperature of 0°C and that the ice is ground into arbitrarily small chunks and thrown into the water. Since the chunks can be made as small as we wish, the resulting mixture, viewed on a normal length scale, is a slush that is neither solid nor liquid. In particular, its average enthalpy lies between that of solid and that of liquid at 0°C. We call such material mushy.

Mushy zones arise in a variety of situations. Certain materials solidify in such a way that the interface between solid and liquid is long and dendritic, with many treelike or columnar branches of arbitrarily small size. In this case, there is effectively a mushy zone between solid and liquid. Similarly, radioactive material is heated by its own decay heat. If a lump of solid radioactive metal is placed in an insulated container, the metal will be heated by its own internally generated heat until it reaches its normal melt temperature. Further heating will not affect its state (in principle) until a quantity of heat equal to the latent heat has been generated. Until this happens, the material is entirely mushy. In the same way, materials heated by electrical resistance, such as by welding or *in situ* vitrification, pass through a mushy stage when their enthalpy lies between that of solid and that of liquid at the melt temperature.

The theoretical treatment of mushy zones is to a large extent open. Each material solidifies in a manner characteristic of its crystal structure. Some salts, for example, exhibit very broad mushy zones under ordinary freezing while others do not; some have crystals that are long and columnar with liquid trapped in between; others have filmy dendritic fronts, again with liquid trapped in between. The use of generalized functions has enabled us to obtain general existence results even in the presence of possible mushy zones, but more precise results may well require a better understanding of the materials involved and the material-dependent aspects of the mushy zone phenomenon.

B. Supercooled Solidification

If we slowly cool a liquid in a very smooth container under very quiet conditions, we may observe that the temperature of the liquid goes below the normal melt temperature for a period of time; it then rises to the melt temperature and the liquid freezes rapidly. A liquid whose temperature is below its normal melting point is said to be supercooled. On thermodynamic grounds, a material in the liquid state but at a temperature below the normal melt temperature is unstable, and given any kind of physical or thermal perturbation, will quickly freeze. Water can be supercooled to temperatures as low as -40°C, while other materials, such as hydrated salts, can be cooled to temperatures much further below their normal melt temperature.

Numerical methods have been developed for simulating the solidification of supercooled material based on the weak solution approach.

C. Natural Convection

The density of a liquid depends on its temperature, generally decreasing with increasing temperature. Thus, any temperature gradient in the vertical direction with temperature increasing with increasing depth usually will induce a flow, or natural convection, much like that seen when heating a thick soup. Thus, the liquid formed in the course of melting a solid from the vertical side or from below will ordinarily begin to flow through natural convection, greatly enhancing heat transfer and altering the form of the relations for heat transfer in the liquid phase. The development of effective numerical methods for simulating this effect in three-dimensional space is an active area of research. In the casting of special metals and alloy systems (see Section III.D), natural convection may be a highly undesired effect. For this reason, there is a long-standing effort to explore casting and metal processing in space, where the effects of gravity are not present. However, in the microgravity of space, convection driven by surface tension, a form of natural convection known as Marangoni convection, occurs at the solid/liquid interface.

D. Alloy Solidification

In recent years, high technology in both civilian and military spheres has pushed us to develop alloys of ever-increasing precise constitution. For example, our needs for special sensors and photovoltaics have resulted in a need for very precisely structured alloy systems in their solid phase.

An alloy is a material formed by the bonding of its constituent elements and differing from both. Thus, a binary alloy of copper and nickel is defined by the relative amounts of copper and nickel present, and for each such composition it is a different material in terms of its thermophysical properties. Moreover, the liquid and solid phases of an alloy have an additional interesting property related to their coexistence. Let us briefly describe this by considering a pure material such as water.

A flat piece of ice and an adjacent body of water can coexist, that is, the ice will remain ice and the water will remain water if the temperature at the interface between them is 0°C. We say that this is a state of thermodynamic equilibrium of the two materials. In contrast, on the same thermodynamic grounds, a solid piece of an alloy with one composition is not generally in thermodynamic equilibrium with a liquid alloy of the same composition. Rather, for each composition of solid, there will be a necessarily different composition needed by a liquid to be in equilibrium with the solid. The implication of this fact for the Stefan problem is clear. When a liquid alloy solidifies, the liquid alloy adjacent to the moving interface must be of a different composition from the solid on the other side of the interface. If not, the system is unstable, as in the case of a supercooled liquid. However, in order that the composition of adjacent solid and liquid be

different there must be diffusion of the component materials, and the diffusion process must be as rapid as that of the heat transfer. Since this is generally not possible, solidification processes for alloy systems occur in the presence of a phenomenon known as constitutional supercooling, where in place of a sharp interfacial surface between solid and liquid we find a generally thick mushy zone of intermixed solid and liquid. Knowing how to adequately model this phenomenon thermodynamically, analytically, and numerically requires significant additional effort, both theoretical and numerical.

E. Ultrarapid Solidification and Melting Processes

In recent years, the tools of thermal processes have evolved from flames at relatively low temperatures and propagation speeds to lasers, which are capable of inputting large amounts of energy in time spans that can be measured in picoseconds. On mathematical grounds, this creates difficulties for the heat equation as a model of heat transfer. The heat equation has the mathematical property of transmitting thermal signals at infinite speed, an impossibility because, of course, they cannot move faster than the speed of light. This is no problem for thermal processes that have been used up to present times, for the predicted amount of heat moving at unrealistically high speeds is ignorably small. This does become a potential problem for such thermal processes as laser pulsing, and will present an ever-larger problem as mechanisms for heat transfer develop further. With this in mind, alternative mechanisms for modeling heat transfer have been developed (e.g., hyperbolic heat transfer), our understanding of which are correct is as yet poor. In addition, because rapid processes are often applied in melting situations, the correct formulation of a Stefan problem in terms of a more realistic heat transfer mechanism is of particular importance.

F. Voids

Density changes accompanying phase change processes can increase or reduce the volume of the material undergoing these changes. In the former case, one sees containers broken (such as water pipes in the winter). In the latter case, one observes containers that have buckled or the formation of voids. Voids are vapor bubbles of material formed as a result of volume reduction accompanying a phase change. For most materials, the density change upon freezing or melting is of the order of 10%. For some materials, such as lithium fluoride, they may be as great as 30%, with their formation inhibiting heat transfer and producing other problems as well.

G. Control

Since melting and freezing processes are of vital importance, their control is of great interest. A typical control situation involves "backtracking" from moving boundary location to boundary conditions (an inverse problem). Research in this area focuses largely on numerical procedures and inverse problems, and is largely open.

IV. FUTURE ACTIVITIES

In the next decades, the Stefan problem and the modeling of melting and freezing processes will be of ever-increasing interest. Future developments can be anticipated in the following areas.

A. Thermodynamics of Phase Change Processes

We will make significant progress in understanding the underlying mechanisms of freezing and melting for pure materials and alloy systems. We will also learn to better understand the thermodynamics of the interface. We note that the interface is itself an important factor in freezing and melting processes. The fact that energy is needed to maintain an interface results in a depression of the equilibrium temperature between a locally spherical crystal and adjacent liquid, a result due to Gibbs-Thomson. These efforts will encompass ultrarapid processes as well.

B. Control of Phase Change Processes

The automation of a broad variety of freezing processes with real-time control of thermal conditions at the boundary will take place in the coming decades under both gravity and microgravity conditions. The mathematics, software, and hardware needed for this control will be developed. One aspect of this capability will be in the area of latent heat thermal energy storage, that is, the storage of heat derived from intermittent sources, such as the sun, for use at a later time or in a different place. Other areas include food preservation, manufacturing of photovoltaics, and metal casting.

C. Large-Scale Simulation

The rapid development of distributed and massive computing capabilities will make it possible to perform simulations of large-scale, long-term phase change processes in three dimensions, including effects of natural convection in the melt, rapidly changing boundary conditions, and local surface effects.

D. Mathematical Analysis of the Stefan Problem

The dimensionality barrier of the analysis of the Stefan problem will be breached with well-posed results obtained for higher dimensional problems arising from pure heat transfer. In addition, significant advances will be made in coupled problems, where the effects of convection, composition, void formation, etc., are felt, and inverse problems for boundary conditions and thermophysical and diffusion parameters.

E. Testing of Simulation Software and Its Implications

The development of increasingly massive and sophisticated software products leads directly to the question of how to test this software. Thus, for example, a software package simulating loss of coolant accidents in nuclear reactors will need to include modules whose purpose is to simulate the melting of metals that come into contact with other molten metals or radioactive metals whose melting is induced by their own internally generated heat. Such modules are small parts of large codes, and the importance of the overall software package will demand their unquestioned correctness. The formulation of tests for these modules based on data input and output analysis is a major challenge because it encompasses all the material-related phenomena noted above (e.g., how to model ultrarapid melting) together with numerical analysis pitfalls (e.g., the consistency or Du-Fort Frankel-like schemes) as well as theoretical considerations (e.g., the proper formulation of inverse problems in which inputs are to be found on the basis of observed melt/freeze processes). We anticipate that the question of software testing will lead necessarily to the need for a more fundamental understanding of all aspects of the Stefan problem.

SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL THERMODYNAMICS • PHASE TRANSFORMATIONS, CRYSTALLOGRAPHIC ASPECTS

BIBLIOGRAPHY

Carslaw, H. S., and Jaeger, J. C. (1959). "Conduction of Heat in Solids," 2nd ed., Oxford University Press, London.

Cheng, K., and Seki, N. (eds.). (1991)."Freezing and Melting Heat Transfer in Engineering: Selected Topics on Ice-Water Systems and Welding and Casting Processes," Hemisphere Publishers, Washington, DC.

Goldman, N. L. (1997). "Inverse Stefan Problems," Kluwer, Dordrecht, The Netherlands.

Hill, J. M. (1987). "One-Dimensional Stefan Problems: An Introduction," Longman, London.

Meirmanov, A., and Crowley, A. (1992). "The Stefan Problem," De Gruyter, Berlin.

Rubinstein, L. I. (1971). "The Stefan Problem," American Mathematical Society, Providence, RI.

Sarler, B., Brebbia, C., and Power, H. (eds.). (1999). "Moving Boundaries V: Computational Modelling of Free and Moving Boundary Problems," WIT Publishers, Wessex, England.

Solomon, A. D., Alexiades, V., and Wilson, D. G. (1992). "Mathematical Modeling of Melting and Freezing Processes," Hemisphere, New York.

Van Keer, R., and Brebbia, C. (eds.). (1997). "Moving Boundaries IV: Computational Modelling of Free and Moving Boundary Problems," Computational Mechanics, Wessex, England.

Wilson, D. G., Solomon, A. D., and Boggs, P. (eds.). (1978). "Moving Boundary Problems," Academic Press, New York.

Wrobel, L. C., and Brebbia, C. A. (eds.). (1993). "Computational Methods for Free and Moving Boundary Problems in Heat and Fluid Flow," Elsevier, Amsterdam.

Wrobel, L. C., Sarler, B., and Brebbia, C. (eds.). (1995). "Computational Modelling of Free and Moving Boundary Problems III," Computational Mechanics, Wessex, England.

Zerroukat, M., and Chatwin, C. (1994). "Computational Moving Boundary Problems," Research Studies Press, London.



Stochastic Processes

Yûichirô Kakihara
California State University, San Bernardino

I. Basics of Stochastic Processes
II. Statistics of Stochastic Processes
III. Some Examples of Stochastic Processes
IV. Weakly Stationary Stochastic Processes
V. Classes of Nonstationary Stochastic Processes
VI. The Calculus of Stochastic Processes
VII. Expansions of Stochastic Processes
VIII. Extensions of Stochastic Processes

GLOSSARY

Probability space A triple consisting of a sample space Ω, a certain collection Σ of subsets of Ω, and a probability P defined on Σ.
Random variable A numerical function defined on the probability space.
Sample space The set of all possible outcomes of an experiment.
Stochastic process A collection of random variables often indexed by time.

IN AN EMPERICAL SENSE, a stochastic process is considered as the description of a random phenomenon evolving in time that is governed by certain laws of probability. Stochastic processes model fluctuations in economic behavior or stock markets, the path of a particle in a liquid, outputs of physical systems, and, in fact, most phenomena exhibiting unpredictable fluctuations. A mathematical abstraction of a stochastic process is any indexed collection of random variables defined on a fixed probability space. Probability theory is a branch of mathematics and the theory of stochastic processes is a part of it. Modern probability theory is based on Kolmogorov's axiomatic treatment using measure theory.

I. BASICS OF STOCHASTIC PROCESSES

The concept of a stochastic process is obtained as a collection of random variables, usually indexed by a time parameter. A random variable is a real- (or complex-) valued function defined on a probability space. Hence the notion of a stochastic process is a generalization of the idea of a random variable.

First we define a probability space according to Kolmogorov's axiomatic formulation. Thus a *probability space* consists of a triple (Ω, Σ, P), where Ω is a *sample space,* Σ is a σ*-algebra* of events, and P is a *probability* on Σ. Each $\omega \in \Omega$ represents an outcome of some experiment and is called a *basic event.* Each $A \in \Sigma$ is a

subset of Ω, called an *event*. Σ satisfies the following conditions:

(i) $\Omega \in \Sigma$.
(ii) $A \in \Sigma$ implies $A^c \in \Sigma$, A^c being the complement.
(iii) If $A_1, A_2, \ldots \in \Sigma$, then $\cup_{n=1}^{\infty} A_n \in \Sigma$.

P satisfies the following:

(iv) $0 \le P(A) \le 1$ for any $A \in \Sigma$.
(v) $P(\Omega) = 1$.
(vi) If $A_1, A_2, \ldots \in \Sigma$ are disjoint, then $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

In other words, the σ-algebra Σ is a collection of subsets of the sample space Ω that contains the entire event Ω and is closed under complementation and countable union. We also can say that the probabilty P is a measure taking values in [0, 1] and countably additive.

Consider an experiment of rolling a die. The sample space Ω consists of six numbers 1, 2, 3, 4, 5, and 6. The σ-algebra Σ is then the set of all subsets of Ω, consisting of $2^6 = 64$ subsets. Probability is assigned 1/6 to each $\omega = \{1\}, \ldots, \{6\}$ if the die is fair. If A is an event that the die faces up with a number less than or equal to 4, then $P(A) = 2/3$. A mapping $X: \Omega \to \mathbb{R} = (-\infty, \infty)$ is a (real) *random variable* if X is measurable, i.e., for any $x \in \mathbb{R}$,

$$\{\omega \in \Omega: X(\omega) < x\} \equiv \{X < x\} \in \Sigma. \tag{1}$$

If we denote the Borel σ-algebra of $\mathbb{R}$ by $\mathfrak{B} = \mathfrak{B}(\mathbb{R})$, then (1) is equivalent to

$$\{\omega \in \Omega: X(\omega) \in A\} \equiv \{X \in A\} \in \Sigma \tag{2}$$

for every $A \in \mathfrak{B}$. [Similarly, a complex random variable is defined as in (2) using the Borel σ-algebra $\mathfrak{B}(\mathbb{C})$ of the complex number field $\mathbb{C}$.] In this section, we shall consider real random variables unless otherwise stated. The *expectation* of a random variable X is defined by the integral of X over Ω if $E\{|X|\} < \infty$, where

$$E\{X\} = \int_{\Omega} X(\omega)\, P(d\omega).$$

Then the *distribution* of X is obtained as a probability on $(\mathbb{R}, \mathfrak{B})$:

$$P_X(A) = P(\{X \in A\}), \qquad A \in \mathfrak{B}.$$

In many cases, the distribution of a random variable determines many of its properties. The *distribution function* F_X of a random variable X is obtained as

$$F_X(x) = P_X((-\infty, x)), \qquad x \in \mathbb{R}.$$

Using the distribution function, we can write the expectation of X as the Stieltjes integral

$$E\{X\} = \int_{-\infty}^{\infty} x\, dF_X(x).$$

In a special case, there exists a *density function* p_X such that

$$F_X(x) = \int_{-\infty}^{x} p_X(r)\, dr, \qquad x \in \mathbb{R},$$

so that $p_X(x) = dF_X(x)/dx$. If

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}, \qquad x \in \mathbb{R}, \tag{3}$$

where $\sigma > 0$ and $m \in \mathbb{R}$, then p is called a *Gaussian* (or *normal*) density function and X is called a *Gaussian* (or *normal*) random variable. The probability distribution given by (3) is denoted as $N(m, \sigma^2)$. Random variables $X_1, \ldots, X_n$ are said to be *mutually independent* if their distribution functions satisfy

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n) \tag{4}$$

for any $x_1, \ldots, x_n \in \mathbb{R}$, where $F_{X_1,\ldots,X_n}$ is called the *joint distribution function* of $X_1, \ldots, X_n$ defined by

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(\{X_1 < x_1, \ldots, X_n < x_n\}).$$

If the density functions exist, then (4) is equivalent to

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n),$$

where $p_{X_1,\ldots,X_n}$ is the density function of $F_{X_1,\ldots,X_n}$. The *characteristic function* of a random variable X is defined by

$$\Phi_X(\lambda) = E\{e^{i\lambda X}\}, \qquad \lambda \in \mathbb{R}.$$

A *stochastic process* is a collection of random variables on the same probability space indexed by a subset T of real numbers. In practice, each $t \in T$ represents a time. We usually denote a stochastic process on T by $\{X(t)\}_{t \in T}$ or simply by $\{X(t)\}$, where the underlying probability space (Ω, Σ, P) is fixed. Thus, for each $t \in T$, $X(t)$ is a random variable on (Ω, Σ, P) and its value at $\omega \in \Omega$ is denoted by $X(t, \omega)$. For each $\omega \in \Omega$, $X(\cdot, \omega)$ represents a function on T and is called a *sample path* or *realization* or *trajectory*. If T is a closed interval such as $\mathbb{R}$, $\mathbb{R}^+ = [0, \infty,)$, or $[0, 1]$, then we consider *continuous time* stochastic processes. If T is a discrete set such as $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$ or $\mathbb{Z}^+ = \{0, 1, 2, \ldots\}$, then $\{X(t)\}$ is called a *discrete time* stochastic process or a *time series*.

Let $\{X(t)\}$ be a stochastic process on $T \subseteq \mathbb{R}$. Then, for $t_1, \ldots, t_n \in T$, the joint distribution function of $X(t_1), \ldots, X(t_n)$ is denoted by $F_{t_1,\ldots,t_n}(x_1, \ldots, x_n)$ and the joint density function by $p_{t_1,\ldots,t_n}(x_1, \ldots, x_n)$, if it exists, so that

$$F_{t_1,\ldots,t_n}(x_1,\ldots,x_n) = \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n} p_{t_1,\ldots,t_n}(r_1,\ldots,r_n)\,dr_1\cdots dr_n.$$

Then, distribution functions should satisfy the following conditions:

$$p_{t_1}(x_1) = \int_{-\infty}^{\infty} p_{t_1,t_2}(x_1,x_2)\,dx_2,$$

$$p_{t_1,\ldots,t_m}(x_1,\ldots,x_m) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} p_{t_1,\ldots,t_n}(x_1,\ldots,x_n)\,dx_{m+1}\cdots dx_n \tag{5}$$

for $m<n$, etc. This means that, if we know the higher joint distribution functions, then the lower joint distribution functions are obtained as marginal distribution functions. *Kolmogorov's* (1933) *consistency theorem* states that, if we are given a system of joint density functions $\{p_{t_1,\ldots,t_n}(x_1,\ldots,x_n)\}$ for which (5) is true and it holds under permutations of time, then there exists a real stochastic process $\{X(t)\}$ whose joint density functions are exactly this system.

II. STATISTICS OF STOCHASTIC PROCESSES

Consider a stochastic process $\{X(t)\}$ on $T\subseteq\mathbb{R}$. The *expectation* of a stochastic process $\{X(t)\}$ is a function of $t\in T$ given by

$$m(t) = E\{X(t)\}.$$

For an integer $k\geq 1$, the *kth absolute moment* of $\{X(t)\}$ is defined by a function of t as

$$M_k(t) = E\{|X(t)|^k\}.$$

$m(t)$ is sometimes called an *ensemble average* at time t. Another type of average is obtained. The *time average* of $\{X(t)\}$ on $\mathbb{R}$ is defined by

$$\overline{X}(\omega) = \lim_{s\to\infty}\frac{1}{2s}\int_{-s}^{s} X(t,\omega)\,dt.$$

$\overline{X}(\omega)$ exists for $\omega\in\Omega$ such that $X(\cdot,\omega)$ is bounded and measurable on $\mathbb{R}$ or, in particular, $X(\cdot,\omega)$ is bounded and continuous on $\mathbb{R}$. Moreover, if this is true for all $\omega\in\Omega$, then $\overline{X}(\omega)$ defines a random variable. If $m(t)\equiv m=\overline{X}(\omega)$ P-a.e., then the process $\{X(t)\}$ is said to be *ergodic*. Thus, a stochastic process is ergodic if its ensemble average is equal to its time average. Here, P-a.e. refers to P-almost everywhere.

If $M_2(t)<\infty$ for all $t\in T$, then we say that $\{X(t)\}$ is a *second-order* stochastic process. The set of all complex random variables of finite second moment forms a Hilbert space $L^2(\Omega,\Sigma,P)=L^2(P)$, where the inner product and the norm are respectively defined by

$$(X,Y)_2 = E\{X\overline{Y}\},\qquad \|X\|_2 = E\{|X|^2\}^{1/2},$$

for $X,Y\in L^2(P)$, where $\overline{z}$ is the complex conjugate of $z\in\mathbb{C}$. The *mixed moment function* of $\{X(t)\}$ is defined by

$$R(s,t) = (X(s),X(t))_2,\qquad s,t\in T.$$

Similarly, the *covariance function* of $\{X(t)\}$ is defined by

$$K(s,t) = (X(s)-m(s),X(t)-m(t))_2,\qquad s,t\in T.$$

K (and also R) is a *positive-definite kernel* on $\mathbb{R}\times\mathbb{R}$, i.e.,

$$\sum_{j,k=1}^{n}\alpha_j\overline{\alpha}_k K(t_j,t_k)\geq 0$$

for any integer $n\geq 1$, $t_1,\ldots,t_n\in T$, and $\alpha_1,\ldots,\alpha_n\in\mathbb{C}$. Moreover, a Schwarz-type inequality holds:

$$|K(s,t)|^2\leq K(s,s)K(t,t),\qquad s,t\in T.$$

A stochastic process $\{X(t)\}$ is said to be *centered* if $m(t)\equiv 0$, in which case we have $K(s,t)=R(s,t)$. The *characteristic function* of $\{X(t)\}$ is defined by

$$\Phi_t(\lambda) = E\left\{e^{i\lambda X(t)}\right\} = \int_{-\infty}^{\infty} e^{i\lambda x}\,dF_t(x),$$

$$\Phi_{t_1,\ldots,t_n}(\lambda_1,\ldots,\lambda_n) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{i(\lambda_1x_1+\cdots+\lambda_nx_n)} \times dF_{t_1,\ldots,t_n}(x_1,\ldots,x_n).$$

It is easily seen that for any $\lambda,\lambda_1,\ldots,\lambda_n\in\mathbb{R}$

$$|\Phi_t(\lambda)|\leq\int_{-\infty}^{\infty} dF_t(x) = \Phi_t(0) = 1,$$

$$|\Phi_{t_1,\ldots,t_n}(\lambda_1,\ldots,\lambda_n)|\leq\Phi_{t_1,\ldots,t_n}(0,\ldots,0)=1.$$

III. SOME EXAMPLES OF STOCHASTIC PROCESSES

We now examine some special stochastic processes which are used to model physical phenomena practically.

A. Gaussian Processes

A real stochastic process $\{X(t)\}$ on $\mathbb{R}$ is said to be *Gaussian* if for every integer $n\geq 1$ its n-dimensional joint distribution is Gaussian, i.e., the density function $p_{t_1,\ldots,t_n}(x_1,\ldots,x_n)$ is of the form

$$p_{t_1,\ldots,t_n}(x_1,\ldots,x_n) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \times\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})V^{-1}(\mathbf{x}-\mathbf{m})^{\mathrm{t}}\right\},$$

where $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{m} = (m(t_1), \ldots, m(t_n)) \in \mathbb{R}^n$,

$$V = ((X(t_i) - m(t_i), X(t_j) - m(t_j))_2)_{i,j}$$

is an invertible $n \times n$ positive-definite Hermitian matrix, and $|V|$ is the determinant of V, with the superscript t being the transpose. One important characteristic of Gaussian processes is that if we are given a function $m(t)$ and a positive-definite function $K(s, t)$, then we can find a (complex) Gaussian process for which $m(t)$ is its expectation and $K(s, t)$ is its covariance function. As is seen from the Central Limit Theorem, the Gaussian distribution is universal and it is reasonable to use the Gaussian assumption on a stochastic process unless there is another specific choice.

B. Markov Processes

A real stochastic process $\{X(t)\}$ on $T = \mathbb{R}, \mathbb{R}^+, \mathbb{Z}$, or $\mathbb{Z}^+ = \{0, 1, 2, \ldots\}$ is said to be a *Markov process* if for any $s_1 < \cdots < s_n < t$ it holds that

$$P\big(X(t) \in A | X(s_1) = x_{s_1}, \ldots, X(s_n) = x_{s_n}\big) = P\big(X(t) \in A \mid X(s_n) = x_{s_n}\big)$$

P-a.e. for any $A \in \Sigma$, where $P(\cdot|\cdot)$ is the conditional probability. In this case, there exits a function $P(s, x, t, A)$, called the *transition probability,* for $s, x, t \in T$ and $A \in \Sigma$, such that the following hold:

(i) For fixed $s, t \in T$, $P(s, x, t, \cdot)$ is a probability for each $x \in T$, and $P(s, \cdot, t, A)$ is $\mathfrak{B}$-measurable for each $A \in \Sigma$.

(ii) $P(s, x, s, A) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$

(iii) $P(X(t) \in A \mid X(s) = x_s) = P(s, x_s, t, A)$, P-a.e.

(iv) (Chapman–Kolmogorovs equation) For fixed $s, u, t \in \mathbb{R}$

$$P(s, x, u, A) = \int_{-\infty}^{\infty} P(s, x, t, dy) P(t, y, u, A),$$

P_{X_s}-a.e.x.

Conversely, if a family of functions $\{P(s, x, t, A)\}$ is given such that (i), (ii), and (iv) hold, and a probability P_0 on $(\mathbb{R}, \mathfrak{B})$ is given, then there exists a Markov process $\{X(t)\}$ for which $P(s, x, t, A)$ is the transition probability and P_0 is the distribution of $X(0)$, i.e., the initial distribution.

There are some special Markov processes. A Markov process $\{X(t)\}$ is said to be *temporally homogeneous* if the transition probability depends on the difference of t and s, i.e., $P(s, x, t, A) = P(t - s, x, A)$. A Markov process $\{X(t)\}$ is said to be *spatially homogeneous* if $P(s, x, t, A) = P(s, t, A - x)$, where $A - x = \{y - x : y \in A\}$. A real stochastic process $\{X(t)\}$ on $\mathbb{R}^+$ is called an *additive process* if $X(0) = 0$ and for any $t_1, \ldots, t_n \in \mathbb{R}^+$ with $t_1 < \cdots < t_n$, $X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1})$ are mutually independent. As can be seen, each spatially homogeneous Markov process is an additive process. A stochastic process $\{X(t)\}$ on $\mathbb{R}^+$ is called a *diffusion process* if it is a Markov process such that the sample path $X(\cdot, \omega)$ is a continuous function on $\mathbb{R}^+$ for almost every $\omega \in \Omega$. If $\{X(t)\}$ is a Markov process and the range of $X(t, \cdot)$, $t \in T$, is a countable set, it is called a *Markov chain.* A *random walk* is a special case of a Markov chain.

C. Brownian Motion

The botanist R. Brown was observing pollens floating on the surface of water and discovered that these particles were moving irregularly all the time. If we consider a particle floating in the air or in the liquid as above, its motion caused by collisions with molecules of the air or the liquid is extremely irregular and regarded as a stochastic process. More precisely, if we denote the position of the particle at time t by $\mathbf{X}(t) = (X_1(t), X_2(t), X_3(t))$, then $\mathbf{X}(t)$ is a three-dimensional random variable such that $\mathbf{X}(t) - \mathbf{X}(s)$ obeys a three-dimensional Gaussian distribution. This kind of motions was considered as a class of stochastic processes and investigated by A. Einstein, N. Wiener, and P. Levy.

Here is a precise definition. A stochastic process $\{\mathbf{X}(t)\}$ on $T \subseteq \mathbb{R}$ is said to be a *Wiener process* or a *Brownian motion* if the following conditions hold:

(i) $\mathbf{X}(t, \omega) \in \mathbb{R}^d$ (d is a positive integer).

(ii) For $t_1 < \cdots < t_n$, $\mathbf{X}(t_2) - \mathbf{X}(t_1), \ldots, \mathbf{X}(t_n) - \mathbf{X}(t_{n-1})$ are mutually independent.

(iii) If $\mathbf{X}(t) = (X_1(t), \ldots, X_d(t))$, then $\{X_1(t)\}, \ldots, \{X_d(t)\}$ are mutually independent stochastic processes such that $X_i(t) - X_i(s) \approx N(0, |t - s|)$ for each i, i.e., $X_i(t) - X_i(s)$ obeys the Gaussian distribution $N(0, |t - s|)$.

It follows from the definition that a Wiener process is a temporally homogeneous additive process. Moreover, a Wiener process is a special case of a diffusion process since $S = \mathbb{R}^d$ is the state space [i.e., the range of the random variables $X(t)$, $t \in \mathbb{R}$] and the transition probabilities are given by

$$P(t, \mathbf{x}, B) = \int_B \frac{1}{(2\pi t)^{d/2}} \exp\left\{-\frac{1}{2t}|\mathbf{x} - \mathbf{y}|^2\right\} d\mathbf{y},$$

where $t > 0$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $B \in \mathfrak{B}(\mathbb{R}^d)$, and $|\mathbf{x}| = \sqrt{x_1^2 + \cdots + x_d^2}$, the norm of $\mathbf{x} = (x_1, \ldots, x_d)$.

D. Stationary Processes

Stationarity is considered as an invariance under the time shift. There are two kinds of stationarity, weak and strong. A stochastic process $\{X(t)\}$ is said to be *strongly stationary* or *stationary in the strict sense* if the joint distribution is invariant under the time shift, i.e., for any $t, t_1, \dots, t_n \in T$ and $E \in \mathfrak{B}(\mathbb{C}^n)$, the Borel σ-algebra of $\mathbb{C}^n$, it holds that

$$P((X(t_1+t), \dots, X(t_n+t)) \in E) = P((X(t_1), \dots, X(t_n)) \in E).$$

A second-order stochastic process $\{X(t)\}$ is said to be *weakly stationary* or *stationary in the wide sense* if its average is constant, if its covariance function $K(s,t)$ depends only on the difference $s-t$, and if K is continuous as a two-variable function. Clearly, if the process is of second order and the covariance function is continuous, then strong stationarity implies weak stationarity. The converse is true for Gaussian processes. In general, a strongly stationary process need not have any moment (e.g., a Cauchy stationary process). Weak stationarity will be singled out and discussed in the next section.

E. White Noise

A discrete or continuous time stochastic process $\{X(t)\}$ is called a *white noise* if $E\{X(t)\} \equiv 0$ and $K(s,t) = \delta(s-t)$, the Kronecker delta function, i.e., $= 1$ for $s = t$ and $= 0$ for $s \neq t$. A white noise does not exist in a real world. But as a mathematical model it is used in many fields. It is known that the derivative of a Wiener process as a generalized stochastic process is a white noise (cf. Section VIII).

IV. WEAKLY STATIONARY STOCHASTIC PROCESSES

Weak stationarity for second-order stochastic processes was initiated by Khintchine (1934). Let us begin with the definition again. A stochastic process $\{X(t)\}$ on $T = \mathbb{R}$ or $\mathbb{Z}$ is said to be *weakly stationary* or *wide sense stationary* if its expectation is constant and if the covariance function is continuous and depends only on the difference of the variables. In other words, $\{X(t)\}$ is weakly stationary if the following hold:

(i) $m(t) = E\{X(t)\} \equiv m$ for $t \in T$.
(ii) $K(s,t) = \tilde{K}(s-t)$ for $s, t \in T$.
(iii) $\tilde{K}(\cdot)$ is continuous on T.

If $T = \mathbb{Z}$, the condition (iii) is automatically satisfied. We may assume that $m = 0$ in (i) [since otherwise we can let $Y(t) = X(t) - m$] and consider processes on $\mathbb{R}$. Denote by $L_0^2(P)$ the closed subspace of $L^2(P)$ consisting of functions with zero expectations, i.e., $L_0^2(P) = \{X \in L^2(P): E\{X\} = 0\}$. So we consider centered processes $\{X(t)\} \subseteq L_0^2(P)$.

Let us proceed to obtain integral representations of a weakly stationary process and its covariance function, and the Kolmogorov isomorphism theorem, which states that the time and the spectral domains are isomorphic.

Thus, suppose $\{X(t)\}$ is a weakly stationary process on $\mathbb{R}$ with the (one-variable) covariance function $K(t)$. Since $K(t)$ is continuous and positive definite, there exists, by Bochner's theorem, a finite positive measure ν on $(\mathbb{R}, \mathfrak{B})$ such that

$$K(t) = \int_{-\infty}^{\infty} e^{itu}\, \nu(du), \qquad t \in \mathbb{R}. \tag{6}$$

ν is called the *spectral measure* of the process. The *time domain* $H(X)$ of $\{X(t)\}$ is defined by

$$H(X) = \mathfrak{S}\{X(t): t \in \mathbb{R}\},$$

the closed subspace of $L_0^2(P)$ spanned by the set $\{X(t): t \in \mathbb{R}\}$. To derive an integral representation of the process itself, define an operator $U(t)$ by

$$U(t)X(s) = X(s+t), \qquad s \in \mathbb{R},$$

or more generally by

$$U(t)\left(\sum_{k=1}^{n} \alpha_k X(t_k)\right) = \sum_{k=1}^{n} \alpha_k X(t_k + t),$$

where $\alpha_1, \dots, \alpha_n \in \mathbb{C}$, $t_1, \dots, t_n \in \mathbb{R}$. Then because of stationarity we see that $U(t)$ is well defined and norm preserving, and can be extended to a unitary operator on the time domain $H(X)$. Hence, by Stone's theorem, there is a spectral measure (i.e., an orthogonal projection-valued measure) E on $(\mathbb{R}, \mathfrak{B})$ such that

$$U(t) = \int_{-\infty}^{\infty} e^{itu}\, E(du), \qquad t \in \mathbb{R}.$$

Consequently we have that

$$X(t) = U(t)X(0) = \int_{-\infty}^{\infty} e^{itu}\, E(du)\, X(0), \qquad t \in \mathbb{R}.$$

If we let $\xi(A) = E(A)X(0)$ for $A \in \mathfrak{B}$, then the random quantity ξ is an $L_0^2(P)$-valued, bounded, countably additive measure, and is *orthogonally scattered*, i.e., $(\xi(A), \xi(B))_2 = 0$ if $A \cap B = \emptyset$. Therefore, we get an integral representation of the process $\{X(t)\}$ as

$$X(t) = \int_{-\infty}^{\infty} e^{itu}\, \xi(du), \qquad t \in \mathbb{R}. \tag{7}$$

By (6) and (7) we see that $\nu(A) = (\xi(A), \xi(A))_2$ for $A \in \mathfrak{B}$.

Consider the Hilbert space $L^2(\nu) = L^2(\mathbb{R}, \mathfrak{B}, \nu)$, which is called the *spectral domain* of the process $\{X(t)\}$. Then

the *Kolmogorov isomorphism theorem* states that the time domain $H(X)$ and the spectral domain $L^2(\nu)$ are isomorphic as Hilbert spaces. Moreover, the isomorphism $V: L^2(\nu) \to H(X)$ is given by

$$V(f) = \int_{-\infty}^{\infty} f(u)\,\xi(du), \qquad f \in L^2(\nu). \tag{8}$$

The importance of this theorem is that each random variable in $H(X) \subseteq L_0^2(P)$ is expressed as a vector integral of a usual function and, especially, $X(t)$ is a Fourier transform of a vector measure ξ. Furthermore, the measure ξ can be obtained from the process $X(t)$ by inversion as

$$\xi((u, v)) = \lim_{t\to\infty} \int_{-t}^{t} \frac{e^{-ivs} - e^{-ius}}{-is} X(s)\,ds, \tag{9}$$

where $u < v$, $\xi(\{u\}) = \xi(\{v\}) = 0$, and the right-hand side is a Bochner integral (cf. Section VI). Thus (9) is called the *inversion formula.*

V. CLASSES OF NONSTATIONARY STOCHASTIC PROCESSES

Stationarity in the strict sense or wide sense is very restrictive in a practical application and somewhat weaker conditions are needed in many cases. For the sake of simplicity we consider only centered second-order stochastic processes and use the space $L_0^2(P)$. The following are classes of nonstationarity.

A. Karhunen Class

Karhunen (1947) introduced a class of stochastic processes given by

$$X(t) = \int_{-\infty}^{\infty} g_t(u)\,\xi(du), \qquad t \in \mathbb{R}, \tag{10}$$

where ξ is an $L_0^2(P)$-valued orthogonally scattered measure and $\{g_t: t \in \mathbb{R}\} \subseteq L^2(\nu)$ with $\nu(\cdot) = \|\xi(\cdot)\|_2^2$. In this case the covariance function is given by

$$K(s, t) = \int_{-\infty}^{\infty} g_s(u)\overline{g_t(u)}\,\nu(du), \qquad s, t \in \mathbb{R}.$$

When $g_t(u) = e^{itu} (t \in \mathbb{R})$, $\{X(t)\}$ reduces to a weakly stationary process.

B. Harmonizable Class

Harmonizability was first introduced by Loève in the middle 1940s. Later a weaker harmonizability was defined by Rozanov (1959). These two notions were distinguished by calling them weak and strong harmonizabilities by Rao (1982).

A second-order stochastic process $\{X(t)\}$ on $\mathbb{R}$ is said to be *strongly harmonizable* if its covariance function K has a representation

$$K(s, t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i(su-tv)}\,\beta(du, dv), \qquad s, t \in \mathbb{R}, \tag{11}$$

for some *positive-definite bimeasure* $\beta: \mathfrak{B} \times \mathfrak{B} \to \mathbb{C}$ of bounded variation, where $\mathfrak{B} \times \mathfrak{B} = \{A \times B: A, B \in \mathfrak{B}\}$ and the total *variation* (= *Vitali variation*) is given by

$$|\beta|(\mathbb{R}, \mathbb{R}) = \sup \sum_{j=1}^{\ell} \sum_{k=1}^{n} |\beta(A_j, B_k)|,$$

the supremum being taken for all finite measurable partitions $\{A_1, \ldots, A_\ell\}$ and $\{B_1, \ldots, B_n\}$ of $\mathbb{R}$. Here, $\beta: \mathfrak{B} \times \mathfrak{B} \to \mathbb{C}$ is a *bimeasure* if $\beta(A, \cdot)$ and $\beta(\cdot, A)$ are $\mathbb{C}$-valued measures on $\mathfrak{B}$ for each $A \in \mathfrak{B}$. When β is of bounded variation, i.e., $|\beta|(\mathbb{R}, \mathbb{R}) < \infty$, β can be extended to an ordinary measure on $\mathfrak{B}(\mathbb{R}^2)$ (the Borel σ-algebra of $\mathbb{R}^2$) and (11) becomes the Lebesgue integral. When β is of unbounded variation, we need to interpret (11) as a bimeasure integral or (a weaker) MT-integral (after Morse and Transue). Since every bimeasure satisfies $\sup\{|\beta(A, B)|: A, B \in \mathfrak{B}\} < \infty$, we can show that (11) is well defined. So, $\{X(t)\}$ is said to be *weakly harmonizable* if its covariance function K is expressed as (11) for some positive-definite bimeasure β. Since the integrand $e^{i(su-tv)} = \varphi_1(u)\varphi_2(v)$, say, in the integral (11) is the product of bounded continuous functions of one variable, we can interpret the MT-integral as follows. For $A, B \in \mathfrak{B}$, φ_1 is $\beta(\cdot, B)$-integrable and φ_2 is $\beta(A, \cdot)$-integrable. Hence, letting

$$\mu_1(A) = \int_{-\infty}^{\infty} \varphi_2(v)\,\beta(A, dv), \qquad A \in \mathfrak{B},$$

$$\mu_2(B) = \int_{-\infty}^{\infty} \varphi_1(u)\,\beta(du, B), \qquad B \in \mathfrak{B},$$

we see that μ_1 and μ_2 are $\mathbb{C}$-valued measures on $\mathfrak{B}$, and φ_j is μ_j-integrable $(j = 1, 2)$. Moreover, when they have the same integral, i.e.,

$$\int_{-\infty}^{\infty} \varphi_1(u)\,\mu_1(du) = \int_{-\infty}^{\infty} \varphi_2(v)\,\mu_2(dv),$$

we denote the value by $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi_1(u)\varphi_2(v)\,\beta(du, dv)$.

An integral representation of a weakly harmonizable process $\{X(t)\}$ is obtained:

$$X(t) = \int_{-\infty}^{\infty} e^{itu}\,\xi(du), \qquad t \in \mathbb{R}, \tag{12}$$

for some $L_0^2(P)$-valued bounded measure ξ, not necessarily orthogonally scattered. The measure ξ is called the

representing measure of $\{X(t)\}$ and the inversion formula (9) is also valid. Furthermore, comparing (11) and (12), we see that

$$\beta(A, B) = (\xi(A), \xi(B))_2, \qquad A, B \in \mathfrak{B}.$$

As is easily seen, a weakly harmonizable process is weakly stationary if and only if the representing measure is orthogonally scattered. If a weakly stationary process is projected onto a closed subspace of its time domain, then the resulting process is not necessarily weakly stationary but always weakly harmonizable. The converse is known as a *stationary dilation.* That is, each weakly harmonizable process can be obtained as an orthogonal projection of some weakly stationary process.

The Kolmogorov isomorphism theorem holds for a weakly harmonizable process $\{X(t)\}$. The spectral domain of $\{X(t)\}$ is defined to be the space $\mathcal{L}^2_*(\beta)$ of all functions $f(u)$ on $\mathbb{R}$ that are *strictly integrable* with respect to the bimeasure β, in a somewhat "restricted sense." Then, the time domain $H(X)$ and the spectral domain $\mathcal{L}^2_*(\beta)$ are isomorphic Hilbert spaces, where the isomorphism is given by

$$V(f) = \int_{-\infty}^{\infty} f(u)\,\xi(du), \qquad f \in \mathcal{L}^2_*(\beta),$$

which is similar to (8). The weakly harmonizable class is a fruitful area of study with many applications because each process in this class is a Fourier transform of a vector measure and is relatively simple to handle.

C. Cramér Class

A further generalization of the Karhunen class was introduced by Cramér (1951). A stochastic process $\{X(t)\}$ is said to be *of strong Cramér class* if its covariance function is written as

$$K(s,t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_s(u)\overline{g_t(v)}\,\beta(du, dv), \qquad s, t \in \mathbb{R}, \tag{13}$$

for some positive-definite bimeasure β of bounded variation and some family $\{g_t : t \in \mathbb{R}\}$ of bounded Borel functions on $\mathbb{R}$. When the bimeasure β in (13) is simply bounded, $\{X(t)\}$ is *of weak Cramér class* relative to the family $\{g_t : t \in \mathbb{R}\}$. Integral representations of such processes are available. If $\{X(t)\}$ is of weakly or strongly Cramér class, then there exists an $L^2_0(P)$-valued measure ξ such that (10) is true. As a harmonizable process has a stationary dilation, each process of Cramér class has a *Karhunen dilation,* i.e., it is expressed as an orthogonal projection of some process of Karhunen class.

D. KF Class

The concept of the spectral measure can be generalized. In the late 1950s, Kampe de Fériet, and Frankiel, and independently Parzen and Rozanov, defined the "associated spectrum" for a process $\{X(t)\}$ for which

$$\tilde{K}(h) = \lim_{t\to\infty} \frac{1}{t}\int_0^t K(s, s+h)\,ds, \qquad h \in \mathbb{R},$$

exists and is positive definite and continuous in h, where K is the covariance function of the process. In this case, by Bochner's theorem there exists a finite positive measure ν on $(\mathbb{R}, \mathfrak{B})$, called the *associated spectrum* of $\{X(t)\}$, such that

$$\tilde{K}(h) = \int_{-\infty}^{\infty} e^{ihu}\,\nu(du), \qquad h \in \mathbb{R}.$$

Such a process is said to be *of KF class* or an *asymptotically stationary* process. Unfortunately, not every process of KF class allows an integral representation of the process itself. A strongly harmonizable process is necessarily of KF class, but a weakly harmonizable process is not in general.

E. Periodically Correlated Class

The number of black spots in the surface of the Sun increases periodically every 11 years. To describe such a phenomenon we use a periodically correlated process. Here, we consider only discrete time processes, namely those on $\mathbb{Z}$. A process $\{X(t)\}$ on $\mathbb{Z}$ is said to be *periodically correlated with period* $p > 0$ if the covariance function K satisfies

$$K(s,t) = K(s+p, t+p), \qquad s, t \in \mathbb{Z}.$$

If $p = 1$, then the process is weakly stationary. So we assume $p \geq 2$.

Consider the Cartesian product space $[L^2_0(P)]^p = L^2_0(P) \times \cdots \times L^2_0(P)$. For $\mathbf{X} = (X_1, \ldots, X_p)$ and $\mathbf{Y} = (Y_1, \ldots, Y_p) \in [L^2_0(P)]^p$ define

$$(\mathbf{X}, \mathbf{Y})_{2,p} = \sum_{j=1}^{p} (X_j, Y_j)_2,$$

so that $(\cdot,\cdot)_{2,p}$ is an inner product in $[L^2_0(P)]^p$. Also define a new process $\{\mathbf{X}(t)\}$, which is $[L^2_0(P)]^p$-valued, by

$$\mathbf{X}(t) = (X(t), X(t+1), \ldots, X(t+p-1)), \qquad t \in \mathbb{Z}.$$

Then we can verify that an $L^2_0(P)$-valued process $\{X(t)\}$ is periodically correlated with period p if and only if the $[L^2_0(P)]^p$-valued process $\{\mathbf{X}(t)\}$ is weakly stationary. Moreover, note that $L^2_0(P)$ can be regarded as a closed subspace of $[L^2_0(P)]^p$ if we identify $L^2_0(P)$ with $L^2_0(P) \times \{0\} \times \cdots \times \{0\}$. Hence, if $\{X(t)\}$ is periodically

correlated with period p and $J: [L_0^2(P)]^p \to L_0^2(P)$ is the orthogonal projection, then $X(t) = J\mathbf{X}(t)$, $t \in \mathbb{Z}$. Therefore, $\{X(t)\}$ is weakly harmonizable and $\{\mathbf{X}(t)\}$ is its stationary dilation. Furthermore, it is known that $\{X(t)\}$ is strongly harmonizable.

VI. THE CALCULUS OF STOCHASTIC PROCESSES

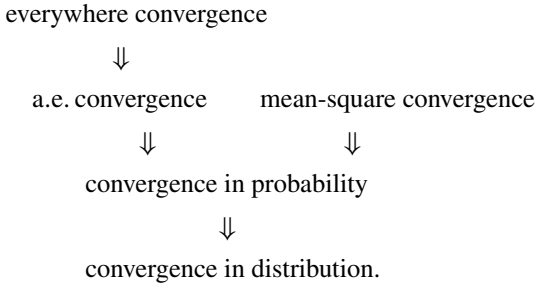
A. Convergence

In order to talk about continuity, differentiability, and integrability of stochastic processes on $\mathbb{R}$, we need to present the notion of convergence of a sequence of random variables. Let $\{X_n\}$ be a sequence of random variables and X a random variable on a probability space (Ω, Σ, P). We say that $\{X_n\}$ converges to X:

1. *Everywhere* if $X_n(\omega) \to X(\omega)$ for every $\omega \in \Omega$
2. *Almost everywhere* (a.e.), *almost surely* (a.s.), or *with probability* 1, if $P(\{\omega \in \Omega: X_n(\omega) \to X(\omega)\}) = 1$
3. *In mean-square* or *in L^2-mean* if $\|X_n - X\|_2 \to 0$
4. *In probability* if for any $\varepsilon > 0$, $P(\{|X_n - X| > \varepsilon\}) \to 0$
5. *In distribution* if $F_n(x) \to F(x)$ for each $x \in \mathbb{R}$ for which F is continuous, where F_n and F are distribution functions of X_n and X, respectively.

When the density functions exist, the last condition is equivalent to $p_n(x) \to p(x)$ for almost all $x \in \mathbb{R}$, where p, p_n are density functions.

The interrelationship among these convergences is as follows:



B. Continuity

Consider a continuous time second-order stochastic process $\{X(t, \omega)\} = \{X(t)\}$ on $T \subseteq \mathbb{R}$. If we choose an $\omega \in \Omega$, then $X(\cdot, \omega)$ is a function of $t \in T$ (a sample path) and its continuity is discussed. We say that $\{X(t)\}$ is *continuous* a.e. or has a *sample path continuity* a.e. if, for a.e. $\omega \in \Omega$, the scalar function $X(\cdot, \omega)$ is a continuous function on T.

Another type of continuity is in the mean-square sense. $\{X(t)\}$ is said to be *continuous in mean-square* at $t_0 \in T$ if

$$\|X(t) - X(t_0)\|_2 \to 0 \qquad \text{as} \quad t \to t_0,$$

which is equivalent to the continuity of the covariance function $K(s, t)$ at (t_0, t_0). Since

$$\begin{aligned} |E\{X(t)\} - E\{X(t_0)\}| &\le E\{|X(t) - X(t_0)|\} \\ &\le \left[E\left\{|X(t) - X(t_0)|^2\right\}\right]^{1/2} \\ &= \|X(t) - X(t_0)\|_2, \end{aligned}$$

we see that the expectation of $X(t)$ is continuous at t_0 if $\{X(t)\}$ is continuous in mean-square at t_0.

C. Differentiability

Let $\{X(t, \omega)\} = \{X(t)\}$ be a second-order stochastic process on an interval $T \subseteq \mathbb{R}$. First consider a sample path differentiability. $\{X(t)\}$ has *a.e. differentiable sample path* if, for a.e. $\omega \in \Omega$, the sample path $X(\cdot, \omega)$ is differentiable. Another type of derivative is in the mean-square sense. $\{X(t)\}$ is *differentiable in mean-square* at t_0 if there is a random variable $X'(t_0)$ such that

$$\left\| \frac{X(t_0 + h) - X(t_0)}{h} - X'(t_0) \right\|_2 \to 0 \qquad \text{as} \quad h \to 0.$$

If $\{X(t)\}$ on $\mathbb{R}$ is weakly stationary with the covariance function $K(t)$, then it is differentiable in mean-square at t_0 if and only if K has the second derivative at t_0. When $\{X(t)\}$ is nonstationary with the covariance function $K(s, t)$, it is differentiable in mean-square at t_0 if $\partial^2 K / \partial s \partial t$ exists at $s = t = t_0$.

D. Integrability

Consider a second-order stochastic process $\{X(t)\}$ on $[a, b]$ and its integral

$$\int_a^b X(t, \omega)\, dt. \tag{14}$$

If, for all $\omega \in \Omega$, the sample path $X(\cdot, \omega)$ is continuous or bounded and measurable, then the integral (14) is defined as a Riemann or Lebesgue integral and defines a random variable $Y(\omega)$. If this is not the case, however, we can consider (14) in a mean-square sense. Then, the integrand is an $L_0^2(P)$-valued function $X(t)$, the measure is the Lebesgue measure dt, and the integral (14) is a *Bochner integral.* The definition of this integral is as follows: Let

$$X_n(t) = \sum_{k=1}^{k_n} 1_{A_{n,k}}(t) X_{n,k},$$

where $X_{n,k} \in L_0^2(P)$, $A_{n,k} \in \mathfrak{B}$, and $1_{A_{n,k}}(t)$ is the indicator function of $A_{n,k}$, i.e., $= 1$ if $t \in A_{n,k}$ and $= 0$ otherwise. Such a function is called a *finitely valued* measurable function. If there exists a sequence $\{X_n(t)\}$ of such functions for which

$$\|X_n(t) - X(t)\|_2 \to 0 \qquad (n \to \infty)$$

for almost all $t \in [a, b]$ in the Lebesgue measure and $\int_a^b \|X(t)\|_2 \, dt < \infty$, then we can define the Bochner integral as

$$\int_a^b X(t)\, dt = \lim_{n\to\infty} \int_a^b X_n(t)\, dt \equiv \lim_{n\to\infty} \sum_{k=1}^{k_n} X_{n,k} L(A_{n,k}),$$

$L(A_{n,k})$ being the Lebesgue measure of $A_{n,k}$, where the limit is in mean-square sense.

VII. EXPANSIONS OF STOCHASTIC PROCESSES

Since a stochastic process is regarded as a two-variable function $X(t, \omega)$ on $T \times \Omega$, it might be interesting and useful to express it as a sum of products of single-variable functions of t and ω:

$$X(t, \omega) = \sum_n \phi_n(t) X_n(\omega), \tag{15}$$

where the right-hand side is convergent in some sense, so that $\phi_n(t)$ is a time function and $X_n(\omega)$ is a random variable for $n \geq 1$. We consider three types of expressions of stochastic processes given by (15).

A. Karhunen–Loève Expansion

Let $\{X(t)\}$ be a second-order stochastic process on a finite closed interval $[a, b]$ with the covariance function $K(s, t)$. Assume that the process $\{X(t)\}$ is mean-square continuous, which is the same as saying that $K(s, t)$ is continuous as a two-variable function on $[a, b]^2$. Then it follows that $K(s, t)$ is measurable on $[a, b]^2$ and

$$\int_a^b \int_a^b |K(s, t)|^2 \, ds\, dt < \infty,$$

i.e., $K \in L^2([a, b]^2)$. Moreover, K defines an integral operator $\mathbf{K}$ on $L^2([a, b])$ in such a way that

$$(\mathbf{K}\varphi)(t) = \int_a^b K(s, t)\varphi(s)\, ds \tag{16}$$

for $\varphi \in L^2([a, b])$. Now Mercer's theorem applied to $K(s, t)$ gives an expansion of $K(s, t)$:

$$K(s, t) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(s)\overline{\varphi_k(t)}, \qquad s, t \in [a, b], \tag{17}$$

where the φ_k are eigenfunctions of the operator $\mathbf{K}$ defined by (16) and the λ_k are corresponding eigenvalues:

$$(\mathbf{K}\varphi_k)(t) = \lambda_k \varphi_k(t), \qquad t \in [a, b], \tag{18}$$

and the convergence in (17) is uniform on $[a, b]^2$ and in $L^2([a, b]^2)$, i.e.,

$$\sup_{s,t\in[a,b]} \left| K(s, t) - \sum_{k=1}^{n} \lambda_k \varphi_k(s)\overline{\varphi_k(t)} \right| \to 0,$$

$$\left\| K - \sum_{k=1}^{n} \lambda_k \varphi_k \otimes \overline{\varphi_k} \right\|_2^2 \equiv \int_a^b \int_a^b \left| K(s, t) - \sum_{k=1}^{n} \lambda_k \varphi_k(s)\overline{\varphi_k(t)} \right|^2 ds\, dt \to 0$$

as $n \to \infty$, where $(\varphi_k \otimes \overline{\varphi_k})(s, t) = \varphi_k(s)\overline{\varphi_k(t)}$. Note that φ_k is continuous on $[a, b]$ and $\lambda_k > 0$ for each $k \geq 1$. For instance, if $K(s, t) = \min\{s, t\}$ on $[0, T]$, the covariance function of a one-dimensional Wiener process, then

$$\min\{s, t\} = \sum_{k=0}^{\infty} \frac{2}{T} \left(\frac{2T}{(2k+1)\pi} \right)^2 \sin\left(\frac{(2k+1)\pi}{2T} s \right) \times \sin\left(\frac{(2k+1)\pi}{2T} t \right),$$

the uniform convergence holding on $[0, T]^2$.

Now for each integer $k \geq 1$ define

$$X_k = \frac{1}{\sqrt{\lambda_k}} \int_a^b X(t)\overline{\varphi_k(t)}\, dt,$$

which is a well-defined random variable in $L^2(P)$. Here, the functions φ_k satisfy (18). Moreover, we see that $\{\varphi_k\}_{k=1}^{\infty}$ forms an orthonormal set in $L^2([a, b])$, i.e.,

$$(\varphi_j, \varphi_k)_2 = \int_a^b \varphi_j(t)\overline{\varphi_k(t)}\, dt = \delta_{jk},$$

the Kronecker delta, and that $\{X_k\}_{k=1}^{\infty}$ also forms an orthonormal set in $L^2(P)$. Finally we obtain an expansion of $X(t)$:

$$X(t) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} \varphi_k(t) X_k, \tag{19}$$

where the convergence is in mean-square. Equation (19) is called the *Karhunen–Loéve expansion.* The characteristic of this expansion is that we are using two orthonormal sets $\{\varphi_k\}$ and $\{X_k\}$, and that $\{X_k\}$ is the set of Fourier

coefficients of $X(t)$ with respect to $\{\varphi_k\}$. Moreover, $\{\varphi_k\}$ is obtained as the eigenfunctions of the operator $\mathbf{K}$ associated with the covariance function of the process.

B. Sampling Theorem

Shannon's (1949) sampling theorem was obtained for deterministic functions on $\mathbb{R}$ (signal functions). This was extended to be valid for weakly stationary stochastic processes by Balakrishnan (1957).

First we consider bandlimited and finite-energy signal functions. A signal function $X(t)$ is said to be *of finite energy* if $X \in L^2(\mathbb{R})$ with the Lebesgue measure. Then its Fourier transform $\mathcal{F}X$ is defined in the mean-square sense by

$$(\mathcal{F}X)(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} X(t) e^{iut}\, dt = \underset{T\to\infty}{\text{l.i.m.}} \int_{-T}^{T} X(t) e^{iut}\, dt,$$

where l.i.m. means "limit in the mean" and $\mathcal{F}: L^2(\mathbb{R}) \to L^2(\mathbb{R})$ turns out to be a unitary operator. A signal function $X(t)$ is said to be *bandlimited* if there exists some constant $W > 0$ such that

$$(\mathcal{F}X)(u) = 0 \qquad \text{for almost every } u \text{ with } |u| > W.$$

Here, W is called a *bandwidth* and $[-W, W]$ a *frequency interval.*

Let $W > 0$ and define a function S_W on $\mathbb{R}$ by

$$S_W(t) = \begin{cases} \frac{W}{\pi} \frac{\sin Wt}{Wt}, & t \neq 0 \\ \frac{W}{\pi}, & t = 0. \end{cases}$$

Then, S_W is a typical example of a bandlimited function and is called a *sample function.* In fact, one can verify that

$$(\mathcal{F}S_W)(u) = \frac{1}{\sqrt{2\pi}} 1_W(u), \qquad u \in \mathbb{R},$$

where $1_W = 1_{[-W,W]}$, the indicator function of $[-W, W]$. Denote by **BLW** the set of all bandlimited signal functions with bandwidth $W > 0$. Then it is easily seen that **BLW** is a closed subspace of $L^2(\mathbb{R})$. Now the *sampling theorem* for a function in **BLW** is stated as follows: Any $X \in$ **BLW** has a sampling expansion in L^2- and L^∞-sense given by

$$X(t) = \sum_{n=-\infty}^{\infty} X\left(\frac{n\pi}{W}\right) \sqrt{\frac{W}{\pi}}\, \varphi_n(t), \tag{20}$$

where the φ_n are defined by

$$\varphi_n(t) = \sqrt{\frac{\pi}{W}}\, S_W\left(t - \frac{n\pi}{W}\right), \qquad t \in \mathbb{R}, \quad n \in \mathbb{Z},$$

$\{\varphi_n\}_{-\infty}^{\infty}$ forms a complete orthonormal system in **BLW**, and it is called a *system of sampling functions.* We can say that a sampling theorem is a Fourier expansion of an L^2-function with respect to this system of sampling functions.

A sampling theorem holds for some stochastic processes. Let $\{X(t)\}$ be an $L_0^2(\Omega)$-valued weakly harmonizable process with the representing measure ξ, i.e.,

$$X(t) = \int_{-\infty}^{\infty} e^{itu}\, \xi(du), \qquad t \in \mathbb{R}.$$

We say that $\{X(t)\}$ is *bandlimited* if there exists a $W > 0$ such that the support of ξ is contained in $[-W, W]$, i.e., $\xi(A) = 0$ if $A \cap [-W, W] = \emptyset$. If this is the case, the *sampling theorem* holds:

$$X(t, \omega) = \sum_{n=-\infty}^{\infty} X\left(\frac{n\pi}{W}, \omega\right) \frac{\sin(W(t - n\pi/W))}{W(t - n\pi/W)}, \qquad t \in \mathbb{R},$$

where the convergence is in $\|\cdot\|_2$ for each $t \in \mathbb{R}$.

C. Series Representation

The above two expansions have some restrictions. Namely, the KL expansion is valid for stochastic processes on finite closed intervals, and the sampling theorem requires bandlimitedness. To relax these conditions we consider separability of the time domain $H(X)$, which is a fairly general condition. Thus let us consider the series representation of a second-order stochastic process $\{X(t, \omega)\}$:

$$X(t, \omega) = \sum_{n=1}^{\infty} \varphi_n(t) X_n(\omega). \tag{21}$$

It is easily verified that $\{X(t, \omega)\}$ has a series representation of the form (21) if and only if the time domain $H(X)$ is separable, i.e., there exists a countable orthonormal base $\{Y_k\}_{k=1}^{\infty}$ for $H(X)$. If we choose one such base $\{X_k\}_{k=1}^{\infty}$, then the $\varphi_n(t)$ are the Fourier coefficients, i.e.,

$$\varphi_n(t) = (X(t), X_n)_2, \qquad t \in \mathbb{R}, \quad n \geq 1.$$

Moreover, the covariance function $K(s, t)$ is expressed as

$$K(s, t) = \sum_{n=1}^{\infty} \varphi_n(s) \overline{\varphi_n(t)}, \qquad s, t \in \mathbb{R}.$$

There are some sufficient conditions for a process $\{X(t)\}$ on T to have such a series representation:

1. $\{X(t)\}$ is weakly continuous on T, i.e., $(X(t), Y)_2$ is a continuous function of t for any $Y \in L_0^2(\Omega)$.
2. $\{X(t)\}$ is mean-square continuous on T.
3. The σ-algebra Σ has a countable generator.
4. T is a countable set, e.g., $\mathbb{Z}$ or $\mathbb{Z}^+$.

Since the covariance function $K(s,t)$ is positive definite, we can associate the RKHS (reproducing kernel Hilbert space) $\mathcal{H}_K$ consisting of $\mathbb{C}$-valued functions on T in a way that the following hold:

(i) $K(\cdot,t)\in\mathcal{H}_K$ for every $t\in T$.
(ii) $f(t)=(f(\cdot),K(\cdot,t))_K$ for every $f\in\mathcal{H}_K$ and $t\in T$.

Here $(\cdot,\cdot)_K$ is the inner product in $\mathcal{H}_K$. Then, an interesting fact is that, if $\{X_k\}_{k=1}^{\infty}$ is an orthonormal base in the time domain $H(X)$ and $\varphi_k(t)=(X(t),X_k)_2$ for $k\geq 1$, then $\{\varphi_k\}_{k=1}^{\infty}$ is also an orthonormal base in the RKHS $\mathcal{H}_K$. Hence we have that $H(X)$ and $\mathcal{H}_K$ are isomorphic Hilbert spaces.

VIII. EXTENSIONS OF STOCHASTIC PROCESSES

A. Finite-Dimensional Extension

So far we have considered mainly one-dimensional second-order stochastic processes $\{X(t)\}$. If there are two such processes $\{X_1(t)\}$ and $\{X_2(t)\}$, we can describe these processes as a single process by letting

$$\mathbf{X}(t)=(X_1(t),X_2(t)).$$

Hence, $\{\mathbf{X}(t)\}$ is regarded as an $[L^2(P)]^2$-valued process. If $k\geq 2$ is an integer and $\{X_j(t)\}$ $(j=1,\ldots,k)$ are one-dimensional second-order stochastic processes, then

$$\mathbf{X}(t)=(X_1(t),\ldots,X_k(t))$$

is a k-dimensional stochastic process, or an $[L^2(P)]^k$-valued process. For such a process $\{\mathbf{X}(t)\}$ the average $\mathbf{m}(t)$ is a k-dimensional vector given by

$$\mathbf{m}(t)=E\{\mathbf{X}(t)\}=(m_1(t),\ldots,m_k(t)),\qquad t\in T,$$

where $m_j(t)=E\{X_j(t)\}$, $1\leq j\leq k$, and the *scalar covariance function* $K(s,t)$ is given by

$$\begin{aligned}K(s,t)&=(\mathbf{X}(s)-\mathbf{m}(s),\mathbf{X}(t)-\mathbf{m}(t))_{2,k}\\&=\sum_{j=1}^{k}(X_j(s)-m_j(s),X_j(t)-m_j(t))_2\\&=\sum_{j=1}^{k}K_j(s,t),\end{aligned}$$

where $(\cdot,\cdot)_{2,k}$ is the inner product in $[L^2(P)]^k$ and $K_j(s,t)$ is the covariance function of $\{X_j(t)\}$. However, the scalar covariance function does not well reflect the correlatedness of $\{X_j(t)\}$ and $\{X_\ell(t)\}$ for $1\leq j,\ \ell\leq k$. The *matricial covariance function* $\mathbf{K}(s,t)$ is defined by

$$\mathbf{K}(s,t)=(K_{j\ell}(s,t))_{j,\ell},$$

where $K_{j\ell}(s,t)=(X_j(s)-m_j(s),X_\ell(t)-m_\ell(t))_2$, the *cross-covariance* of $\{X_j(t)\}$ and $\{X_\ell(t)\}$, and it expresses correlatedness among all the processes $\{X_j(t)\}$, $1\leq j\leq k$. The k-dimensional weakly stationary class and nonstationary classes (given in Section IV and V) are defined using the matricial covariance function $\mathbf{K}(s,t)$, and for many processes in these classes, the integral representations of the processes are obtained in terms of $L_0^2(P)]^k$-valued measures.

B. Infinite-Dimensional Extension

Finite-dimensional extension is rather straightforward, as seen above. However, infinite-dimensional extension needs considerably more work. One way to do this is to identify $[L_0^2(P)]^k$ with $L_0^2(P;\mathbb{C}^k)$ for each $k\geq 1$, where the latter is the space of all $\mathbb{C}^k$-valued random vectors with finite second moment and zero expectation. Then, when $k=\infty$, we replace $\mathbb{C}^k$ by an infinite-dimensional Hilbert space H. Thus we consider the Hilbert space $L_0^2(P;H)$ of all H-valued random variables X on (Ω,Σ,P) with finite second moment $\|X\|_2^2=\int_\Omega\|X(\omega)\|_H^2\,P(d\omega)<\infty$ and zero expectation, where $\|\cdot\|_H$ is the inner product norm in H. As in the finite-dimensional case, we have two kinds of covariance functions, the scalar and the operator (instead of matricial) ones. Then, we obtain the stationary and nonstationary classes mentioned above.

C. Stochastic Fields

Another type of extension of stochastic processes is to consider a time parameter set T to be other than a subset of $\mathbb{R}$. It is sometimes appropriate to replace $\mathbb{R}$ by $\mathbb{R}^k$ ($k\geq 2$, an integer) and consider stochastic processes $\{X(\mathbf{t})\}$ on $\mathbf{T}\subseteq\mathbb{R}^k$. These processes are called *stochastic fields.* When the stochastic process is depending on time t and two-dimensional location (x_1,x_2), then by letting $\mathbf{t}=(t,x_1,x_2)\in\mathbb{R}^3$, we may describe the process as $\{X(\mathbf{t})\}$. (The more abstract case is obtained by replacing $\mathbb{R}^k$ by a locally compact Abelian group G.) Most of the results obtained so far can be extended to processes on $\mathbb{R}^k$. For instance, if $\{X(\mathbf{t})\}$ is a weakly harmonizable process on $\mathbb{R}^k$ with the covariance function $K(\mathbf{t})$, then there exists an $L_0^2(\Omega)$-valued measure ξ and a positive-definite bimeasure β on $\mathfrak{B}(\mathbb{R}^k)\times\mathfrak{B}(\mathbb{R}^k)$ such that

$$X(\mathbf{t})=\int_{\mathbb{R}^k}e^{i(\mathbf{t},\mathbf{u})}\,\xi(d\mathbf{u}),\qquad\mathbf{t}\in\mathbb{R}^k,$$

$$K(\mathbf{s},\mathbf{t})=\int_{\mathbb{R}^{2k}}e^{i((\mathbf{s},\mathbf{u})-(\mathbf{t},\mathbf{v}))}\,\beta(d\mathbf{u},d\mathbf{v}),\qquad\mathbf{s},\mathbf{t}\in\mathbb{R}^k,$$

where $(\mathbf{s},\mathbf{u})=\sum_{j=1}^{k}s_ju_j$, the inner product in $\mathbb{R}^k$.

D. Generalized Stochastic Processes

Generalized stochastic processes were initiated by Itô (1953) and Gel'fand (1955). Let $\{X(t)\}$ be a stochastic process on $\mathbb{R}$ and $\mathfrak{D}$ be the set of all *test functions* on $\mathbb{R}$, i.e., $\phi \in \mathfrak{D}$ if ϕ is infinitely many times differentiable and is zero outside of some finite closed interval. Hence the dual space of $\mathfrak{D}$ is the set of *distributions* in the sense of Schwartz and not probability distributions. If

$$X(\phi) = \int_{-\infty}^{\infty} X(t)\phi(t)\,dt \tag{22}$$

is defined for each $\phi \in \mathfrak{D}$, so that $X(\phi)$ represents a random variable, then $\{X(\phi)\}$ is regarded as a stochastic process on $\mathfrak{D}$, called a *generalized stochastic process* induced by an ordinary stochastic process $\{X(t)\}$.

Here is a formal definition. Let $X(\phi, \omega)$ be a function of $\phi \in \mathfrak{D}$ and $\omega \in \Omega$ such that the following hold:

(i) $X(\cdot, \omega)$ is a distribution for a.e. $\omega \in \Omega$.
(ii) $X(\phi, \cdot)$ is a random variable for every $\phi \in \mathfrak{D}$.

Then $\{X(\phi)\}$ is called a *generalized stochastic process* on $\mathfrak{D}$. It is known that not every such process is induced by an ordinary stochastic process as in (22).

The derivative of a generalized stochastic process $\{X(\phi)\}$ can be obtained using that of distribution:

$$X'(\phi) = \langle \phi, X' \rangle = \langle -\phi', X \rangle = X(-\phi')$$

for $\phi \in \mathfrak{D}$, where $\langle \cdot, \cdot \rangle$ is a duality pair. In general, we have

$$X^{(p)}(\phi) = (-1)^p \langle \phi^{(p)}, X \rangle = (-1)^p X\left(\phi^{(p)}\right)$$

for $p \geq 2$ and $\phi \in \mathfrak{D}$. If $\{B(t)\}$ is a Brownian motion (or a Wiener process) on $\mathbb{R}$, then this induces a generalized stochastic process $\{B(\phi)\}$ on $\mathfrak{D}$. Its derivative $\{B'(\phi)\}$ is regarded as a white noise Gaussian process.

SEE ALSO THE FOLLOWING ARTICLE

CALCULUS • EVOLUTIONARY ALGORITHMS AND METAHEURISTICS • MATHEMATICAL MODELING • OPERATIONS RESEARCH • PROBABILITY • QUEUEING THEORY • STOCHASTIC DESCRIPTION OF FLOW IN POROUS MEDIA

BIBLIOGRAPHY

Balakrishnan, A. V. (1957). A note on the sampling principle for continuous signals, *IRE Trans. Inform. Theory* **IT-3,** 143–146.

Cramér, H. (1951). A contribution to the theory of stochastic processes, *Proc. Second Berkeley Symp. Math. Statist. Probab*. (J. Neyman, ed.) Univ. of California Press, Berkeley, pp. 329–339.

Gel'fand, I. M. (1955). Generalized random processes, *Dokl. Acad. Nauk SSSR* **100,** 853–856.

Itô, K. (1953). Stationary random distributions, *Mem. Coll. Sci. Univ. Kyoto* **28,** 209–223.

Kakihara, Y. (1997). "Multidimensional Second Order Stochastic Processes," World Scientific, Singapore.

Karatzas, I., and Shreve, S. E. (1991). "Brownian Motion and Stochastic Calculus," Springer, New York.

Karhunen, K. (1947). Über lineare Methoden in der Wahrscheinlichkeitsrechung, *Ann. Acad. Sci. Fenn. Ser. A. I. Math.* **37,** 1–79.

Khintchine, A. Ya. (1934). Korrelationstheorie der stationäre stochastischen Prozesse, *Math. Ann*. **109,** 605–615.

Kolmogorov, A. N. (1933). "Grundbegriffe der Wahrscheinlichkeitsrechung," Springer-Verlag, Berlin.

Loéve, M. (1948). Fonctions aléatoires du second ordre, Appendix to P. Lévy, "Processus Stochastiques et Movement Brownian," Gauthier-Villas, Paris, pp. 299–352.

Rao, M. M. (1982). Harmonizable processes: structure theory, *L'Enseign. Math*. **28,** 295–351.

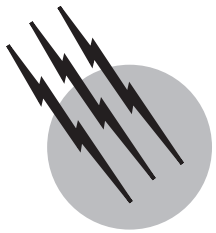
Rao, M. M. (1984). "Probability Theory with Applications," Academic Press, New York.

Rao, M. M. (1995). "Stochastic Processes: General Theory," Kluwer Academic, New York.

Rao, M. M. (2000). "Stochastic Processes: Inference Theory," Kluwer Academic, New York.

Rozanov, Yu. A. (1959). Spectral analysis of abstract functions, *Theory Probab. Appl*. **4,** 271–287.

Shannon, C. E. (1949). Communication in the presence of noise, *Proc. IRE* **37,** 10–21.



Tilings

Egon Schulte
Northeastern University

I. Preliminary Concepts
II. Fundamental Concepts of Tilings
III. General Considerations
IV. Plane Tilings
V. Monohedral Tilings
VI. Nonperiodic Tilings

GLOSSARY

Anisohedral tile Shape that admits a monohedral tiling but no isohedral (tile-transitive) tiling.
Aperiodic prototile set Set of prototiles that admits a tiling, but each tiling is nonperiodic.
Archimedean tiling Edge-to-edge tiling of the plane by convex regular polygons whose symmetry group is vertex transitive.
Lattice tiling Tiling by translates of a single tile such that the corresponding translation vectors form a lattice.
Monohedral tiling Tiling in which the tiles are congruent to a single prototile.
Normal tiling Tiling in which the tiles are uniformly bounded in size.
Parallelotope Convex polytope that is the prototile of a lattice tiling.
Penrose tilings Certain nonperiodic tilings of the plane with an aperiodic prototile set (discovered by Penrose).
Periodic tiling Tiling for which the symmetry group is a crystallographic group (contains translations in d independent directions, where d is the dimension of the ambient space).
Prototiles A minimal collection of shapes such that each tile in the tiling is congruent to a shape in the collection.
Space-filler Convex polytope that is the prototile of a monohedral tiling.
Voronoi region For a point x in a discrete set L, the Voronoi region of x consists of all points in space that are at least as close to x as to any other point in L.

TILINGS (or tessellations) have been investigated since antiquity. Almost all variants of the question "*How can a given space be tiled by copies of one or more shapes?*" have been studied in some form or another. The types of spaces permitted have included Euclidean spaces, hyperbolic spaces, spheres, and surfaces, and the types of shapes have varied from simple polygonal or polyhedral shapes to sets with strange topological properties. Today, tiling theory is a rapidly growing field that continues to pose challenging mathematical problems to scientists and mathematicians. The discovery of quasicrystals in 1984 has sparked a surge in interest in tiling theory because of its relevance for mathematical modeling of crystals and quasicrystals.

For more than 200 years, crystallographers such as Fedorov, Voronoi, Schoenflies, and Delone have greatly influenced tiling theory and the study of crystallographic groups. Another important impetus came from the geometry of numbers, notably from Minkowski's work. Hilbert's famous list of fundamental open problems in mathematics, posed in 1900, contained an unsolved problem on discrete groups and their fundamental regions that has had a strong impact on tiling theory. In the 1970s, Grünbaum and Shephard began their comprehensive work on tilings that resulted in a beautiful book on plane tilings. To date, no complete account on tilings in higher dimensional spaces is available in the literature. The recent developments on nonperiodicity of tilings are among the most exciting in tiling theory. The mathematics of aperiodic order is still in its infancy but, when developed, should be useful for understanding quasicrystals and other disordered solid materials.

The purpose of this chapter is to give a short survey on tilings in Euclidean spaces. The discussion is essentially limited to relatively well-behaved tilings and tiles.

I. PRELIMINARY CONCEPTS

It is necessary to establish some preliminary concepts before beginning the subject of tilings itself. Unless stated otherwise, the underlying space of a tiling will always be d-dimensional Euclidean space, or simply *d-space*, E^d. This is real d-space R^d equipped with the standard inner (scalar) product given by:

$$x \cdot y = \sum_{i=1}^{d} x_i y_i,$$

$$x = (x_1, \ldots, x_d), \quad y = (y_1, \ldots, y_d) \in \mathrm{E}^d$$

The length of a vector x in E^d is denoted by $|x|$.

An *orthogonal transformation* of E^d is a linear self-mapping of E^d which preserves the inner product and hence lengths and distances. A *similarity transformation* σ of E^d is a self-mapping of E^d of the form:

$$\sigma(x) = c\lambda(x) + t \qquad (x \in \mathrm{E}^d)$$

where λ is an orthogonal transformation of E^d, c a positive scalar (giving the expansion factor), and t a vector in E^d (determining the *translation part* of σ). A (Euclidean) *isometry* of E^d is a similarity transformation with $c = 1$.

Two subsets P and Q of E^d are *congruent* (respectively, *similar*), if there exists an isometry (similarity transformation) of E^d which maps P onto Q.

Let x be a point in E^d, and let $r > 0$. Then the d-dimensional (Euclidean) *ball* $B(x, r)$ of radius r centered at x consists of all points y in E^d whose distance from x is at most r. The *unit ball* in E^d is the ball of radius 1 centered at the origin.

A subset P of E^d is *open* if every point of P admits a neighborhood (a d-dimensional ball of small positive radius) which is entirely contained in P. The union of any number of open sets is again open. A subset P of E^d is *closed* if its complement $\mathrm{E}^d \setminus P$ in E^d is open. The intersection of any number of closed sets is again closed. A subset P of E^d is *compact* if it is closed and bounded (contained in a ball of large radius). For example, every ball $B(x, r)$ is compact.

Let P be a subset of E^d. The *interior* $\mathrm{int}(P)$ of P consists of those points, called the *interior points*, of P which admit a neighborhood entirely contained in P; this is the largest open set contained in P. The interior of a ball is also called an *open ball*, in contrast to the ball itself, which is a *closed ball*. The *closure* $\mathrm{cl}(P)$ of a subset P is the intersection of all closed subsets that contain P; this is the smallest closed subset which contains P. The *boundary* $\mathrm{bd}(P)$ of P is the set $\mathrm{int}(P) \setminus \mathrm{cl}(P)$, consisting of the *boundary points* of P. For example, the *unit sphere* (centered at the origin) in E^d is the boundary of the unit ball in E^d.

Two open (respectively, closed) subsets P and Q of E^d are *homeomorphic* if there exists a bijective mapping f from P onto Q such that both f and its inverse f^{-1} are continuous; such a mapping is called a *homeomorphism* from P onto Q. A subset P of E^d is called a *topological k-ball* (respectively, *topological k-sphere*) if it is homeomorphic to a Euclidean k-ball (k-sphere). A *topological disc* is a topological 2-ball.

A subset P of E^d is *convex* if, for every two points x and y in P, the line segment joining x and y is contained in P. The intersection of any number of convex sets is again convex.

Let P be a subset of E^d. The *convex hull*, $\mathrm{conv}(P)$, of P is the intersection of all convex sets which contain P; it is the smallest convex set which contains P.

A convex *k-polyhedron* P in E^d is the intersection of finitely many closed half-spaces in E^d which is k-dimensional. If $k = d$, then P has nonempty interior in E^d, and vice versa. A convex *k-polytope* is a bounded convex k-polyhedron. A subset of E^d is a convex polytope if and only if it is the convex hull of finitely many points in E^d. Every convex d-polytope in E^d is a topological d-ball. In applications, the tiles of a tiling will often be convex polytopes.

The boundary of a convex d-polytope P in E^d splits into finitely many lower dimensional polytopes called the *faces* of P. These include the empty set (*empty face*) and P itself as *improper* faces, of dimensions -1 and d, respectively. A *proper* face F of P is the (nonempty) intersection of P with a *supporting hyperplane* of P; this is a hyperplane H of E^d that intersects P in F such that P lies entirely in

one of the two closed half-spaces bounded by H. A face of dimension 0, 1, i, or $d-1$ is also called a *vertex*, an *edge*, an *i-face*, or a *facet*, respectively. The family of all proper and improper faces of P is called the *face lattice* of P. (When ordered by set-theoretic inclusion, this partially ordered set is a lattice.) The following are two important properties of the face lattice of P: first, a face of a face is again a face (that is, if F is a face of P and G is a face of the polytope F, then G is also a face of P); and, second, the intersection of any two faces of P is again a face of P, possibly the empty face. The *boundary complex* of P, consisting of all proper faces of P (and the empty face), yields a decomposition of the boundary of P, which is a topological sphere.

Two convex d-polytopes P and Q are *combinatorially equivalent* if there exists a mapping f from the face lattice of P onto the face lattice of Q which is one-to-one and inclusion preserving; such a mapping is called a (*combinatorial*) *isomorphism*. An *automorphism* of P is a self-mapping of the face lattice of P which is an isomorphism. For example, a rectangle is combinatorially equivalent to a square and has the same number of automorphisms as the square.

A *parallelopiped* P is a convex d-polytope that is spanned by d linearly independent vectors $b_1, \ldots, b_d$ of E^d; that is, P is the set of linear combinations $\sum_{i=1}^{d} \lambda_i b_i$ with $0 \le \lambda_i \le 1$ for all $i = 1, \ldots, d$. The simplest example is the standard d-dimensional (unit) *cube* (or *hypercube*), obtained when $b_1, \ldots, b_d$ are the canonical base vectors of E^d. Any parallelopiped is combinatorially equivalent to the cube.

Groups are certain sets Γ with an algebraic structure. We shall not define the term in full generality but restrict ourselves to groups where the elements of the underlying set Γ are one-to-one self-mappings of another set X, such as nonsingular affine transformations of $X = \mathrm{E}^d$, isometries of $X = \mathrm{E}^d$, or simply bijections of X. A set Γ of this kind is called a *group* if the following two properties are satisfied: first, the composition of any two mappings in Γ is again a mapping in Γ, and, second, for every mapping in Γ, the inverse mapping also belongs to Γ. A group Γ is *finite* if the set Γ has finitely many elements; otherwise, it is *infinite*. The *order* of a finite group Γ is the number of elements of Γ.

A *subgroup* Γ' of a group Γ is a subset of Γ which itself is a group. A (*left*) *coset* of a subgroup Γ' is a subset of Γ of the form $\sigma\Gamma' := \{\sigma\gamma \mid \gamma \in \Gamma'\}$. The number of cosets of a subgroup Γ' is called the *index* of Γ' in Γ.

Two subgroups Γ' and Γ'' of a group Γ are said to be *conjugate subgroups* of Γ if there exists an element $\sigma \in \Gamma$ such that $\Gamma'' = \sigma\Gamma'\sigma^{-1} (= \{\sigma\gamma\sigma^{-1} \mid \gamma \in \Gamma'\})$.

Two groups Γ_1 and Γ_2 are said to be *isomorphic* if there exists a bijective mapping f from Γ_1 onto Γ_2 such that $f(\gamma\gamma') = f(\gamma)f(\gamma')$ for all $\gamma, \gamma' \in \Gamma_1$. Then f is called an *isomorphism* between Γ_1 and Γ_2.

Let P be a nonempty subset of E^d. A *symmetry* of P is an isometry which maps P onto itself. The set of all symmetries of P forms a group, the *symmetry group* $S(P)$ of P.

The set of all automorphisms of a convex polytope P also forms a group, the *automorphism group* $\Gamma(P)$ of P. Since each symmetry of a polytope maps faces to faces, the symmetry group $S(P)$ is a subgroup of $\Gamma(P)$; in general, $S(P)$ is a proper subgroup of $\Gamma(P)$.

As before, let Γ be a group of self-mappings of a set X. Then Γ is said to *act transitively*, or to be *transitive*, on X if for any two elements x and y in X there exists a mapping σ in Γ such that $\sigma(x) = y$. The *orbit* of an element x in X is the set of all images of x under mappings in Γ. Then Γ acts transitively on X if and only if X is the orbit of any (indeed, every) element x in X. For a given x in X, the subgroup of Γ consisting of the mappings that fix x is called the *stabilizer* of x in Γ.

For example, the *dihedral group* D_n, of order $2n$, is the symmetry group of the convex regular n-gon. The group consists of n rotations and n reflections, and it acts transitively on both the set X_1 of vertices and the set X_2 of edges. The subgroup of D_n consisting of the n rotations is denoted C_n and is a *cyclic* group of order n; that is, all elements are powers of one element, the rotation by $2\pi/n$ in this case. In these examples, as in many other instances described later, the original group Γ is a group of isometries of Euclidean space (in this case, the plane) X, but Γ can also be viewed as a group of self-mappings of the vertex-set X_1 or edge-set X_2 of the regular n-gon.

A subset P of E^d is called *discrete* if each point x in P has an open neighborhood that does not contain any other point of P.

A group Γ of isometries of E^d is called *discrete* (or, more exactly, is said to *act discretely*) if the orbit of every point x in E^d is a discrete subset of E^d. If Γ is a discrete group of isometries of E^d, then a *fundamental region* for Γ is an open set D of E^d satisfying the following two properties: first, the images of D under elements of Γ are mutually disjoint, and, second, the union of the closures of the images of D under Γ is the entire space E^d. Every discrete group of isometries has a fundamental region in E^d (whose boundary is a set of measure zero).

A *lattice* L in E^d is the group of all integral linear combinations of a set of d linearly independent vectors in E^d; these vectors form a *basis* of L. A lattice has many different bases. The lattices in E^d are precisely the discrete subgroups of (the additive group) E^d which contain d linearly independent vectors. By Z^d we denote the standard integer lattice in E^d consisting of all vectors with only integral coordinates.

If z is a point in E^d, then the *point reflection* in z is the affine mapping that takes each x in E^d to $2z - x$. A subset P of E^d is called *centrally symmetric* if it is invariant under the point reflection in a point z, called a *center of symmetry* of P. A set can have more than one center of symmetry.

II. FUNDAMENTAL CONCEPTS OF TILINGS

A. What Is a Tiling?

A *tiling* (or *tessellation* or *honeycomb*) $\mathcal{T}$ of Euclidean d-space E^d is a countable family of closed subsets T of E^d, the *tiles* of $\mathcal{T}$, which cover E^d without gaps and overlaps. This means that the union of all tiles of $\mathcal{T}$ is the entire space, and that any two distinct tiles do not have interior points in common. To rule out pathological situations we shall assume that the tiles of $\mathcal{T}$ are (closed) topological d-balls. In applications, the tiles will often be convex d-polytopes.

A tiling $\mathcal{T}$ of E^d is called *locally finite* if each point of E^d has a neighborhood that meets only finitely many tiles. Then, in fact, every bounded region of space meets only finitely many tiles. We will always assume that a tiling is locally finite.

A tiling $\mathcal{T}$ of E^d by topological d-balls is called *normal* if its tiles are uniformly bounded (that is, there exist positive real numbers r and R such that each tile contains a Euclidean ball of radius r and is contained in a Euclidean ball of radius R) and the intersection of every pair of tiles is a connected set. The latter condition is trivially satisfied if the tiles are convex d-polytopes. On the other hand, in a tiling by arbitrary topological d-balls, the intersection pattern of any two or more tiles can be rather complicated, so connectedness is a reasonable requirement for normality. Note that a normal tiling is necessarily locally finite. Figure 1 shows an example of a locally finite plane tiling by convex heptagons which is not normal.

We often do not distinguish two tilings $\mathcal{T}_1$ and $\mathcal{T}_2$ of E^d which are obtained from each other by a similarity transformation of E^d; such tilings are said to be *equal* or *the same* (this, of course, is an abuse of the standard notion of equality). In particular, we use the following notation. Two tilings $\mathcal{T}_1$ and $\mathcal{T}_2$ of E^d are called *congruent* (respectively, *similar*) if there is an isometry (similarity transformation) of E^d which maps the tiles of $\mathcal{T}_1$ onto the tiles of $\mathcal{T}_2$.

A central notion is that of a symmetry of a tiling. A Euclidean isometry of E^d is a *symmetry* of a tiling $\mathcal{T}$ if it maps each tile of $\mathcal{T}$ onto a tile of $\mathcal{T}$. The set of all symmetries of $\mathcal{T}$ forms a group, the *symmetry group* $S(\mathcal{T})$ of $\mathcal{T}$.

A *protoset*, or *prototile set*, of a tiling $\mathcal{T}$ of E^d is a minimal subset of tiles of $\mathcal{T}$ such that each tile of $\mathcal{T}$ is congruent to one of those in the subset. The tiles in the set are the *prototiles* of $\mathcal{T}$, and the protoset is said to *admit* the tiling $\mathcal{T}$. By abuse of notation, we also use this terminology for shapes and family of shapes that are under consideration for being prototiles or protosets of a tiling, respectively.

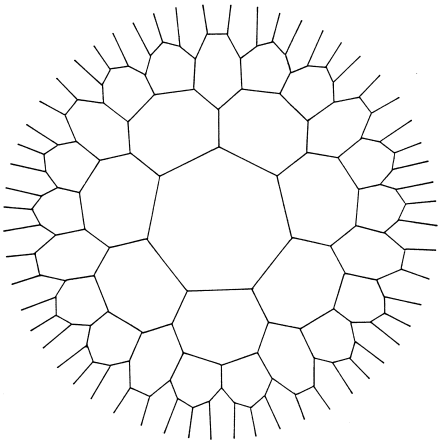


FIGURE 1 A non-normal tiling of the plane by convex heptagons, three meeting at each vertex. The tiles become longer and thinner the farther away they are from the center. [From Grünbaum, B., and Shephard, G. C. (1986). *Tilings and Patterns*, Freeman & Co., San Francisco.]

Tilings are closely related to packings and coverings of space. A tiling is a family of sets which cover E^d without gaps and overlaps. A *packing* is a family of sets in E^d which do not overlap; that is, no two distinct sets have interior points in common. On the other hand, a *covering* is a family of sets in E^d which leave no gaps; that is, the union of all sets is the entire space E^d. The efficiency of a packing or covering is measured by its *density*, which is a well-defined quantity provided the sets are distributed in a sufficiently regular way. The density of a packing measures the fraction of space that is covered by the sets; the density of a covering measures the average frequency with which a point of space is covered by the sets. Thus, tilings are at the same time packings and coverings of space, each with density 1.

B. Tilings by Polytopes

The best-behaved tilings of E^d are the face-to-face tilings by convex polytopes. More precisely, a tiling $\mathcal{T}$ by convex d-polytopes is said to be *face-to-face* if the intersection of any two tiles is a face of each tile, possibly the empty face. In a face-to-face tiling, the intersection of any number of tiles is a face of each of the tiles.

In a face-to-face tiling $\mathcal{T}$ of E^d by convex d-polytopes, the i-faces of the tiles are also called the *i-faces* of $\mathcal{T}$ ($i = 0, \ldots, d$). The d-faces of $\mathcal{T}$ are then the tiles of $\mathcal{T}$.

The faces of dimensions 0 and 1 are the *vertices* and *edges* of $\mathcal{T}$, and the empty set and E^d itself are considered to be *improper faces* of $\mathcal{T}$, of dimensions -1 and $d+1$, respectively. The set of all faces of $\mathcal{T}$ is called the *face lattice* of $\mathcal{T}$. (When ordered by set-theoretic inclusion, this is a lattice.) This terminology carries over to more general tilings in which the tiles are topological d-polytopes (homeomorphic images of convex d-polytopes).

Two face-to-face tilings $\mathcal{T}_1$ and $\mathcal{T}_2$ by convex d-polytopes are *combinatorially equivalent* if there exists a mapping from the face lattice of $\mathcal{T}_1$ onto the face lattice of $\mathcal{T}_2$ which is one-to-one and inclusion preserving; such a mapping is called a (*combinatorial*) *isomorphism*. An *automorphism* of a tiling $\mathcal{T}$ is a self-mapping of the face lattice of $\mathcal{T}$ which is an isomorphism. The set of all automorphisms of $\mathcal{T}$ forms a group, the *automorphism group* $\Gamma(\mathcal{T})$ of $\mathcal{T}$. This group is generally larger than the symmetry group of $\mathcal{T}$.

A mapping between the face lattices of $\mathcal{T}_1$ and $\mathcal{T}_2$ is called a *duality* if it is one-to-one and inclusion reserving. If such a mapping exists, then $\mathcal{T}_2$ is said to be a *dual* of $\mathcal{T}_1$. A tiling can have many metrically distinct duals but any two duals are isomorphic. Except for highly symmetrical tilings, not much is known about the existence and properties of dual tilings.

C. Tilings of the Plane

A lot more is known about tilings in the plane than about tilings in spaces of three or more dimensions. Let $\mathcal{T}$ be a (locally finite) plane tiling in E^2 by topological discs. We now introduce the notion of a ($\mathcal{T}$-induced) vertex or a ($\mathcal{T}$-induced) edge of $\mathcal{T}$. A point x in E^2 is called a *vertex* of $\mathcal{T}$ if it is contained in at least three tiles. Since the tiles are topological discs, each simple closed curve which bounds a tile T is divided into a finite number of closed arcs by the vertices of $\mathcal{T}$, where any two arcs are disjoint except for possibly vertices of $\mathcal{T}$. These arcs are said to be the *edges of the tile* T and are also referred to as *edges of* $\mathcal{T}$. Each vertex x of $\mathcal{T}$ is contained in finitely many edges of $\mathcal{T}$. The number of edges emanating from a vertex x is called the *valence* $v(x)$ of x; clearly, $v(x) \geq 3$ for each vertex x. Note that two tiles can intersect in several edges, where pairs of edges may or may not have vertices in common. Two tiles with a common edge are called *adjacents* of each other.

If a tile of $\mathcal{T}$ is a planar polygon, then the $\mathcal{T}$-induced vertices and edges of the tile will generally not coincide with the original vertices and edges of the tile. To avoid confusion we shall refer to the latter as the *corners* and *sides* of the polygonal tile. In a tiling $\mathcal{T}$ by polygons, the vertices and edges of the tiles may or may not coincide with the corners and sides of the polygons, respectively; if they do, then we call $\mathcal{T}$ an *edge-to-edge* tiling.

In higher dimensions, no general terminology has been introduced that deals with the distinction of an *a priori* facial structure and the $\mathcal{T}$-induced facial structure on the tiles of $\mathcal{T}$.

The notions of isomorphism, automorphism and duality introduced in the previous subsection carry over to general plane tilings, with the understanding that we now have inclusion preserving or inclusion reversing one-to-one mappings between the sets of all ($\mathcal{T}$-induced) vertices, edges, and tiles of the tilings. For every normal plane tiling $\mathcal{T}$ there exists a normal tiling which is dual to $\mathcal{T}$. In the dual, every vertex is an interior point of a tile of $\mathcal{T}$, every tile contains one vertex of $\mathcal{T}$, and any two vertices that correspond to a pair of adjacent tiles in $\mathcal{T}$ are joined by an edge (arc) that crosses the edge common to both tiles.

For normal plane tilings, there is a close relationship between combinatorial and topological equivalence of tilings. Two tilings of E^d are said to be of the *same topological type* or to be *topologically equivalent* if there is a homeomorphism of E^d which maps one onto the other. Two normal plane tilings are topologically equivalent if and only if they are combinatorially equivalent.

D. Monohedral Tilings

A tiling $\mathcal{T}$ of E^d is *monohedral* if all its tiles are congruent to a single set T, the prototile of $\mathcal{T}$. The simplest examples of monohedral tilings $\mathcal{T}$ are those in which the tiles are translates of T. If T admits such a tiling, then we say that T *tiles by translation*. In such a tiling $\mathcal{T}$, if the corresponding translation vectors form a d-dimensional lattice L in E^d, then $\mathcal{T}$ is called a *lattice tiling* (*with lattice* L). If $\mathcal{T}$ is a lattice tiling of E^d with convex d-polytopes as tiles, then the prototile T is called a *parallelotope*, or *parallelohedron* if $d = 3$. Every parallelopiped in E^d is a parallelotope, and the vectors that span it also generate the corresponding lattice.

With every lattice in E^d, and indeed with every discrete set L in E^d, is associated a tiling with tiles called *Voronoi regions* or *Dirichlet regions*. Given a point x in L, the Voronoi region $V(L, x)$ of x is the set of all points in E^d that are at least as close to x as to any other point in L; that is,

$$V(L, x) = \{y \in \mathrm{E}^d \mid |y - x| \leq |y - z| \text{ for all } z \in L\}$$

The family of all Voronoi regions $V(L, x)$ with $x \in L$ gives a face-to-face tiling $\mathcal{T}$ of E^d by convex polyhedra (polytopes if L is a lattice), called the *Voronoi tiling* for L. Figure 2 shows the Voronoi regions for a finite set of points in the plane, including those regions that are unbounded.

If L is a lattice in E^d, then the Voronoi regions are translates of the Voronoi region $V(L) := V(L, 0)$ obtained

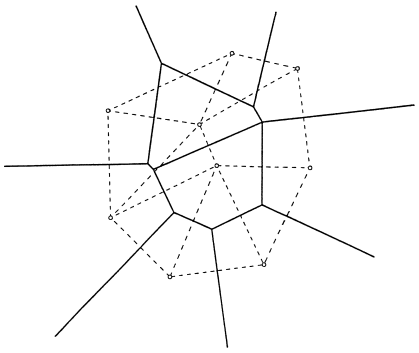


FIGURE 2 The Voronoi regions of a finite set of points in the plane. [From Schattschneider, D., and Senechal, M. (1997). *Handbook of Discrete and Computational Geometry*, (Goodman, J.E. and O'Rourke, J., eds.). CRC Press, Boca Raton.]

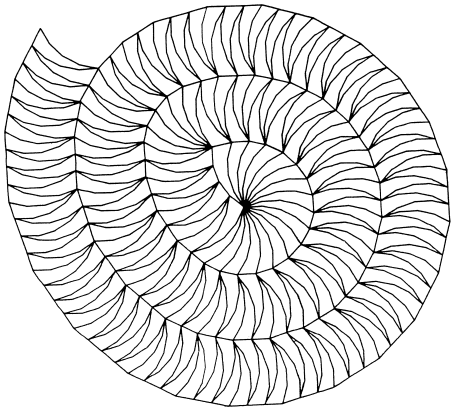


FIGURE 3 A spiral tiling of the plane. Many other spiral tilings can be constructed with the same prototile. [From Grünbaum, B., and Shephard, G. C. (1986). *Tilings and Patterns*, Freeman & Co., San Francisco.]

for $x = 0$, and $\mathcal{T}$ is a lattice tiling with $V(L)$ as prototile. In particular, the Voronoi regions are convex d-polytopes and $V(L)$ is a parallelotope. The translations by vectors in L all belong to the symmetry group of $\mathcal{T}$. The structure of the Voronoi regions of a lattice depends a great deal on the structure of the lattice itself.

The Voronoi tiling for a lattice L has a dual, called the *Delone tiling* for L, whose vertex set is L and whose tiles are called *Delone cells* (not to be confused with Delone sets defined later). The Delone cell corresponding to a vertex v of a Voronoi tiling is the convex hull of the lattice points x whose Voronoi regions $V(L, x)$ have v as a vertex.

If L is the standard integer lattice Z^d in E^d, then the Voronoi regions and Delone cells are unit cubes, and the Voronoi tiling and Delone tiling are cubical tessellations of E^d that are translates of each other.

A subset L of E^d is called a *Delone set* if there exist two positive real numbers r and R such that every ball of radius r contains at most one point of L, and every ball of radius R contains at least one point of L. The Delone sets are an important class of discrete point sets. Their Voronoi regions are again convex polytopes, but generally there is more than one shape.

There are many other important classes of monohedral tilings that we meet later on. From an artistic perspective, "spiral" tilings of the plane are among the most beautiful; see Fig. 3 for an example.

E. The Symmetry Group of a Tiling

The symmetry properties of a tiling $\mathcal{T}$ of E^d are captured by its symmetry group $S(\mathcal{T})$, which is a discrete group of Euclidean isometries of E^d. The structure of $S(\mathcal{T})$ depends essentially on its *rank* k, the number of independent translations in $S(\mathcal{T})$. Since the tiles in $\mathcal{T}$ are assumed to be topological d-balls, the group cannot contain arbitrarily small translations. In fact, the subgroup of all translations in $S(\mathcal{T})$, called the *translation subgroup* of $S(\mathcal{T})$, is in one-to-one correspondence with a k-dimensional lattice in E^d consisting of the corresponding translation vectors; here, $k = 0, 1, \ldots$ or d. A tiling $\mathcal{T}$ in E^d is called *periodic* if $S(\mathcal{T})$ contains translations in d linearly independent directions (that is, $k = d$). For example, all the tilings in Fig. 4 are periodic plane tilings; the translation subgroup is generated by two independent translations. Trivially, each lattice tiling of E^d is periodic. A tiling $\mathcal{T}$ is called *nonperiodic* if $S(\mathcal{T})$ contains no translation other than the identity (that is, $k = 0$). Note that a tiling that is not periodic need not be nonperiodic.

A group G of Euclidean isometries in E^d is called a *crystallographic group* or a *space group* if it is discrete and contains translations in d independent directions. The crystallographic groups are precisely the discrete groups of isometries in E^d whose fundamental region is compact. Two discrete groups are said to be *geometrically isomorphic* if they are conjugate subgroups in the group of all nonsingular affine transformations of E^d. Clearly, geometric isomorphism of discrete groups implies isomorphism as abstract groups. But for crystallographic groups, the converse is also true; that is, two crystallographic groups are geometrically isomorphic if and only if they are isomorphic as abstract groups. If $d = 2$, 3, or 4, the number of (geometric isomorphism) types of crystallographic groups in E^d is 17, 219, or 4783, respectively. If $d \geq 5$, the number of types in E^d is finite but is not explicitly known.

Every infinite orbit of a crystallographic group is a Delone set. The Delone sets that arise from crystallographic groups are also called *regular systems of points*.

For plane tilings $\mathcal{T}$, the three possible choices for the rank k of $S(\mathcal{T})$ lead to well-known classes of plane isometry groups:

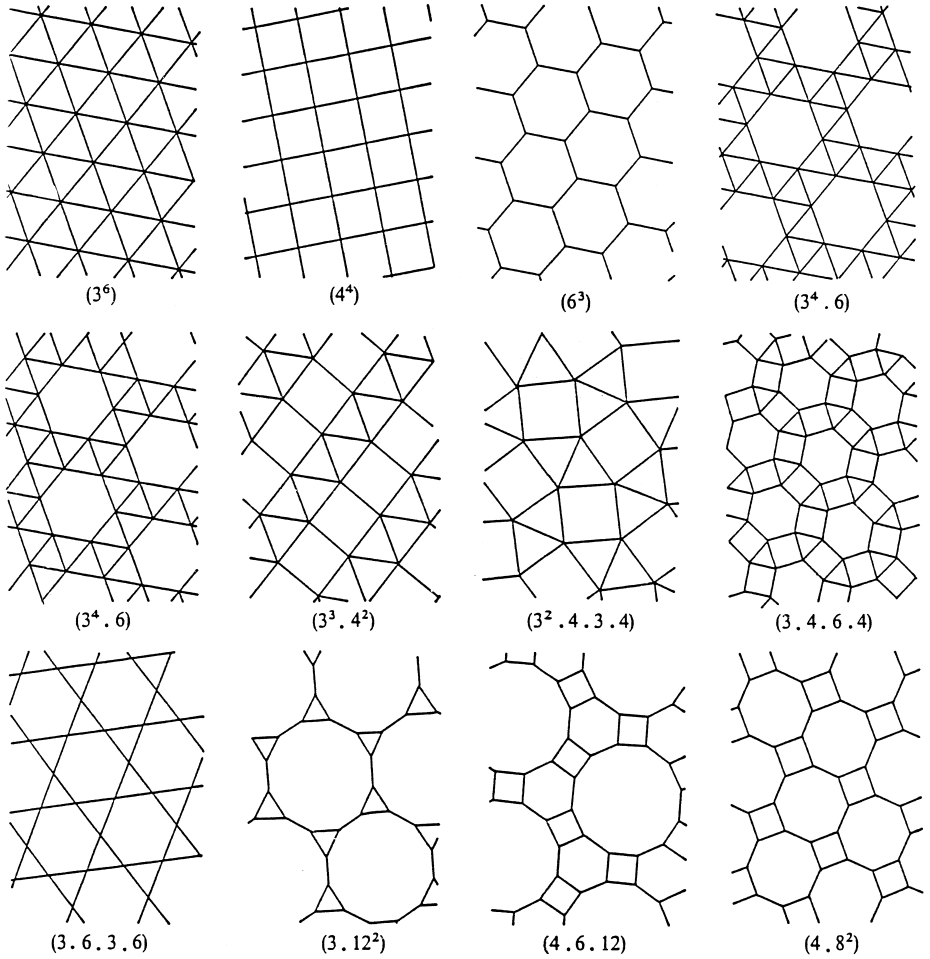


FIGURE 4 The 11 Archimedean tilings in the plane. [From Grünbaum, B., and Shephard, G. C. (1986). *Tilings and Patterns*, Freeman & Co., San Francisco.]

1. If $S(\mathcal{T})$ contains no translation other than the identity (that is, $k = 0$), then $S(\mathcal{T})$ is (isomorphic to) a cyclic group C_n or a dihedral group D_n for some $n \geq 1$. In this case, $\mathcal{T}$ has a *center*; that is, there is at least one point (exactly one if $n \geq 3$) of the plane that is left fixed by every symmetry of $\mathcal{T}$.

2. If $S(\mathcal{T})$ contains only translations in one independent direction, then $S(\mathcal{T})$ is (isomorphic to) one of seven groups called *strip groups* or *frieze groups*.

3. If $S(\mathcal{T})$ contains translations in two independent directions, then $\mathcal{T}$ is periodic and $S(\mathcal{T})$ is (isomorphic to) one of the 17 (*plane*) *crystallographic groups*, also known as *periodic groups* or *wallpaper groups*. (There are different sets of notations for these groups; the most common is that of the International Union of Crystallography, another is the orbifold notation proposed by Conway.)

Two tilings $\mathcal{T}_1$ and $\mathcal{T}_2$ of E^d are said to be of the *same symmetry type* if $S(\mathcal{T}_1)$ and $S(\mathcal{T}_2)$ are *geometrically isomorphic* as discrete groups.

An important problem in tiling theory is the classification of tilings $\mathcal{T}$ with respect to certain transitivity properties of $S(\mathcal{T})$. A tiling $\mathcal{T}$ of E^d is called *isohedral* if $S(\mathcal{T})$ acts transitively on the tiles of $\mathcal{T}$. Clearly, an isohedral tiling is monohedral but the converse is not true. A plane tiling $\mathcal{T}$ is said to be *isogonal* (*isotoxal*) if $S(\mathcal{T})$ acts transitively on the vertices (edges, respectively) of $\mathcal{T}$. Isogonality and isotoxality (and indeed, transitivity on the faces of any dimension) can also be studied more generally for face-to-face tilings $\mathcal{T}$ of E^d by convex d-polytopes.

A face-to-face tiling $\mathcal{T}$ of E^d is called *regular* if $S(\mathcal{T})$ acts transitively on the flags (maximal chains of mutually incident faces) of $\mathcal{T}$. These are the tilings with maximum possible symmetry. In the plane, there are only three regular tilings—namely, the well-known tilings by regular triangles, squares, and hexagons; see Fig. 4, where these tilings occur as (3^6), (4^4), and (6^3), respectively. For each dimension d, we also have the regular tessellation of E^d by d-cubes. Except for two additional exceptional

tilings in E^4, there are no other regular tilings in any dimension.

III. GENERAL CONSIDERATIONS

In this section we discuss a number of basic results which hold in any dimension. Throughout, $\mathcal{T}$ will be a locally finite tiling of E^d whose tiles are topological d-balls (that is, discs if $d = 2$). The *tiling problem* in E^d asks if there exists an algorithm which, when applied to any finite prototoset $\mathcal{S}$ (of topological d-balls) in E^d, decides whether or not $\mathcal{S}$ admits a tiling of E^d. The solution is given by the following *Undecidability Theorem*:

Theorem 1. *The tiling problem in the plane or in higher dimensional space is undecidable; that is, there exists no algorithm that decides whether or not an arbitrary finite set of prototiles admits a tiling.*

In constructing tilings we often come across the problem of having to extend a patch of tiles to a larger patch or a tiling of the entire space. By a *patch* in a tiling $\mathcal{T}$ of E^d we mean a finite collection of tiles of $\mathcal{T}$ whose union is a topological d-ball. Figure 5 shows a patch of a plane tiling.

A finite protoset $\mathcal{S}$ is said to *tile over arbitrarily large d-balls* in E^d if, for every d-ball B in E^d, the protoset admits a patch such that the union of the tiles in the patch contains B. A patch of tiles which covers a ball will generally not be extendable to a patch which covers a larger ball; that is, in constructing a global tiling it may be necessary to rearrange the tiles after each step. However, the following *Extension Theorem* holds:

Theorem 2. *Let $\mathcal{S}$ be any finite set of prototiles in E^d, each of which is a topological d-ball. If $\mathcal{S}$ tiles over arbitrarily large d-balls, then $\mathcal{S}$ admits a tiling of the entire space E^d.*

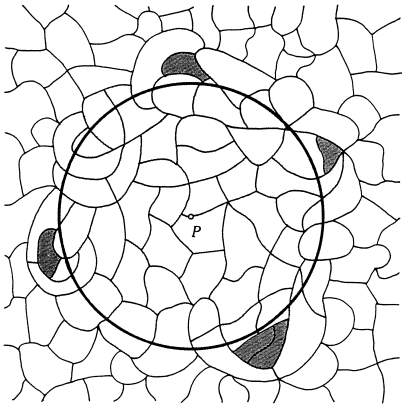


FIGURE 5 A patch of tiles in a plane tiling. The patch consists of the tiles that meet the circular disc, along with the dark-shaded tiles needed to make the arrangement a patch. [From Grünbaum, B., and Shephard, G. C. (1986). *Tilings and Patterns*, Freeman & Co., San Francisco.]

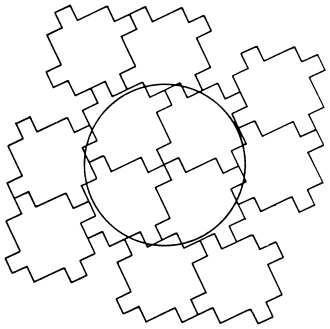


FIGURE 6 A nonextendable patch. Suitable tiles have been added to the 2×2 block of tiles in the center to make this patch nonextendable. There are similar examples based on $n \times n$ blocks for any $n \geq 2$. [From Grünbaum, B., and Shephard, G. C. (1986). *Tilings and Patterns*, Freeman & Co., San Francisco.]

To illustrate that rearranging of tiles may be necessary after each step, consider the patch in Fig. 6, which is the initial patch in an increasing sequence of similar patches. No patch is obtained by extending a preceding patch, yet the extension theorem implies that the given prototile admits a tiling of the entire plane.

An important class of tilings is given by the periodic tilings. For tiles that are convex polytopes, we have the following periodicity result:

Theorem 3. *Let $\mathcal{S}$ be a finite set of prototiles in E^d that are convex d-polytopes. If $\mathcal{S}$ admits a face-to-face tiling of E^d which possesses a (nontrivial) translational symmetry, then $\mathcal{S}$ also admits a tiling that is periodic (and may or may not be the same as the first tiling).*

We now discuss normality of tilings in E^d. Normality has strong implications on the frequency with which tiles are distributed along the boundary of large spherical regions in space. A *spherical patch* $P(x, r)$ in a tiling is the collection of tiles whose intersection with the ball $B(x, r)$ of radius r centered at the point x is nonempty, together with any additional tiles needed to complete the patch (that is, to make the union of its tiles homeomorphic to a topological d-ball). By $t(x, r)$ we denote the number of tiles in $P(x, r)$.

The following *Normality Lemma* implies that a normal tiling cannot have "singularities" at finite points or at infinity:

Theorem 4. *In a normal tiling $\mathcal{T}$ of E^d, the ratio of the number of tiles that meet the boundary of a spherical patch to the number of tiles in the patch itself tends to zero as the radius of the patch tends to infinity. More precisely, if x is a point in E^d and $s > 0$, then:*

$$\lim_{r\to\infty}\frac{t(x,r+s)-t(x,r)}{t(x,r)}=0$$

In a normal plane tiling, the tiles cannot have too many edges. In fact, as a consequence of the Normality Lemma we have

Theorem 5. *If all the tiles in a normal plane tiling have the same number k of edges, then* $k=3$, 4, 5, *or* 6.

Finally we note the following implication of convexity of the tiles in a tiling:

Theorem 6. *If* $\mathcal{T}$ *is a tiling of* E^d *with (compact) convex tiles, then necessarily each tile in* $\mathcal{T}$ *is a convex d-polyhedron (d-polytope, respectively).*

IV. PLANE TILINGS

Much of the attraction of plane tilings comes from their appearance in nature and art. They have been widely studied, and much more is known about tilings in the plane than about tilings in spaces of higher dimensions. We begin with some well-known classes of plane tilings; see Section II.C for basic notation.

A. Archimedean Tilings and Laves Tilings

Historically, tilings by (convex) regular polygons were the first kind to be investigated. If the tiles in an edge-to-edge tiling are congruent regular polygons, then the tiling must be one of the three regular tilings in the plane, by triangles, squares, or hexagons.

The Archimedean tilings are edge-to-edge tilings by regular polygons, not necessarily all of the same kind, in which all the vertices are surrounded in the same way by the tiles that contain it. More precisely, an edge-to-edge tiling $\mathcal{T}$ by regular polygons is said to be of *type* $(n_1.n_2.\ \ldots\ .n_r)$ if each vertex x of $\mathcal{T}$ is of *type* $n_1.n_2.\ \ldots\ .n_r$, meaning that, in a cyclic order, x is surrounded by an n_1-gon, an n_2-gon, and so on. Then, the type $(n_1.n_2.\ \ldots\ .n_r)$ of $\mathcal{T}$ is unique up to a cyclic permutation of the numbers n_i, and we will denote $\mathcal{T}$ itself by this symbol. The *Archimedian tilings* then are the eleven tilings described by the next theorem and depicted in Fig. 4; they were already enumerated by Kepler.

Theorem 7. *There exist precisely 11 distinct edge-to-edge tilings of the plane by convex regular polygons such that all vertices are of the same type*: (3^6), $(3^4.6)$, $(3^3.4^2)$, $(3^2.4.3.4)$, $(3.4.6.4)$, $(3.6.3.6)$, (3.12^2), (4^4), $(4.6.12)$, (4.8^2), and (6^3).

An edge-to-edge tiling $\mathcal{T}$ of the plane is called *uniform* if it is isogonal and its tiles are convex regular polygons.

Theorem 8. *The uniform plane tilings are precisely the 11 Archimedean tilings.*

A tiling $\mathcal{T}$ by regular polygons is called *equitransitive* if each set of mutually congruent tiles forms one transitivity class with respect to the symmetry group of $\mathcal{T}$. All but one Archimedean tilings are equitransitive; the exception is $(3^4.6)$, which has two congruence classes but three transitivity classes of tiles. There are many further equitransitive tilings of both kinds, edge-to-edge or not edge-to-edge.

The Laves tilings, named after the crystallographer Laves, are monohedral tilings related to the Archimedean tilings by duality. To explain this, let $\mathcal{T}$ be a monohedral edge-to-edge tiling in which the tiles are convex r-gons. Call a vertex x of $\mathcal{T}$ *regular* if the edges emanating from x dissect a small neighbourhood of x into equiangular parts, the angle being $2\pi/v$ where v is the valence of x. If all vertices of $\mathcal{T}$ are regular and the vertices in each tile have valences $v_1, \ldots, v_r$ in $\mathcal{T}$, then we denote $\mathcal{T}$ by the symbol $[v_1.v_2.\ \ldots\ .v_r]$. Again, the symbol is unique up to a cyclic permutation of its entries. The *Laves tilings* then are the eleven tilings described by the next theorem and depicted in Fig. 7:

Theorem 9. *(a) If* $\mathcal{T}$ *is a monohedral edge-to-edge tiling of the plane by convex polygons which has only regular vertices, then its symbol is one of the 11 symbols mentioned in part b. (b) To each of the eight symbols* $[3^4.6]$, $[3^2.4.3.4]$, $[3.6.3.6]$, $[3.4.6.4]$, $[3.12^2]$, $[4.6.12]$, $[4.8^2]$, *and* $[6^3]$ *corresponds a unique such tiling* $\mathcal{T}$*; to each of* $[3^3.4^2]$ *and* $[4^4]$ *corresponds a family of such tilings depending on a single real-valued parameter; and to* $[3^6]$ *corresponds a family of tilings with two such parameters.*

The Laves tiling $[v_1.v_2.\ \ldots\ .v_r]$ and the Archimedean tiling $(v_1.v_2.\ \ldots\ .v_r)$ correspond to each other by duality. The Laves tilings are isohedral; this is analogous to the Archimedean tilings being uniform.

B. Euler's Theorem

The well-known Euler Theorem for convex polytopes in ordinary space E^3 says that, if a convex polytope has v vertices, e edges, and f two-dimensional faces, then:

$$v-e+f=2$$

In this section we discuss its analogue and relatives for normal tilings $\mathcal{T}$ in the plane.

The Euler-type theorem in the plane now involves the two limits:

$$v(\mathcal{T})=\lim_{r\to\infty}\frac{v(x,r)}{t(x,r)}\qquad\text{and}\qquad e(\mathcal{T})=\lim_{r\to\infty}\frac{e(x,r)}{t(x,r)}$$

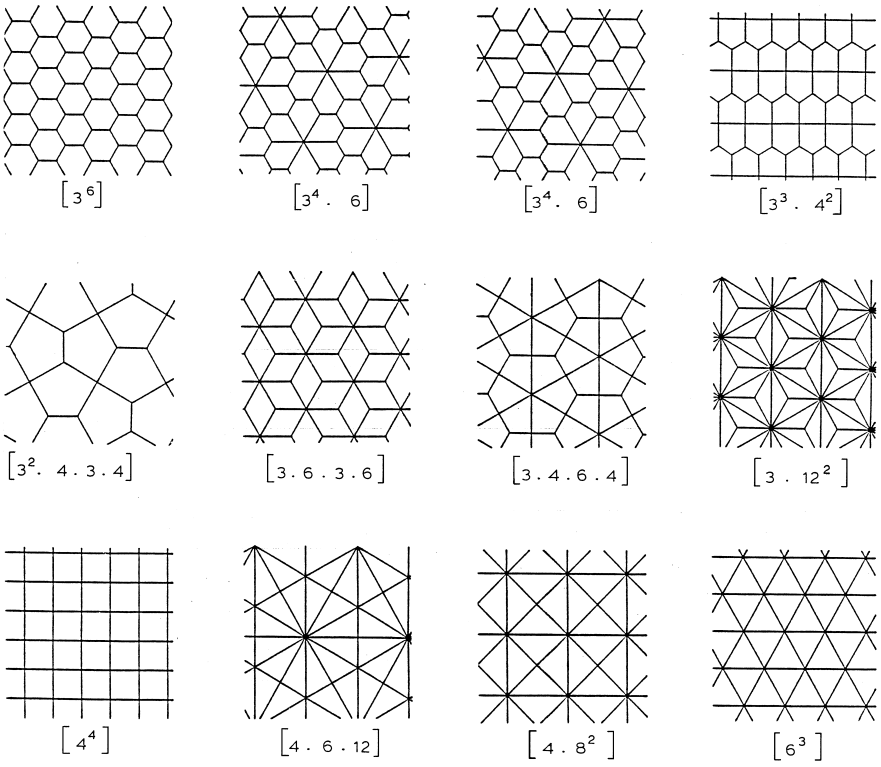


FIGURE 7 The 11 Laves tilings in the plane. [From Grünbaum, B., and Shephard, G. C. (1986). *Tilings and Patterns*, Freeman & Co., San Francisco.]

where x is a point in E^2, and $t(x, r)$, $e(x, r)$, and $v(x, r)$ denote the numbers of tiles, edges, and vertices, respectively, in the spherical patch $P(x, r)$. If these limits exist and are finite for some reference point x, then so for every point x in E^2, and the limits are independent of the reference point. Tilings for which these limits exist and are finite are called *balanced tilings*.

The following is known as *Euler's Theorem for Tilings*:

Theorem 10. *For a normal plane tiling $\mathcal{T}$, if one of the limits $v(\mathcal{T})$ or $e(\mathcal{T})$ exists and is finite, then so does the other. In particular, $\mathcal{T}$ is balanced and*

$$v(\mathcal{T}) - e(\mathcal{T}) + 1 = 0$$

For periodic tilings $\mathcal{T}$, this theorem can be restated as follows. Let L be the two-dimensional lattice of vectors that correspond to translations in $S(\mathcal{T})$. Choose any fundamental parallelogram P for L such that no vertex of $\mathcal{T}$ lies on a side of P, and no corner of P lies on an edge of $\mathcal{T}$. Let V, E, and T denote the numbers of vertices, edges, and tiles in P, respectively, where fractions of edges and tiles are counted appropriately. Then these numbers do not depend on the choice of the fundamental parallelogram.

The following result is known as *Euler's Theorem for Periodic Plane tilings*, and corresponds to an Euler-type theorem for tessellations of the two-dimensional torus:

Theorem 11. *If $\mathcal{T}$ is a normal plane tiling that is periodic, then $V - E - T = 0$.*

Euler's Theorem can be extended in various ways. One such extension involves the limits,

$$v_j(\mathcal{T}) = \lim_{r\to\infty} \frac{v_j(x, r)}{t(x, r)} \quad \text{and} \quad t_k(\mathcal{T}) = \lim_{r\to\infty} \frac{t_k(x, r)}{t(x, r)}$$

where $v_j(x, r)$ is the number of vertices of valence j in the patch $P(x, r)$, and $t_k(x, r)$ is the number of tiles with k adjacents in $P(x, r)$. A normal tiling $\mathcal{T}$ is called *strongly balanced* if, for some point x, all the limits $v_j(\mathcal{T})$ $(j \geq 3)$ and $t_k(\mathcal{T})$ $(k \geq 3)$ exist; then they will be finite by normality. As before, this definition does not depend on the reference point x. In any strongly balanced tiling, $v(\mathcal{T}) = \sum_{j\geq 3} v_j(\mathcal{T}) \leq \infty$ and $\sum_{k\geq 3} t_k(\mathcal{T}) = 1$; in particular, such a tiling is balanced. The following are examples of strongly balanced plane tilings: periodic tilings; tilings in which each tile has k vertices, with valences $j_1, \ldots, j_k$ (in some order); and tilings with j-valent vertices each incident with tiles that have $k_1, \ldots, k_j$ adjacents (in some order).

The equations in the next two theorems involve the relative frequences of vertices of various valences and tiles with various number of adjacents.

Theorem 12. *If $\mathcal{T}$ is a strongly balanced plane tiling, then:*

$$2\sum_{j\geq 3}(j-3)v_j(\mathcal{T})+\sum_{k\geq 3}(k-6)t_k(\mathcal{T})=0$$

$$\sum_{j\geq 3}(j-4)v_j(\mathcal{T})+\sum_{k\geq 3}(k-4)t_k(\mathcal{T})=0$$

$$\sum_{j\geq 3}(j-6)v_j(\mathcal{T})+2\sum_{k\geq 3}(k-3)t_k(\mathcal{T})=0$$

Theorem 13. *If $\mathcal{T}$ is a strongly balanced plane tiling, then:*

$$\frac{1}{\sum_{j\geq 3} j w_j(\mathcal{T})}+\frac{1}{\sum_{k\geq 3} k t_k(\mathcal{T})}=\frac{1}{2}$$

where $w_j(\mathcal{T}) := v_j(\mathcal{T})/v(\mathcal{T})$.

Note that the denominators on the left in Theorem 13 are the average valence of the vertices of $\mathcal{T}$ and the average number of edges of the tiles of $\mathcal{T}$. For example, for the regular tessellation of the plane by regular p-gons, q meeting at each vertex, the latter equation takes the simple form $\frac{1}{p}+\frac{1}{q}=\frac{1}{2}$. From this equation we obtain the well-known solutions $\{p, q\}=\{3, 6\}$, $\{4, 4\}$, and $\{6, 3\}$, which correspond to the tessellations by triangles, squares, or hexagons, respectively.

C. Classification by Symmetry

In this section we discuss the classification (enumeration) of plane tilings $\mathcal{T}$ by combinatorial or metrical (euclidean) symmetry. The tiles are as usual topological discs, but convexity of the tiles is not assumed. We begin with the coarser classification by combinatorial symmetry which employs the combinatorial automorphism group $\Gamma(\mathcal{T})$ and its transitivity properties on $\mathcal{T}$. As we remarked earlier, the concepts of combinatorial equivalence and topological equivalence are the same for normal plane tilings. Any automorphism in $\Gamma(\mathcal{T})$ can be realized by a homeomorphism of the plane that preserves $\mathcal{T}$, and vice versa. The classification by combinatorial symmetry is thus the same as the classification by topological symmetry.

1. Combinatorial Symmetry

Combinatorial symmetry can occur locally or globally. In a locally symmetric tiling, the combinatorial data for the neighborhoods of tiles or vertices are all the same. In a globally symmetric tiling $\mathcal{T}$, this is also true, but now the symmetry is furnished by a combinatorial automorphism group $\Gamma(\mathcal{T})$, which acts transitively on the tiles or vertices of the tiling. We first discuss symmetry properties with respect to tiles.

Let $\mathcal{T}$ be a plane tiling, and let T be a tile of $\mathcal{T}$. With T we can associate its *valence-type* $j_1.j_2.\ \ldots\ .j_k$ provided T has k vertices which, in cyclic order, have valences $j_1, j_2, \ldots, j_k$. Then $\mathcal{T}$ is said to be *homogeneous of type* $[j_1.j_2.\ \ldots\ .j_k]$ if $\mathcal{T}$ is normal and each tile of $\mathcal{T}$ has valence-type $j_1.j_2.\ \ldots\ .j_k$. As before, these symbols for $\mathcal{T}$ and its tiles are unique up to a cyclic permutation of the entries.

A plane tiling $\mathcal{T}$ is called *homeohedral* or *combinatorially tile-transitive* if $\mathcal{T}$ is normal and $\Gamma(\mathcal{T})$ acts transitively on the tiles of $\mathcal{T}$. Each homeohedral tiling is also homogeneous.

Theorem 14. (a) *If $\mathcal{T}$ is a homogeneous plane tiling, then it has one of the 11 types* $[3^6]$, $[3^4.6]$, $[3^3.4^2]$, $[3^2.4.3.4]$, $[3.4.6.4]$, $[3.6.3.6]$, $[3.12^2]$, $[4^4]$, $[4.6.12]$, $[4.8^2]$, $[6^3]$. (b) *Each homogeneous plane tiling is homeohedral. Any two homogeneous plane tilings of the same type are combinatorially equivalent, and each type can be represented by a Laves tilings (see Theorem 9 and Fig. 7).*

We now discuss the dual concept. A tiling $\mathcal{T}$ is called *homeogonal* or *combinatorially vertex-transitive* if $\mathcal{T}$ is normal and $\Gamma(\mathcal{T})$ acts transitively on the vertices of $\mathcal{T}$. As with uniform tilings, we can associate with a homeogonal tiling $\mathcal{T}$ its *type* $(k_1.k_2.\ \ldots\ .k_j)$; here, j is the valence of any vertex x, and the j tiles that contain x have, in a suitable cyclic order, k_1 vertices, k_2 vertices, and so on. If $\mathcal{T}$ and $\mathcal{T}'$ are normal dual tilings, then $\mathcal{T}$ is homeogonal if and only if $\mathcal{T}'$ is homeohedral.

Theorem 15. (a) *If $\mathcal{T}$ is a homeogonal plane tiling, then it has one of the 11 types* (3^6), $(3^4.6)$, $(3^3.4^2)$, $(3^2.4.3.4)$, $(3.4.6.4)$, $(3.6.3.6)$, (3.12^2), (4^4), $(4.6.12)$, (4.8^2), (6^3). (b) *Any two homeogonal plane tilings of the same type are combinatorially equivalent, and each type can be represented by an Archimedean (uniform) tiling (see Theorems 7 and 8, and Fig. 4).*

A normal plane tilings $\mathcal{T}$ is said to be *homeotoxal* if its automorphism group $\Gamma(\mathcal{T})$ is edge transitive. It can be shown that there are just five "types" of homeotoxal tilings.

Normality of the tilings is essential in all these classifications. If the requirement of normality is dropped, then many further possibilities arise.

2. Metrical Symmetry

Before we investigate tilings, we introduce some terminology that applies in the wider context of classifying certain geometric objects with respect to metrical symmetries or, as we will say, *by homeomerism*. In particular, this will explain when two geometric objects are considered to be the same (of the same homeomeric type) with respect to classification purposes.

Let $\mathcal{R}$ and $\mathcal{R}'$ be two geometric objects in Euclidean space E^d, and let $S(\mathcal{R})$ and $S(\mathcal{R}')$ be their symmetry groups, respectively. Consider a homeomorphism φ of E^d which maps $\mathcal{R}$ onto $\mathcal{R}'$. Then φ is said to be *compatible*

with a symmetry σ in $S(\mathcal{R})$ if there exists a symmetry σ' in $S(\mathcal{R}')$ such that $\sigma'\varphi = \varphi\sigma$; that is, up to the one-to-one correspondence determined by the homeomorphism φ, a symmetry σ moves elements of the object $\mathcal{R}$ in the same way as σ' moves elements of $\mathcal{R}'$. We call φ *compatible with* $S(\mathcal{R})$ if φ is compatible with each symmetry σ in $S(\mathcal{R})$.

Two geometric objects $\mathcal{R}$ and $\mathcal{R}'$ in E^d are said to be *homeomeric*, or *of the same homeomeric type*, if there exists a homeomorphism φ which maps $\mathcal{R}$ onto $\mathcal{R}'$ such that φ is compatible with $S(\mathcal{R})$ and its inverse φ^{-1} is compatible with $S(\mathcal{R}')$.

These concepts apply to the classification of normal plane tilings $\mathcal{T}$ with certain transitivity properties of their symmetry group $S(\mathcal{T})$. The following three results summarize the enumeration of all homeomeric types of tilings with a tile-transitive, vertex-transitive, or edge-transitive symmetry group, respectively, yielding the homeomeric types of isohedral, isogonal, or isotoxal tilings:

Theorem 16. *There exist precisely 81 homeomeric types of normal isohedral plane tilings. Precisely 47 types can be realized by a normal isohedral edge-to-edge tiling with convex polygonal tiles.*

Theorem 17. *There exist precisely 91 homeomeric types of normal isogonal plane tilings. Precisely 63 types can be realized by normal isogonal edge-to-edge tilings with convex polygonal tiles.*

Theorem 18. *There exist precisely 26 homeomeric types of normal isotoxal plane tilings. Precisely six types can be realized by a normal isotoxal edge-to-edge tiling with convex polygonal tiles.*

The enumeration of the tilings can be accomplished in various ways. One approach is through *incidence symbols* which encode data about the local structure of the plane tilings. For example, isohedral tilings fall into 11 combinatorial classes, typified by the Laves tilings. In an isohedral tiling, every tile is surrounded in the same way, and its vertex degree sequence is given by the symbol for the corresponding Laves tiling. The incidence symbol records (in terms of edge labels and edge orientations) how each tile meets its (congruent) neighbors. The situation is similar for isogonal and isotoxal tilings. The enumeration proceeds by identifying all possible incidence symbols.

Another algorithmic approach is through *Delaney–Dress symbols*. With every tiling is associated a "barycentric subdivision," and the symbol now stores information about the way in which the symmetry group acts on this subdivision. Then the enumeration of the tilings amounts to the enumeration of all Delaney–Dress symbols of a certain type. This approach is more flexible and allows generalizations to tilings in higher dimensions. For example, the method has been applied to show that there are 88 combinatorial classes of periodic tilings in ordinary three-space for which the symmetry group acts transitively on the two-dimensional faces of the tiling.

V. MONOHEDRAL TILINGS

This section deals with monohedral tilings $\mathcal{T}$ in Euclidean spaces E^d of any dimension $d \geq 2$. Monohedral tilings have a single prototile. To determine which shapes occur as prototiles is one of the main open problems in tiling theory. The answer is not even known for polygonal prototiles in the plane, which necessarily cannot have more than six vertices. In its full generality, the prototile enumeration problem seems to be intractable in higher dimensions, so suitable restrictions must be imposed on the tiles or on the kind of tilings.

A. Lattice Tilings

Lattice tilings in E^d by convex d-polytopes have been widely studied. The interest in such tilings originated from crystallography and the geometry of numbers. Early contributions on the subject can be found in the works of Dirichlet, Fedorov, Minkowski, Voronoi, and Delone.

A tiling $\mathcal{T}$ by translates of a single prototile T is the simplest kind of monohedral tiling. Prototiles which admit such a tiling are very restricted. For example, if the prototile T is a convex d-polytope, then T and all its facets must necessarily be centrally symmetric. The translation vectors for the tiles in a tiling by translates need not necessarily come from a lattice. However, if a prototile tiles by translation, then it is natural to ask if it also admits a lattice tiling.

Theorem 19. *If a convex d-polytope T tiles E^d by translation, then T also admits (uniquely) a face-to-face lattice tiling of E^d and thus is a parallelotope.*

For nonconvex prototiles T, the analogue of Theorem 19 is generally not true. There are nonconvex star-shaped polyhedral sets T which tile E^d by translation but do not admit a lattice tiling. Examples have been found in almost all dimensions d, including $d = 3$.

Parallelotopes are the most basic space fillers. Simple examples are the Voronoi regions for lattices. The previous theorem is based on the following characterization of parallelohedra. A *belt* of a convex d-polytope T is a sequence of facets $F_0, F_1, \ldots, F_{k-1}, F_k = F_0$ of T, such that $F_{i-1} \cap F_i$ is a $(d-2)$-face of T for each $i = 1, \ldots, k$, and all these $(d-2)$-faces are parallel. If T is a centrally symmetric convex d-polytope with centrally symmetric facets, then every $(d-2)$-face G of T determines a belt of facets whose $(d-2)$-faces are parallel to G.

Theorem 20. *A convex d-polytope T is a parallelotope if and only if T and its facets are centrally symmetric and each belt of T contains four or six facets.*

In each dimension d, there is only a finite number of distinct combinatorial types of parallelotope. This follows from Minkowski's observation that a parallelotope in E^d can have at most $2(2^d - 1)$ facets. A complete list of parallelotopes is only known for $d \leq 4$. For $d = 2$, 3, or 4, the number of distinct combinatorial types of parallelotopes is 2, 5, or 52, respectively. Figure 8 shows representatives for the parallelotopes in the plane and ordinary space. In the plane, the parallelogram and the centrally symmetric hexagon are the only possible shapes. For $d = 3$, Minkowski's bound predicts that a parallelotope cannot have more than 14 facets; the maximum of 14 is attained by the truncated octahedron, which is the last parallelotope in Fig. 8.

Call a parallelotope T *primitive* if, in its unique face-to-face tiling $\mathcal{T}$, each vertex of $\mathcal{T}$ is contained in exactly $d + 1$ tiles. In E^2 and E^3, the centrally symmetric hexagon and truncated octahedron are the only types of primitive parallelotopes; in E^4, there are three distinct combinatorial types. The total number of combinatorial types of parallotopes grows very fast in dimensions $d \geq 5$, and the number of primitive types is known to be at least 223 if $d = 5$. It has been conjectured (already by Voronoi) that each parallelotope is in fact an affine image of the Voronoi region for some lattice. This is known to be true in dimensions $d \leq 4$, as well as for primitive parallelohedra in all dimensions.

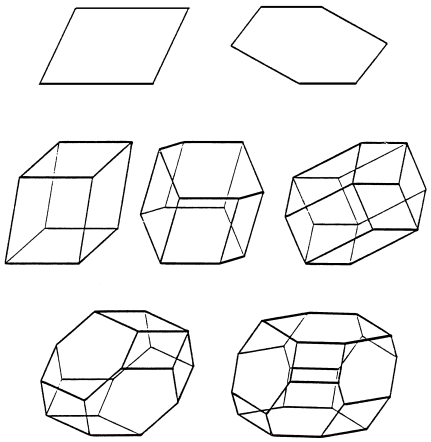


FIGURE 8 The parallelotopes in two and three dimensions. The top row shows the two possible shapes in the plane, a parallelogram, and a centrally symmetric hexagon. There are five shapes of parallelotopes in ordinary space—the parallelopiped, hexagonal prism, and rhombic dodecahedron shown in the middle row and the elongated dodecahedron and truncated octahedron shown in the bottom row. [From Schulte, E. (1993). *Handbook of Convex Geometry*, (Gruber, P. M., and Wills, J. M., eds.). Elsevier, Amsterdam.]

The d-dimensional cube obviously is a parallelotope in each dimension d. In every lattice tiling of E^d by unit d-cubes, there always is a "stack" of cubes in which each two adjacent cubes meet in a whole facet. However, for $d \geq 10$, there are nonlattice tilings of E^d by translates of unit d-cubes in which no two cubes share a whole facet. On the other hand, for $d \leq 6$, any tiling of E^d by unit d-cubes must contain at least one pair of cubes that share a whole facet.

B. Space Fillers in Dimension *d*

Convex d-polytopes which are prototiles of monohedral tilings in E^d are called *space fillers* of E^d, or *plane fillers* if $d = 2$. The enumeration of all space fillers is far from being complete, even for the planar case.

Isohedral tilings are monohedral tilings with a tile-transitive symmetry group. Convex polytopes which are prototiles of isohedral tilings are called *stereohedra*. Each stereohedron is a space filler, but the converse is not true. Indeed, a famous problem by Hilbert (posed in 1900) asked whether there exists a three-dimensional polyhedral shape that admits a monohedral tiling but no isohedral tiling of E^3. Such a prototile (in any dimension) is said to be *anisohedral*. Rephrased in terms of crystallographic groups, Hilbert's problem asked whether there exists a polyhedral space-filling tile in E^3 which is not a fundamental region for a discrete group of Euclidean isometries. Hilbert's question can be reduced to the planar case by observing that three-dimensional (and higher dimensional) anisohedral prototiles can be constructed as prisms over anisohedral planar prototiles. Several authors have discovered examples of anisohedral prototiles, beginning with Reinhardt in 1928 and Heesch in 1935. In the plane, there are anisohedral convex pentagons which admit periodic monohedral tilings. The tile in Fig. 9 is

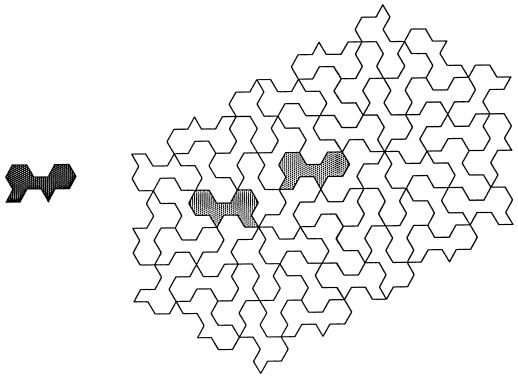


FIGURE 9 A planar anisohedral tile and its unique plane tiling. The two shaded tiles are surrounded in different ways. [From Schattschneider, D., and Senechal, M. (1997). *Handbook of Discrete and Computational Geometry*, (Goodman, J. E., and O'Rourke, J., eds.). CRC Press, Boca Raton, FL.]

anisohedral and admits a unique plane tiling; in fact, the tiling must be nonisohedral because the two shaded tiles are surrounded in different ways.

Theorem 21. *For every $d \geq 2$, there exist anisohedral space fillers that are convex d-polytopes.*

Isohedrality is a global property that a monohedral tiling may or may not have. Our next theorem is related to Hilbert's problem and describes a way to detect isohedrality of a tiling locally. We first introduce some terminology.

Let $\mathcal{T}$ be a monohedral tiling in E^d. Given a tile T of $\mathcal{T}$, the *kth corona* $C^k(T)$ of T in $\mathcal{T}$ is the set of all tiles T' in $\mathcal{T}$ for which there exists a sequence of tiles $T = T_0, T_1, \ldots, T_{m-1}, T_m = T'$ with $m \leq k$ such that $T_i \cap T_{i+1} \neq \emptyset$ for $i = 0, \ldots, m-1$. So $C^0(T)$ consists of T itself, $C^1(T)$ of the tiles that meet T, and $C^k(T)$ of the tiles that meet a tile in $C^{k-1}(T)$. It is possible that two distinct tiles T and T' of $\mathcal{T}$ have the same kth corona for some k; that is, $C^k(T) = C^k(T')$. A *centered kth corona* in $\mathcal{T}$ is a pair $(T, C^k(T))$, consisting of a tile T and its *kth* corona $C^k(T)$. A centered corona determines the original tile uniquely. Two centered kth coronas $(T, C^k(T))$ and $(T', C^k(T'))$ are said to be *congruent* (as centered coronas) if there exists an isometry γ of E^d such that $\gamma(T) = T'$ and $\gamma(C^k(T)) = C^k(T')$; here, it is not required that γ maps $\mathcal{T}$ onto itself. By $S_k(T)$ we denote the group of isometries of E^d which map the centered corona $(T, C^k(T))$ onto itself; this is the subgroup of the symmetry group $S(T)$ of the tile T which maps $C^k(T)$ onto itself. Since $S(T)$ is a finite group, the chain of subgroups $S(T) = S_0(T) \supseteq S_1(T) \supseteq \cdots \supseteq S_k(T) \supseteq \cdots$ can only contain a finite number of distinct groups.

We now have the following *Local Theorem for Tilings*, which characterizes isohedrality locally:

Theorem 22. *Let $\mathcal{T}$ be a monohedral tiling of E^d. Then $\mathcal{T}$ is isohedral if and only if there exists a positive integer k for which the following two conditions hold: first, any two centered kth coronas in $\mathcal{T}$ are congruent, and, second, $S_{k-1}(T) = S_k(T)$ for some tile (and hence all tiles) T of $\mathcal{T}$. Moreover, if the two conditions hold for k, and if T is a tile of $\mathcal{T}$, then $S_k(T)$ is the stabilizer of T in $S(\mathcal{T})$. In particular, if the prototile of $\mathcal{T}$ is asymmetric (that is, has no nontrivial symmetry), then $\mathcal{T}$ is isohedral if and only if the first condition holds with $k = 1$.*

The classification of the space fillers is one of the big open problems in tiling theory. Its complexity is already evident in the planar case, which is still unsettled. Many authors who attempted the classification for the plane believed, or even stated explicitly, that their enumeration is complete. Often this was generally accepted until a new prototile was discovered by someone else.

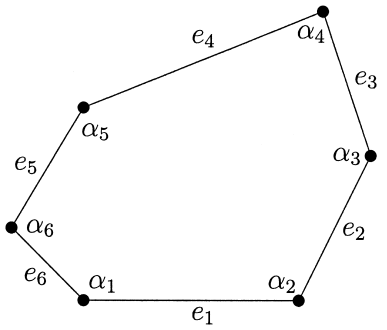


FIGURE 10 A convex hexagon with sides $e_1, \ldots, e_6$ and angles $\alpha_1, \ldots, \alpha_6$. [From Senechal, M. (1995). *Quasicrystals and Geometry*, Cambridge University Press, Cambridge.]

A plane filler must necessarily be a triangle, quadrangle, pentagon or hexagon. At present, the list of convex plane fillers comprises all triangles and all quadrangles, as well as 14 kinds of pentagons and three kinds of hexagons. To see how each triangle can tile, join two copies to form a parallelogram, and then tile the plane with congruent parallelograms. Similarly, two copies of a convex quadrangle can be joined along an edge to form a hexagon of which congruent copies tile the plane. Any convex pentagon with a pair of parallel sides is a plane filler. To describe the three kinds of hexagons that tile, label the sides of the hexagon by $e_1, \ldots, e_6$ and the angles by $\alpha_1, \ldots, \alpha_6$, as in Fig. 10. Then a convex hexagon tiles if and only if it satisfies one of the following conditions:

- $\alpha_1 + \alpha_2 + \alpha_3 = \alpha_4 + \alpha_5 + \alpha_6 = 2\pi,\ e_6 = e_3$
- $\alpha_1 + \alpha_2 + \alpha_4 = \alpha_3 + \alpha_5 + \alpha_6 = 2\pi,\ e_6 = e_3,\ e_4 = e_2$
- $\alpha_1 = \alpha_3 = \alpha_5 = 2\pi/3,\ e_6 = e_1,\ e_3 = e_2,\ e_5 = e_4$

Among these convex plane fillers, all the triangles, quadrangles, and hexagons and exactly five kinds of the pentagons admit isohedral tilings. No other convex polygons can admit isohedral plane tilings.

There is a wealth of further interesting monohedral plane tilings whose prototiles are nonconvex shapes. Examples are tilings by polyominos or polyhexes, respectively, which are composed of squares or hexagons of the regular square or hexagonal tiling of the plane. Noteworthy is also the existence of monohedral *spiral* plane tilings (see Fig. 3), whose prototiles have the remarkable property that two copies of it can completely surround a third.

The space-filler problem is most appealing in three dimensions. Many examples of space fillers have been discovered by crystallographers as Voronoi regions for suitable discrete point sets (dot patterns) in E^3. In the past, there have been several contradictory claims in the literature as to how many facets a three-dimensional space filler can have. The current world record is held by a spectacular space filler with 38 facets, which is a Voronoi region

for some discrete point set (discovered using computers). Among the five Platonic solids (the regular tetrahedron, cube, octahedron, dodecahedron, and icosahedron), only the cube tiles ordinary space, in an obvious manner.

The classification of space fillers becomes even more difficult in higher dimensions. Not even the parallelotopes, which are the convex polytopes that tile by translation, have been completely enumerated for $d \geq 5$. Tilings by certain kinds of simplices have been studied, but it is still not known, even in three dimensions, which simplices tile space. There are several kinds of tetrahedra which do admit monohedral tilings of E^3, among them tetrahedra that are fundamental regions for discrete groups generated by reflections.

A more modest problem is the classification of all combinatorial types of space-filling convex polytopes in E^d. This problem is trivial for the plane, where triangles, quadrangles, pentagons, and hexagons are the only possible solutions. For dimensions $d \geq 3$ it is not even known if there are only finitely many combinatorial types of space-filling polytopes. The number of types would be finite if there was an upper bound on the number of facets of space-filling polytopes, but the existence of a general bound has not been established. The only general result available in the literature is the following theorem which provides an upper bound for the number of facets of stereohedra that admit isohedral face-to-face tilings:

Theorem 23. *If a stereohedron admits an isohedral face-to-face tiling of* E^d, *then its number of facets is bounded by* $2^d(h - \frac{1}{2}) - 2$, *where* h *is the index of the translation subgroup in the symmetry group of the tiling.*

The theorem gives a finite bound for the number of facets of d-dimensional stereohedra that admit an isohedral face-to-face tiling. In fact, the symmetry group of the tiling must be among the finitely many crystallographic groups in E^d, and so we know that the index h of the translation subgroup is uniformly bounded. In three dimensions, the maximum index is 48 and the bound for the number of facets is 378; it is likely that the true bound is considerably lower, possibly as low as 38.

The strong requirement of congruence of the tiles considerably restricts the various possibilities for designing monohedral tilings. If this requirement is relaxed to combinatorial equivalence, there is generally much freedom for choosing the metrical shape of the tiles to arrive at a tiling of the whole space. Call a tiling $\mathcal{T}$ of E^d by convex polytopes *monotypic* if each tile of $\mathcal{T}$ is combinatorially equivalent to a convex d-polytope T, the *combinatorial prototile* of $\mathcal{T}$. In a monotypic tiling, there generally are infinitely many different metrical shapes of tiles, but the tiles are all convex and combinatorially equivalent to a single combinatorial prototile. It is not difficult to see that the plane admits a monotypic face-to-face tiling $\mathcal{T}$ by convex n-gons for each $n \geq 3$. Figure 1 illustrates how such a tiling can be constructed for $n = 7$. It is a consequence of Euler's Theorem that these tilings cannot be normal if $n \geq 7$, and so infinitely many different shapes of n-gons are needed in the process.

Much less obvious are the following rather surprising results about monotypic tilings in E^3:

Theorem 24. *Every convex* 3*-polytope* T *is the combinatorial prototile of a monotypic tiling* $\mathcal{T}$ *of* E^3. *If all facets of* T *are triangles, then* $\mathcal{T}$ *can be made face-to-face.*

In other words, in three dimensions, every polytope tiles, as long as combinatorially equivalent (rather than congruent) copies are allowed. In general, these tilings will not be face-to-face. In fact, for each dimension $d \geq 3$, there are convex d-polytopes which cannot occur as combinatorial prototiles of monotypic face-to-face tilings of E^d.

VI. NONPERIODIC TILINGS

Tiling models are an important tool in the geometrical study of crystals. Crystal growth is a modular process: from a relatively tiny cluster (seed) of atoms, a crystal grows by the accretion of modules (atoms, molecules). In the geometric modeling of crystal structure, these modules are sometimes represented by space-filling polyhedra yielding a tiling of ordinary space. For nearly 200 years, until 1984, it had been an *axiom* of crystallography that the internal structure of a crystal was periodic. The presence of lattice structure in a crystal had been strongly supported by the geometry of the X-ray diffraction images of the crystal, but this "fundamental law" of crystal structure was overturned in 1984 when an alloy of aluminum and manganese was discovered that had diffraction images exhibiting icosahedral symmetry; pentagonal, and thus icosahedral, symmetry was known not to be compatible with lattice structure. Soon other crystal-like structures were found, and eventually the name "quasicrystals" emerged.

Since 1984, crystallography has gone through a process of redefining itself, as well as the objects it studies. In the world of crystals and quasicrystals, order is no longer synonymous with periodicity, but exactly how far one has to go beyond periodicity is still a big open problem. Tilings once again have become an important modeling tool, but now the interest is in nonperiodic tilings and in shapes that do not admit periodic tilings.

A. Aperiodicity

Aperiodicity is a fascinating phenomenon in tiling theory. Recall that a tiling $\mathcal{T}$ in Euclidean d-space E^d is called *nonperiodic* if it does not admit a nontrivial translational

symmetry. For example, the spiral tiling shown in Fig. 3 is nonperiodic. There are many shapes in E^d that admit both periodic tilings and nonperiodic tilings. For example, already the d-dimensional cube admits tilings whose symmetry group contains translational symmetries in any preassigned number k of independent directions, with $k = 1, \ldots, d$. In fact, if $\mathcal{S}$ is any finite set of prototiles in E^d which are convex polytopes, and if $\mathcal{S}$ admits a face-to-face tiling of E^d which possesses a nontrivial translational symmetry, then $\mathcal{S}$ also admits a tiling of E^d which is periodic (see Theorem 3).

A set $\mathcal{S}$ of prototiles in E^d is said to be *aperiodic* if $\mathcal{S}$ admits a tiling of E^d, but all such tilings are nonperiodic. In discussing aperiodic prototile sets it is convenient to generalize the notion of a tiling and allow decorations or markings on the tiles. This leads to *decorated tiles* and *decorated tilings*, respectively. Here we shall not investigate these concepts in full generality but restrict ourselves to applications in the theory of nonperiodic tilings. In a typical application in two dimensions, the tiles in a prototile set $\mathcal{S}$ are polygons with some corners colored or with orientations on some sides (that is, the colors and orientations are the decorations). The condition then is that, in constructing a tiling with the given set of decorated prototiles, we must place tile against tile in an edge-to-edge manner, such that colored corners are placed against colored corners with the same color, and oriented sides are placed against oriented sides with the same orientation; in short, the colors at corners and the orientations on sides must match. Figure 11 shows the Penrose aperiodic prototile set consisting of two decorated tiles known as *kite* and *dart*; here the corners are colored with two colors, black and white (say), as shown, but there are no orientations given to the sides. We elaborate on this example in the next subsection.

Decorated tiles and matching rules are the price we have to pay if we want tiles of simple geometric.
By altering the shape of the decorated tiles in a prototile set to reflect their decorations and the matching rules, we can often construct a prototile set with nondecorated tiles whose tiling properties are equivalent to those of the set of decorated tiles. However, the geometric shape of the nondecorated tiles is generally more complicated and less appealing than the underlying shape of the corresponding decorated tiles.

There is a close connection between aperiodicity and the Undecidability Theorem for tilings (see Theorem 1). In fact, the Undecidability Theorem was originally proved by Berger in 1966 by establishing the existence of an aperiodic set consisting of 20,426 so-called *Wang tiles*; these are square tiles with colored edges, which must be tiled in an edge-to-edge manner by translation only such that colors of adjacent tiles match. The existence of such a set is enough to ensure that no algorithm can exist which, for any given finite set $\mathcal{S}$ of prototiles, decides in a finite number of steps whether $\mathcal{S}$ admits a tiling or not. The number of tiles in an aperiodic set of Wang tiles has since then been reduced considerably.

It is still an open question whether there exists an aperiodic set in the plane consisting of a single decorated or nondecorated tile. Over the years, many examples of small aperiodic sets have been discovered, both in the plane and in higher dimensions. The most famous examples are the three Penrose aperiodic sets in the plane described in the next subsection; they have greatly influenced the study of aperiodicity in higher dimensional spaces. There is a pair of decorated aperiodic tiles in any dimension $d \geq 2$. There are also planar examples with only three nondecorated tiles (one hexagon and two pentagons) that force aperiodicity. In E^3, the aperiodic pair of *Penrose rhombs* depicted in Fig. 12 has been generalized to yield families of decorated (Ammann) rhombohedra which again only admit non-periodic tilings. There also exists an aperiodic set of only four decorated tetrahedral tiles in E^3 that admit tilings with global icosahedral symmetry.

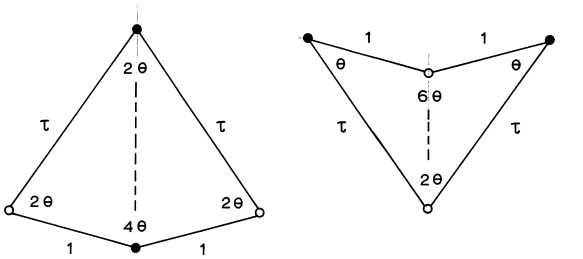


FIGURE 11 The Penrose kite and dart. The corners are colored black or white (solid or open circles, respectively), and the angles and side lengths are as indicated, with $\tau = (1+\sqrt{5})/2$ and $\theta = \pi/5$. [From Grünbaum, B., and Shephard, G. C. (1986). *Tilings and Patterns*, Freeman & Co., San Francisco.]

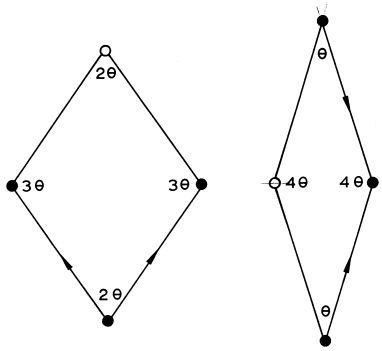


FIGURE 12 The Penrose rhombs. The rhombs have black and white corners (solid or open circles, respectively) and have orientations given to some of their sides; the angles are as indicated, with $\theta = \pi/5$. [From Grünbaum, B., and Shephard, G. C. (1986). *Tilings and Patterns*, Freeman & Co., San Francisco.]

A remarkable three-dimensional tile, known as the *Schmitt–Conway–Danzer tile*, has been discovered that is aperiodic when only direct congruent copies (that is, no mirror images) are allowed in tilings of E^3. The tile is a certain biprism with a cell-complex on its boundary, whose structure forces the tilings to have screw rotational symmetry through an irrational angle in one direction. The form of aperiodicity exhibited by this example is weaker than in the previous examples; in fact, although the tilings do not have translational symmetry, they still have a symmetry of infinite order.

B. The Penrose Tilings

A striking advance in tiling theory was Penrose's discovery in 1973 and 1974 of three aperiodic sets of decorated prototiles and matching rules in the plane, which are closely related to each other in that tilings with tiles from one set can be converted into tilings with tiles from another set. Two of these sets are depicted in Figs. 11 and 12; they are the Penrose kite and dart and the Penrose rhombs, both mentioned earlier. A *Penrose tiling* is a tiling of the entire plane that is constructed from any of the three sets by obeying the matching rules. Figure 13 shows a patch of a Penrose tiling by Penrose rhombs.

It is not difficult to make your own Penrose tiles from cardboard and build finite patches of tiles. However, even if you obey the matching rules, you will probably run into untileable regions at some stage, and you will have to remove or rearrange some tiles and try again. The decorations and matching rules are crucial in this process; in fact, the underlying geometric shapes do admit *periodic* tilings of the plane if the decorations and matching rules are ignored.

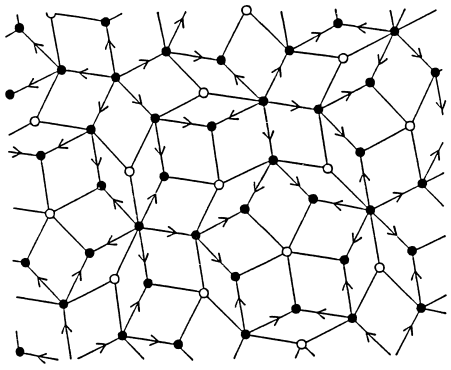


FIGURE 13 A Penrose tiling by Penrose rhombs. The tiling has been constructed according to the matching rule for Penrose tilings by Penrose rhombs (that is, the colors at corners and the orientations of sides of adjacent tiles must match). [From Grünbaum, B., and Shephard, G. C. (1986). *Tilings and Patterns*, Freeman & Co., San Francisco.]

The kite and dart of Fig. 11 have sides of two lengths in the ratio τ:1 (with $\tau = (1 + \sqrt{5})/2$ the golden ratio), and the angle θ is $\pi/5$. The corners are colored with two colors, black and white. The matching condition is that equal sides must be put together, so that the colors at the corners match. The two Penrose rhombs of Fig. 12 also have black and white corners and have orientations given to some of their sides; the angles are as indicated, where again $\theta = \pi/5$. The condition is that in constructing a tiling the colors of corners as well as lengths and orientations of sides must match.

Plane tilings by kites and darts can be transformed into plane tilings by rhombs, and vice versa. The transition from a kite and dart tiling to a rhomb tiling is illustrated in Fig. 14; the kites and darts are bisected, and the resulting triangles are merged into rhombs. The two tilings depicted in Fig. 14 are *mutually locally derivable*; that is, the tiles in either tiling can, through a process of decomposition into smaller tiles or regrouping with adjacent tiles, or a combination of both processes, form the tiles of the other. The concept of mutual local derivability links the tilings with one prototile set to those with another. Thus, characteristic properties of one prototile set translate into similar properties of the other set.

As for many other aperiodic prototile sets, the aperiodicity of the Penrose sets is based on two important transformations of tilings, namely *composition* and the corresponding inverse process of *decomposition*, which exhibit the existence of hierarchical structure. Figure 15 illustrates the composition process for a tiling by Penrose rhombs.

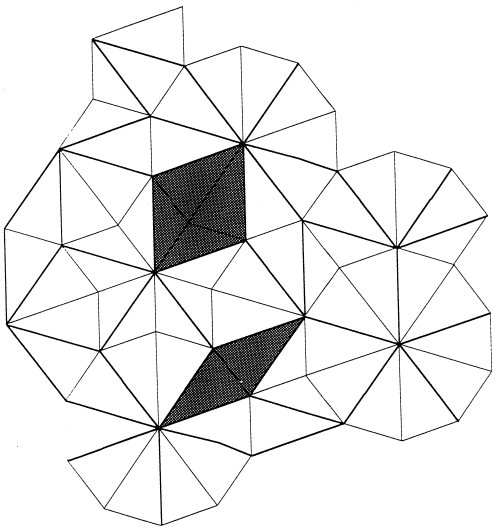


FIGURE 14 The transition from a Penrose kite and dart tiling to a Penrose rhomb tiling. The kites and darts of the original tiling are bisected, and the resulting triangles are merged into rhombs. [From Senechal, M. (1995). *Quasicrystals and Geometry*, Cambridge University Press. Cambridge.]

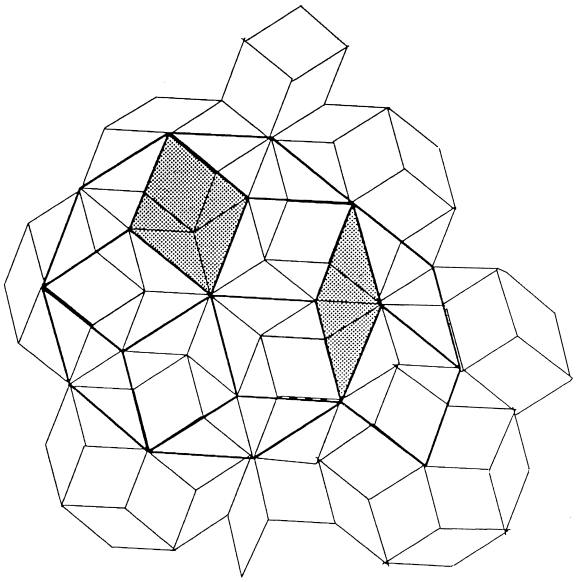


FIGURE 15 Composition for the Penrose rhombs. First, certain tiles of the original Penrose tiling are bisected to yield triangular tiles. Then, the triangular tiles and remaining rhombic tiles are regrouped with adjacent tiles to form the rhombic tiles of a new Penrose tiling of larger scale. [From Senechal, M. (1995). *Quasicrystals and Geometry*, Cambridge University Press. Cambridge.]

By composition we mean the process of taking unions of tiles (or pieces of tiles) to form larger tiles of basically the same shape as those of the original tiles, such that the decorations of the original tiles specify a matching condition equivalent to the original one. Decomposition is the basis for the process of *inflation* which, when iterated, can be used to generate arbitrarily large patches or even global tilings. Here inflation is the operation of, first, homothetically enlarging a given patch of tiles (in the case of the Penrose tiles, by some factor involving τ), and then decomposing it into tiles of the original size. Inflation and its inverse process, *deflation*, are characteristic features of many aperiodic sets and their tilings.

A tiling $\mathcal{T}$ in E^d is said to be *repetitive* if every bounded configuration of tiles appearing anywhere in $\mathcal{T}$ is repeated infinitely many times throughout $\mathcal{T}$; more precisely, for every bounded patch of $\mathcal{T}$ there exists an $r > 0$ such that every ball of radius r in E^d contains a congruent copy of the patch. Two tilings $\mathcal{T}_1$ and $\mathcal{T}_2$ are called *locally isomorphic* if for every bounded patch of $\mathcal{T}_1$ there exists a congruent copy of the patch in $\mathcal{T}_2$. The Penrose tilings of the plane are repetitive tilings, as are the nonperiodic tilings for many other aperiodic prototile sets. Furthermore, any two Penrose tilings with the same prototile set are locally isomorphic. So, for example, since there exist Penrose tilings of the plane with global (pentagonal) D_5-symmetry, there must be arbitrarily large finite patches with D_5-symmetry in any Penrose tiling with the same prototile set.

The remaining (but historically first) Penrose aperiodic set consists of six decorated prototiles and suitable matching rules.

C. The Projection Method

The projection method and its variants are important tools for the construction of nonperiodic tilings. These tilings are nondecorated.

In the *strip projection method*, the d-dimensional tilings are obtained as projection images of polyhedral surfaces that are embedded in a Euclidean "superspace" E^n. Let E be a d-dimensional linear subspace of E^n, and let $E^\perp$ be its orthogonal complement in E^n. Then, $\mathrm{E}^n = E \oplus E^\perp$. Let π and $\pi^\perp$ denote the orthogonal projections of E^n onto E or $E^\perp$, respectively. Recall that Z^n is the vertex set of the standard tessellation $\mathcal{C}$ of E^n by n-dimensional cubes, whose lower dimensional faces are also cubical. Let C denote the n-dimensional cube whose vertex set consists of all 0–1 vectors. A vector $t \in \mathrm{E}^n$ is said to be *regular* if the boundary of the *tube*,

$$E + C + t := \{x + c + t \mid x \in E,\ c \in C\}$$

does not have a point in common with Z^n; otherwise, z is *singular*. Note that the shape of the tube is determined by its cross-section with $E^\perp$, which is the projection image under $\pi^\perp$ of the cube $C + t$. If $d = n - 1$, the tube is a strip in E^n bounded by a pair of hyperplanes parallel to E.

Now, for every regular vector t, the d-dimensional faces of $\mathcal{C}$ that are fully contained in $E + C + t$ make up a d-dimensional polyhedral surface in E^n. This surface is tessellated by d-dimensional cubes and contains all i-dimensional faces of $\mathcal{C}$ contained in $E + C + t$, for $i = 0, \ldots, d$. When restricted to the surface, the projection π onto E is one-to-one and maps the cubical polyhedral complex on the surface onto a tiling $\mathcal{T}_t$ of E. Figure 16 illustrates the case $n = 2$ and $d = 1$, which yields tilings of the real line by projecting the staircase onto the line bounding the strip from below. The tiles of $\mathcal{T}_t$ are the projection images under π of the d-dimensional cubes on the surface; that is, the tiles are d-dimensional parallelopipeds. If $E \cap \mathrm{Z}^n = \{0\}$ (that is, if E is "totally irrational"), then the resulting tilings $\mathcal{T}_t$ with t regular are nonperiodic. The structure of these tilings depends on the parameter vector t (and the subspace E, of course), but any two such tilings (with the same E) are locally isomorphic. Note that the tilings are nondecorated.

The same tilings can also be obtained by the *canonical cut method* as follows. Let $e_1, \ldots, e_n$ denote the canonical basis of E^n, and let π and $\pi^\perp$ be as above. In applications, any d of the vectors $\pi(e_i)$ in E (or equivalently, any $n - d$ of the vectors $\pi^\perp(e_i)$ in $E^\perp$), with $i = 1, \ldots, n$, are linearly independent, and so we assume this from now on.

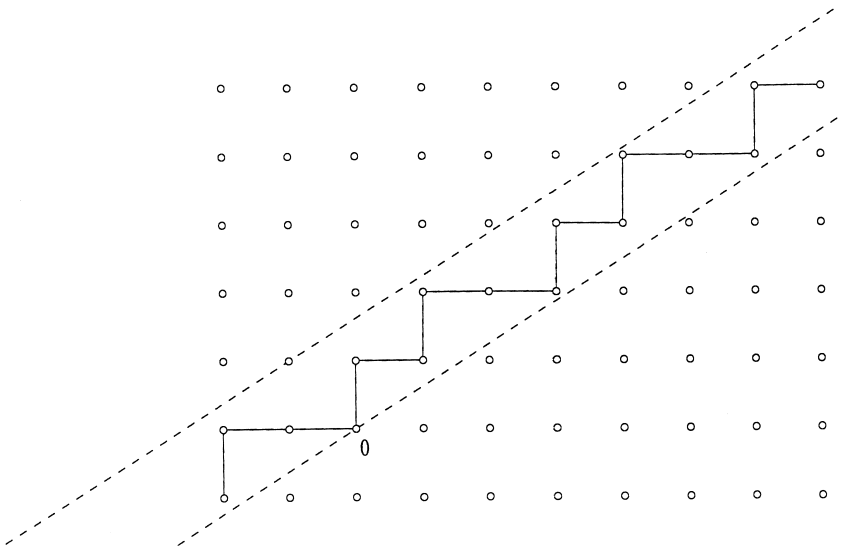


FIGURE 16 Illustration of the strip projection method for $n=2$ and $d=1$. The projection of the staircase onto the (totally irrational) line bounding the strip from below yields a nonperiodic tiling on the line by intervals of two sizes. These intervals are the projections of the horizontal or vertical line segments that are contained in the strip and connect two adjacent grid points. [From Senechal, M. (1995). *Quasicrystals and Geometry*, Cambridge University Press. Cambridge.]

We first construct a periodic tiling of E^n by prismatic tiles. For a d-element subset I of $\{1, \ldots, n\}$, let B_I and $B_I^\perp$, respectively, denote the parallelopipeds in E or $E^\perp$ that are generated by the sets of vectors $\{\pi(e_i) \mid i \in I\}$ and $\{-\pi^\perp(e_i) \mid i \notin I\}$ (the minus sign is significant). The convex n-polytope $P_I := B_I \oplus B_I^\perp$ is a "prism" with bases B_I in E and $B_I^\perp$ in $E^\perp$. Then the prisms $P_I + z$, with $z \in \mathrm{Z}^n$ and I a d-element subset of $\{1, \ldots, n\}$, are the tiles in a periodic tiling $\mathcal{O}$ of E^n called the *oblique tiling*.

The nonperiodic tilings now are obtained as sections of $\mathcal{O}$ with certain subspaces $E + t$. More precisely, if t is regular as before, then the intersection of the affine d-dimensional subspace $E + t$ with a tile $P_I + z$ of $\mathcal{O}$ either is empty, or of the form $B_I + y + z$ with y a relative interior point of $B_I^\perp$. Thus, the intersections of $E + t$ with the tiles of $\mathcal{O}$ yield the tiles in a tiling of $E + t$ by d-dimensional parallelopipeds. Then, via projection by π onto E, this gives a tiling of E by parallelopipeds. This is the same tiling $\mathcal{T}_t$ as before. Figure 17 illustrates the case $n = 2$ and $d = 1$; now $\mathcal{O}$ is a plane tiling with two kinds of square tiles (small and large), and nonperiodic tilings on the line are obtained by cutting $\mathcal{O}$ by a (regular) translate of E.

Further properties of these tilings generally depend on the particular choice of subspace E. For example, to obtain the *generalized Penrose tilings* of the plane, take $n = 5$ and consider the group G of isometries of E^5 that permutes the coordinates like a cyclic group of order 5. Then G has three invariant subspaces, namely the line spanned by $e_1 + \cdots + e_5$, and two planes, including the plane E spanned by the vectors:

$$x_1 = (4, \sqrt{5} - 1, -\sqrt{5} - 1, -\sqrt{5} - 1, \sqrt{5} - 1)$$
$$x_2 = (\sqrt{5} - 1, 4, \sqrt{5} - 1, -\sqrt{5} - 1, -\sqrt{5} - 1)$$

The resulting nondecorated plane tilings $\mathcal{T}_t$ in E have only two prototiles, namely the nondecorated Penrose rhombs. If the (regular) parameter vector t is contained in the invariant plane E' of G distinct from E (E' is a subspace of $E^\perp$), then we obtain an ordinary nondecorated Penrose tiling by rhombs (that is, a tiling obtained from a decorated Penrose tiling by rhombs by removing the decorations). The tilings $\mathcal{T}_t$ with $t \in E'$ have the property that every finite patch in any ordinary nondecorated Penrose tiling by rhombs already occurs in one of them. It is known that

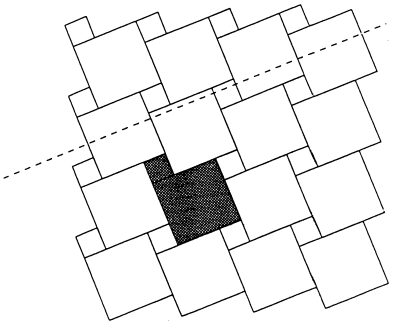


FIGURE 17 The oblique tiling $\mathcal{O}$ for $n=2$ and $d=1$. There are two kinds of square tiles in $\mathcal{O}$. A nonperiodic tiling on the line by two kinds of intervals is obtained by cutting $\mathcal{O}$ with the dotted line. [From Senechal, M. (1995). *Quasicrystals and Geometry*, Cambridge University Press. Cambridge.]

every ordinary nondecorated Penrose tiling by rhombs can be decorated by arrows in exactly one way to become an ordinary decorated Penrose tiling by rhombs.

SEE ALSO THE FOLLOWING ARTICLES

INCOMMENSURATE CRYSTALS AND QUASICRYSTALS • MATHEMATICAL MODELS

BIBLIOGRAPHY

Bezdek, K. (2000). "Space filling," *In* "Handbook of Discrete and Combinatorial Mathematics" (K. H. Rosen, ed.), pp. 824–830, CRC Press, Boca Raton, FL.

Conway, J. H., and Sloane, N. J. A. (1988). "Sphere Packings, Lattices and Groups," Springer-Verlag, New York.

Grünbaum, B., and Shephard, G. C. (1986). "Tilings and Patterns," Freeman, San Francisco, CA.

Gruber, P. M., and Lekkerkerker, C. G. (1987). "Geometry of Numbers," 2nd ed., North-Holland, Amsterdam.

Le, T. T. Q. (1995). "Local rules for quasiperiodic tilings," *In* "The Mathematics of Long-Range Aperiodic Order" (R. V. Moody, ed.), pp. 331–366, Kluwer Academic, Dordrecht/Norwell, MA.

Moody, R. V., ed. (1995). "The Mathematics of Long-Range Aperiodic Order," NATO ASI Series, Series C: Mathematical and Physical Sciences, Vol. 489, Kluwer Academic, Dordrecht/Norwell, MA.

Patera, J., ed. (1998). "Quasicrystals and Discrete Geometry," Fields Institute Monographs, Vol. 10, American Mathematical Society, Providence, RI.

Radin, C. (1999). "Miles of Tiles," Student Mathematical Library, Vol. 1, American Mathematical Society, Providence, RI.

Schattschneider, D., and Senechal, M. (1997). "Tilings," *In* "Handbook of Discrete and Computational Geometry" (J. E. Goodman and J. O'Rourke, eds.), pp. 43–62, CRC Press, Boca Raton, FL.

Schulte, E. (1993). "Tilings," *In* "Handbook of Convex Geometry" (P. M. Gruber and J. M. Wills, eds.), pp. 899–932, Elsevier, Amsterdam.

Senechal, M. (1995). "Quasicrystals and Geometry," Cambridge University Press, Cambridge, U.K.

Stein, S., and Szabo, S. (1994). "Algebra and Tiling," The Carus Mathematical Monographs, Vol. 25, Mathematical Association of Amererica, Providence, RI.

Tarasov, A. S. (1997). "Complexity of convex stereohedra," *Mathematical Notes* **61**(5), pp. 668–671.



Topology, General

Taqdir Husain
McMaster University

I. Topology and Topological Spaces
II. Maps and Functions
III. Methods of Constructing New Topological Spaces
IV. Separation Axioms
V. Special Topological Spaces
VI. Fixed Points
VII. Function Spaces
VIII. Marriage of Topology and Algebra
IX. Banach Algebras
X. Algebraic Topology

GLOSSARY

Closure $\bar{A} = A \cup A'$ is called the closure of A, where A' is the set of all limit points of A.

Compact A topological space X is called compact if in each open covering of X there is a finite subcollection that covers X.

Connected A topological space is said to be connected if it cannot be written as the union of two disjoint nonempty open sets.

Continuous map A map f of a topological space X into a topological space Y is called continuous if for each open set Q of Y, $f^{-1}(Q) = \{x \in X : f(x) \in Q\}$ is open in X.

Covering A family $\{P_\alpha\}$ of sets is said to be a covering of $A \subset X$ if $A \subset \bigcup_\alpha A_\alpha$. If all P_α's are open, it is called an open covering.

Homeomorphism A continuous, open, and bijective map of a topological space X into a topological space Y is called a homeomorphism.

Limit point An element a is said to be a limit or accumulation point of a set A if each open set containing a contains at least one point of A other than a.

Metric d is said to define a metric for a set X if for all pairs (x, y), $x, y \in X$, $d(x, y)$ is a real number satisfying the following:

(i) $d(x, y) \geq 0$.
(ii) $d(x, y) = 0$, if $x = y$.
(iii) $d(x, y) = 0$ implies $x = y$.
(iv) $d(x, y) = d(y, x)$.
(v) $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality).

Neighborhood A set U is a neighborhood of a point

a if there exists an open set P such that $a \in P \subset U$.

Open set Each member of a topology $\mathcal{T}$ on a set X is called an open set.

Set A collection of "distinguished" objects is called a set.

Topological group A group G endowed with a topology is called a topological group if the map $(x, y) \to xy^{-1}$ of $G \times G$ into G is continuous.

GENERAL TOPOLOGY is a branch of mathematics that deals with the study of topological spaces and maps, especially continuous maps. This subject is regarded as a natural extension of part of classical analysis (a major field of mathematics) insofar as the study of continuous maps is concerned. Classical analysis deals primarily with functions on the real line $\mathbb{R}$ or, more generally, n-dimensional Euclidean spaces $\mathbb{R}^n$, in which the notions of limit and continuity are very basic. With the help of topology, these concepts can be studied more generally in topological spaces. In short, the genesis of the subject general topology is rooted in the fundamental properties of $\mathbb{R}$ or $\mathbb{R}^n$ and functions thereon.

The reader is cautioned not to confuse "topology" with "topography" and "map" with a "map" of a country. For an explanation of basic set-theoretic concepts used herein, see Table I.

TABLE I Symbols and Set-Theoretic Notations

Set	A collection of distinguished objects: capital letters denote sets in the text
$\varnothing$	Null or empty set
$a \in A$	a is an element of the set A
$A \cup B$	Union of A and B
$A \cap B$	Intersection of A and B
$A \cap B = \varnothing$	A, B are disjoint
A^c	Complement of A
$A \backslash B$	The set of elements in A that are not in B
$\bigcup_{i=1}^{n} A_i$	(Finite) union of sets $A_1, A_2, \ldots, A_n$
$\bigcap_{i=1}^{n} A_i$	(Finite) intersection of sets $A_1, \ldots, A_n$
$\bigcup_{i=1}^{\infty} A_i$	Countable union of sets $\{A_1, A_2, \ldots\}$
$\bigcap_{i=1}^{\infty} A_i$	Countable intersection
$\bigcup_{\alpha\in\Gamma} A_\alpha$ or $\bigcup_\alpha A_\alpha$	(Arbitrary) union of a family $\{A_\alpha\}_{\alpha\in\Gamma}$ of sets
$\bigcap_{\alpha\in\Gamma} A_\alpha$ or $\bigcap_\alpha A_\alpha$	Arbitrary intersection
$f : X \to Y$	f is a map of X into Y
iff (or $\Leftrightarrow$)	If and only if (or implies and implied by)
$\mathbb{R}$ (or $\mathbb{C}$)	The set of all real (or complex) numbers
$\sum_{i=1}^{n} a_i$	$a_1 + a_2 + \cdots + a_n$
De Morgan's laws	$(\bigcup_\alpha A_\alpha)^c = \bigcap_\alpha A_\alpha^c$ and $(\bigcap_\alpha A_\alpha)^c = \bigcup_\alpha A_\alpha^c$

I. TOPOLOGY AND TOPOLOGICAL SPACES

A. Basic Notions of Topology and Examples

Let X be a set. A collection $\mathcal{T} = \{U_\alpha\}$ of subsets of X is said to define or is itself called a *topology* if the following axioms hold:

(i) $X, \varnothing$ belong to $\mathcal{T}$.
(ii) Finite intersections of elements in $\mathcal{T}$ are also in $\mathcal{T}$.
(iii) Arbitrary unions of elements in $\mathcal{T}$ are also in $\mathcal{T}$.

A set X endowed with a topology $\mathcal{T} = \{U_\alpha\}$ is called a *topological space* (TS) and is often denoted $(X, \mathcal{T})$. Each member of $\mathcal{T}$ is called an *open* set or $\mathcal{T}$*-open* (if the topology is to be emphasized). Thus, by definition the whole set X and the null set $\varnothing$ are open.

If the topology $\mathcal{T}$ consists of exactly two open sets $X, \varnothing$, then the topology is termed *indiscrete*. If, however, $\mathcal{T}$ consists of all subsets of X, it is called the *discrete* topology.

Let $\mathcal{T}, \mathcal{T}'$ be topologies on a set X. If each $\mathcal{T}'$-open set is also $\mathcal{T}$-open, then $\mathcal{T}$ is said to be *finer* than $\mathcal{T}'$ or, equivalently, $\mathcal{T}'$ is *coarser* than $\mathcal{T}$. Thus, the discrete topology is finer than the indiscrete one. Actually, the discrete topology is the *finest*, whereas the indiscrete one is the weakest or *coarsest* of all topologies on X.

A set C of a topological space $(X, \mathcal{T})$ is called *closed* or $\mathcal{T}$*-closed* (when the topology is to be emphasized) if the complement $C^c = X \backslash C$ of C is open. By De Morgan's laws (see Table I), arbitrary intersections and finite unions of closed sets are closed, since by definition finite intersections and arbitrary unions of open sets are open.

A subset P of a topological space $(X, \mathcal{T})$ is called a *neighborhood of a point* $x \in X$ if there is an open set U such that $x \in U \subset P$. If P itself is an open or a closed set, it is called an *open* or *closed neighborhood* of x. The collection $\mathcal{N}_x$ of all neighborhoods of x is called the *neighborhood system* at x. The following properties hold: $x \in U$ for all U in $\mathcal{N}_x$; for all U, V in $\mathcal{N}_x$, $U \cap V \in \mathcal{N}_x$; if W is any subset of X such that for some $U \in \mathcal{N}_x$, $U \subset W$, then $W \in \mathcal{N}_x$; and each U in $\mathcal{N}_x$ contains an open neighborhood of x.

Let A be a subset of a topological space $(X, \mathcal{T})$. The union of all open sets contained in A is called the *interior* of A and is denoted A^0. Clearly, $A^0 \subset A \cdot A = A^0$ iff (if and only if) A is open.

A point $x \in X$ is called a *limit* or *accumulation point* of a set $A \subset X$ if each neighborhood of x contains points of A other than x. The set of all limit points of a set A is usually denoted A' and is called the *derived set* of A. The set $\bar{A} = A \cup A'$ is called the *closure* of A. Closed sets can be characterized by closures as follows. A set A is closed iff

$A = \bar{A}$. In particular, each closed set contains all its limit points. In general, however, $A \subset \bar{A}$.

Properties of interior and closure operations are complementary and comparable: $A^0 \subset A$ (respectively, $A \subset \bar{A}$); $A^{00} = A^0$ (respectively, $\bar{\bar{A}} = \bar{A}$); and $A^0 \cap B^0 = (A \cap B)^0$ (respectively, $\bar{A} \cup \bar{B} = \overline{A \cup B}$).

The set $\partial A = \bar{A} \cap \overline{A^c}$ is called the *boundary* of A. It immediately follows that A is open (respectively, closed) iff $A \cap \partial A = \varnothing$ (respectively, $\partial A \subset A$).

A subset A of a topological space $(X, \mathcal{T})$ is said to be *dense* in X if $\bar{A} = X$. Since X is always closed, $\bar{X} = X$ implies that the whole space X is always a dense subset of itself. However, if the subset A is genuinely smaller than X, the topological space assumes some special features. A noteworthy case is the following. If A is a countable dense subset of X, then $(X, \mathcal{T})$ is called a *separable* topological space.

In practice, the collection of all open sets in a topological space is very huge. To cut it down to a smaller size without changing the topology, the notion of a base of a topology is useful. A subcollection $\mathcal{B}$ of open sets in a topological space $(X, \mathcal{T})$ is said to form a *base* of the topology if each open set of X is the union of sets from $\mathcal{B}$. To cut the size of $\mathcal{B}$ down even farther without changing the topology, one introduces the following concept. A subcollection $\mathcal{B}'$ of open sets in a topological space $(X, \mathcal{T})$ is said to form a *subbase* of the topology if the collection of all finite intersections of elements form $\mathcal{B}'$ forms a base of the topology. One notes the following facts. A collection $\mathcal{B}$ of open sets forms a base of the topology $\mathcal{T}$ iff for each $x \in X$ and each neighborhood U of x there is an element B in $\mathcal{B}$ such that $x \in B \subset U$. Furthermore, any nonempty family $\mathcal{B}'$ of subsets of a given set X such that $\mathcal{B}'$ covers X defines a unique topology on X or is a subbase of a unique topology.

If a topology $\mathcal{T}$ on X has a countable base, then $(X, \mathcal{T})$ is said to be a *second countable* (SC) space, but if each $x \in X$ has a *countable base* $\{U_n(x)\}$ of *neighborhoods*, that is, each neighborhood of x contains some $U_n(x)$, then the topological space is said to be a *first countable* (FC) space. It is clear that each second countable space is first countable, but the converse is not true. Moreover, each second countable space is separable. Again, the converse is not true. However, for a special class of topological spaces called metric space, which is introduced in Section V.A, the converse does hold. To ascertain the existence of limit points of certain subsets in a topological space, one has the following. Each uncountable subset of a second countable space has a limit point.

To make the above-mentioned abstract notions more concrete, we consider some examples. In the set $\mathbb{R}$ of all real numbers, for $a, b \in \mathbb{R}$, the subsets $(a, b) = \{x \in \mathbb{R} : a < x < b\}$ and $[a, b] = \{x \in \mathbb{R} : a \le x \le b\}$ are called open and closed intervals, respectively. A subset U of $\mathbb{R}$ is called open if for each $x \in U$ there is an open interval (a, b) such that $x \in (a, b) \subset U$. The collection of all open sets so designated defines a topology on $\mathbb{R}$, sometimes called the *natural* or *Euclidean topology* of $\mathbb{R}$. Clearly, each open (respectively, closed) interval of $\mathbb{R}$ is an open (respectively, closed) set. Thus, the collection of all open intervals forms a base of the natural topology of $\mathbb{R}$. Furthermore, it is easy to see that the collection of all half-open lines, $(a, \infty) = \{x \in \mathbb{R} : a < x\}$ and $(-\infty, b) = \{x \in \mathbb{R} : < b\}$, forms a subbase of the natural topology. If a, b run over rational numbers of $\mathbb{R}$, the collection $\{(a, b)\}$ of all these open intervals being a countable base of the natural topology of $\mathbb{R}$ makes $\mathbb{R}$ a second countable space. Since the set $\mathbb{Q}$ of rational numbers is a countable dense subset of $\mathbb{R}$, $\mathbb{R}$ is separable.

Apart from the natural topology of $\mathbb{R}$, as mentioned above, there are many other topologies, among which are the discrete and indiscrete. In the discrete case, each subset of $\mathbb{R}$ is open, in particular, each singleton (consisting of a single point) is open. Certainly, however, no singleton is open in the natural topology of $\mathbb{R}$. Thus, it follows that the discrete topology of $\mathbb{R}$ is strictly finer than the natural topology. $\mathbb{R}$ with the discrete topology is not separable and hence not second countable. Since for each x, $\{x\}$ forms a countable base of the neighborhood system at x, $\mathbb{R}$ with the discrete topology is first countable, but $\mathbb{R}$ with the indiscrete topology is trivial, because each singleton $\{x\}$ $(x \in \mathbb{R})$ is dense in $\mathbb{R}$.

Another important example of a useful topological space is the *n-dimensional* Euclidean real (or complex) space $\mathbb{R}^n$ (or $\mathbb{C}^n$) which is the set of all *n-tuples* $(x_1, x_2, \ldots, x_n)$ in which all $x_1, \ldots, x_n$ are real (respectively, complex) numbers. If $n = 1$, then $\mathbb{R}^1$ coincides with $\mathbb{R}$; for $n = 2$, $\mathbb{R}^2$ is called the *Euclidean plane*; for $n = 3$, $\mathbb{R}^3$ is called the *3-space*; and so on. The subsets defined by

$$B_\varepsilon(a_1, \ldots, a_n) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : (x_1 - a_1)^2 + \cdots + (x_n - a_n)^2 < \varepsilon^2\}$$

and

$$B'_\varepsilon(a_1, \ldots, a_n) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : (x_1 - a_1)^2 + \cdots + (x_n - a_n)^2 \le \varepsilon^2\}$$

are called *open* and *closed disks* of $\mathbb{R}^n$, respectively. A subset P of $\mathbb{R}^n$ is said to be open if for each $(x_1, \ldots, x_n) \in P$ there is a real number $\varepsilon > 0$ such that $(x_1, \ldots, x_n) \in B_\varepsilon(x_1, \ldots, x_n) \subset P$. Thus, the collection of all open disks forms the base of a topology called the *Euclidean topology* of $\mathbb{R}^n$. In this topology, each open (respectively, closed) disk is an open (respectively, closed) set. Like $\mathbb{R}$, $\mathbb{R}^n$ is second countable and hence

separable because the set of n-tuples $(x_1, \ldots, x_n)$ in which all $x_1, \ldots, x_n$ are rational numbers is a countable dense subset of $\mathbb{R}^n$. Here, too, the discrete topology is strictly finer than the Euclidean topology.

Another important property of $\mathbb{R}^n$ $(n \geq 1)$ that lends itself to abstraction is the following. A topological space $(X, \mathcal{T})$ is called *connected* if X cannot be written as the union of two disjoint, nonempty open sets of X. In other words, if $X = P \cup Q$ for some open sets P, Q such that $P \cap Q = \emptyset$, then one of P, Q must be a null set. $\mathbb{R}$ and $\mathbb{R}^n$ with their natural topology are connected, but not so in the discrete topology. $(X, \mathcal{T})$ is called *locally connected* if each neighborhood of each point of X contains a connected open neighborhood. Indeed, $\mathbb{R}^n$ $(n \geq 1)$ is locally connected.

A subset A of a topological space $(X, \mathcal{T})$ is called *connected* if it is so in the induced topology (see Section III.A). The closure of a connected set is connected. Moreover, an arbitrary union of connected subsets is connected provided that no two members of the collection are disjoint. A connected subset that is not properly contained in any other connected subset is called a *component*. Clearly, if a topological space $(X, \mathcal{T})$ is connected, it is its only component. Each component of the real numbers with the discrete topology consists of a singleton, whereas $\mathbb{R}$ with the natural topology is its only component.

B. Methods of Defining Topology

The method of defining topology by means of open sets is not the only one. There are several methods used in the mathematical literature. Four of them are given below.

1. Kuratowski Axioms

One can define topology by closure operations. This method was introduced by K. Kuratowski. Let X be a set and let $\mathcal{P}(X)$ denote the collection of all subsets of X (often called the *power set* of X). Let ϕ be an operation that to each subset A of X assigns a subset of X. In other words ϕ maps $\mathcal{P}(X)$ into itself. Suppose that ϕ satisfies the so-called Kuratowski axioms: $\phi(\emptyset) = \emptyset$; $A \subset \phi(A)$; $\phi(\phi(A)) = \phi(A)$; $\phi(A \cup B) = \phi(A) \cup \phi(B)$ for all A, B in $\mathcal{P}(X)$. Then there exists a unique topology $\mathcal{T}$ on X such that $\phi(A)$ becomes the closure $\bar{A}$ of A with respect to the topology $\mathcal{T}$. Specifically, a set A is called closed with respect to $\mathcal{T}$ if $\phi(A) = A$. Hence, the complement A^c of a closed set A will, as usual, be open, and the topology $\mathcal{T}$ is thus determined.

The reader will have noticed that Kuratowski axioms are satisfied by the closure operation defined in a topological space as shown in the preceding section.

2. Neighborhood Systems

The second method of defining topology is via neighborhood systems. Let X be a set. Suppose that for each $x \in X$ there is a collection $\mathcal{U}_x$ of subsets of X satisfying the following axioms. For each $x \in X$ and each U in $\mathcal{U}_x$, $x \in U$; if a subset W of X contains some $U \in \mathcal{U}_x$, then $W \in \mathcal{U}_x$; finite intersections of elements from $\mathcal{U}_x$ are also in $\mathcal{U}_x$; and for each U in $\mathcal{U}_x$ there is a V in $\mathcal{U}_x$ such that $V \subset U$ and $U \in \mathcal{U}_y$ for all $y \in V$. Then there exists a unique topology $\mathcal{T}$ on X such that for each $x \in X$, $\mathcal{U}_x$ becomes a neighborhood system at x with respect to the topology $\mathcal{T}$.

3. Nets

Another method of defining topology is by means of nets. The notion of net generalizes the concept of sequence, which is so useful and popular in the case of real numbers. To define *net* we first need some other notions.

Let Γ denote a nonempty set. Γ is called a partially ordered set if there is a binary relation ">" (or "<") defined for elements of Γ such that whenever $\alpha > \beta$ and $\beta > \gamma$ for $\alpha, \beta, \gamma \in \Gamma$, we have $\alpha > \gamma$. (Clearly, the set of all positive integers is a partially ordered set.) A partially ordered set Γ with the binary relation $>$ is called *directed* if for $\alpha, \beta \in \Gamma$ there is a $\gamma \in \Gamma$ such that $\gamma > \alpha$, $\gamma > \beta$. Γ is called linearly ordered if for all $\alpha, \beta \in \Gamma$ either $\alpha > \beta$ or $\beta > \alpha$ and $\alpha > \beta$, $\beta > \alpha \Rightarrow \alpha = \beta$. An element $\gamma \in \Gamma$ is called an *upper bound* of a subset $\Gamma' \subset \Gamma$ if $\gamma \geq \alpha$ for all $\alpha \in \Gamma'$. γ is called the *least upper bound* if γ is the smallest upper bound of Γ'.

A map ϕ of a directed partially ordered set Γ into a set X is called a *net*. If we write $\phi(\alpha) = x_\alpha \in X$ for all $\alpha \in \Gamma$, then often a net is denoted by its range $\{x_\alpha : \alpha \in \Gamma\}$ in X. In short, one denotes a net $\{x_\alpha\}$ instead of $\{x_\alpha : \alpha \in \Gamma\}$. Clearly, if we replace Γ by $\mathbb{N}$ (positive integers), then the net $\{x_n : n \geq 1\} = \{x_n\}$ is a sequence in X. Thus, a sequence is a particular case of a net.

A net $\{x_\alpha\}$ is called a *constant net* if $x_\alpha = x$ for all $\alpha \in \Gamma$. If $\{x_n\}$ is a sequence and $n_1 < n_2 < \cdots$ are infinitely many positive integers, then $\{x_{n_k}\} = \{x_{n_1}, x_{n_2}, \ldots\}$ is called a *subsequence* of $\{x_n\}$.

A net $\{x_\alpha\}$ in a topological space $(X, \mathcal{T})$ is said to *converge* to a point $x \in X$ if for each neighborhood U of x there exists $\alpha_0 \in \Gamma$ such that for $\alpha \geq \alpha_0$ $(\alpha \in \Gamma)$, $x_\alpha \in U$. Now the closure of a set A in X can be characterized by means of nets as follows: $x \in \bar{A}$ iff there is a net $\{x_\alpha\}$ in A converging to x. Clearly, a constant net is always convergent. Moreover, a subsequence of a convergent sequence is also convergent. Without going into finer details, let us say that nets can be used to define limit points and then closures of subsets leading to the definition of topology. For details the

interested reader is referred to any standard book on topology listed in the Bibliography.

4. Filters

Another method of defining a topology on a set is via filters. The notion of filters was introduced and popularized by French mathematicians, especially those belonging to the French group Bourbaki, named after a fictitious mathematician, Nicholas Bourbaki.

Let X be a nonempty set. A collection $\mathcal{F} = \{F_\alpha\}$ of subsets of X is called a *filter* if the following axioms hold. Each F_α is nonempty; for F_α, F_β in $\mathcal{F}$, $F_\alpha \cap F_\beta \in \mathcal{F}$; and if H is any subset of X such that $F_\alpha \subset H$ for some $F_\alpha \in \mathcal{F}$, then $H \in \mathcal{F}$. A collection $\mathcal{F} = \{F_\alpha\}$ of subsets F_α in X is called a *filter base* if each F_α is nonempty and if, for F_α, F_β in $\mathcal{F}$, there is F_γ in $\mathcal{F}$ such that $F_\gamma \subset F_\alpha \cap F_\beta$. Clearly, a filter base generates a filter consisting of all subsets of X, each containing a member of the filter base.

A neighborhood system $\mathcal{N}_x$ of x in a topological space $(X, \mathcal{T})$ is a filter called the *neighborhood filter*.

A filter $\mathcal{F} = \{F_\alpha\}$ in a topological space $(X, \mathcal{T})$ is said to *converge to* $x \in X$ if for each neighborhood U of x there exists $F_\alpha \in \mathcal{F}$ such that $F_\alpha \subset U$. An element $y \in X$ is called a *limit point* of a filter $\mathcal{F} = \{F_\alpha\}$ if $y \in \bar{F}_\alpha$ for all F_α in $\mathcal{F}$.

Theories of filters and nets are equivalent. Specifically, if $\{x_\alpha : \alpha \in \Gamma\}$ is a net, the collection of all $F_\alpha = \{x_\beta : \beta > \alpha, \beta \in \Gamma\}$ is a filter base. Conversely, given a filter $\mathcal{F} = \{F_\alpha\}$, it is possible to construct a net $\{x_\alpha\}$ by considering the set Γ of all pairs (x_α, F_α) in which $x_\alpha \in F_\alpha$, and by defining $(x_\alpha, F_\alpha) \leq (x_\beta, F_\beta) \Leftrightarrow F_\beta \subset F_\alpha$, we see that $\leq$ defines a partial ordering and Γ becomes a directed set because $\mathcal{F}$ is a filter. Now if we put $\phi(x_\alpha, F_\alpha) = x_\alpha$, we obtain a net $\{x_\alpha\}$. Details can be found in standard texts on topology. Thus, filters can be used to define a topology as nets do.

A filter can be contained in a *maximal* filter (one that is not a proper subcollection of a filter). A maximal filter is called an *ultrafilter*. (The existence of maximal filters is guaranteed by Zorn's lemma: Each partially ordered set in which each linearly ordered subset has a least upper bound has a maximal element.) A filter $\mathcal{F}$ in X is an ultrafilter iff for each subset A of X, one of the two sets A, A^c is in $\mathcal{F}$. Ultrafilters are useful for the study of extensions of topological spaces as well as compactness among others. We shall see one application of ultrafilters in Section V.B.

II. MAPS AND FUNCTIONS

Given two sets X and Y, if to each element x of X there corresponds a unique element y of Y, then this correspondence is called a *map* and often written as $f : X \to Y$, meaning that $y = f(x)$ for $x \in X$. X is called the *domain* of the map f, and $f(X) = \{y \in Y : y = f(x), x \in X\}$ the *range* of f. If $f(x) = f(x')$ implies that $x = x'$ for all $x, x' \in X$, then f is called *one to one* (or *injective*). If for each $y \in Y$ there is $x \in X$ such that $y = f(x)$, then f is called *onto* (or *surjective*). A map that is both injective and surjective is called *bijective*.

In particular cases when either $Y = \mathbb{R}$ (real numbers) or $X = Y = \mathbb{R}$, a map $f : X \to Y$ is often called a *real-valued function* or simply *function*.

The notion of map or function is very basic to understanding the physical phenomena of the universe, since each physical or other kind of event for a mathematical study must be expressed as a function. Indeed, the properties of real-valued functions studied in classical analysis are used in almost all scientific studies. Here we look at some special properties of maps between topological spaces.

A. Continuous, Open Maps and Homeomorphisms

Let $(X, \mathcal{T})$ and $(Y, \mathcal{T}')$ be two topological spaces. A map $f : X \to Y$ is said to be *continuous at a point* $x_0 \in X$ if for each neighborhood V of $f(x_0)$ in Y there exists a neighborhood U of x_0 in X such that $f(x) \in V$ for all $x \in U$. If f is continuous at each point of X, it is called *continuous*.

Indeed, each continuous map is continuous at each point of X by definition, but there are functions that are continuous at one point but not at another. For example, $f(x) = 1$ if $x \geq 0$ and $= 0$ if $x < 0$ is a real-valued function on $\mathbb{R}$ that is continuous at all $x \neq 0$ but not continuous at $x = 0$.

Since the continuity of a map depends on the topologies of the topological spaces concerned and since there are several methods of defining topologies, it is natural that the continuity of a map can be expressed in more than one way. Specifically, we have the following equivalent statements:

(i) f is continuous (see the above definition).
(ii) For each open set Q of Y, $f^{-1}(Q) = \{x \in X : f(x) \in Q\}$ is open in X.
(iii) For each net $\{x_\alpha\}$ in X converging to $x \in X$, $\{f(x_\alpha)\}$ converges to $f(x)$ in Y.
(iv) For each filter $\mathcal{F} = \{F_\alpha\}$ converging to $x \in X$, $\{f(F_\alpha)\}$ converges to $f(x) \in Y$.
(v) For each subset A of X, $f(\bar{A}) \subset \overline{f(A)}$.
(vi) For each subset B of Y, $f^{-1}(\overline{B}) \subset f^{-1}(\bar{B})$.

The continuity property is transitive in the following sense. If $f : X \to Y$ and $g : Y \to Z$ are continuous

maps, respectively, of topological spaces from $(X, \mathcal{T})$ into $(Y, \mathcal{T}')$ and from $(Y, \mathcal{T}')$ into $(Z, \mathcal{T}'')$, then the *composition map* $g \circ f : X \to Z$ defined by $g \circ f(x) = g(f(x))$, $x \in X$, is also continuous.

Although the continuity property of maps between topological spaces is the most popular and useful, there are other notions that complement or augment continuity.

A map f of a topological space $(X, \mathcal{T})$ into another topological space $(Y, \mathcal{T}')$ is called *open* if for each open set P of X, $f(P) = \{y = f(x) : x \in P\}$ is open in Y. It follows that f is open iff $[f(A)]^0 \subset f(A^0)$ for all subsets A of X. (Note that A^0 denotes the interior of A defined earlier.) A continuous map need not be open (e.g., $i : (\mathbb{R}, \mathcal{D}) \to (\mathbb{R}, \mathcal{T})$ in which $\mathcal{D}$ is the discrete topology and $\mathcal{T}$ the natural topology), or vice versa.

Similarly a map $f : (X, \mathcal{T}) \to (Y, \mathcal{T}')$ is called *closed* if for each closed set C of X, $f(C)$ is a closed set in Y. Here, too, we have a characterization: f is closed iff for each subset A of X, $\overline{f(A)} \subset f(\bar{A})$. Again, a continuous map need not be closed, nor is a closed map necessarily continuous.

A continuous, open, and bijective map of a topological space $(X, \mathcal{T})$ into another topological space $(Y, \mathcal{T}')$ is called a *homeomorphism*. If $f : (X, \mathcal{T}) \to (Y, \mathcal{T}')$ is a homeomorphism, then $(X, \mathcal{T})$ is said to be *homeomorphic* to $(Y, \mathcal{T}')$. In this case, there is no topological difference between the spaces $(X, \mathcal{T})$ and $(Y, \mathcal{T}')$, and hence they are sometimes called *topologically equivalent*.

It is one of the most important aspects of topological studies to discover conditions for homeomorphisms. For example, a simple result is that a bijective map between two discrete spaces is a homeomorphism. We present further examples of this kind in Section V.B.

Continuity can be used to define topology, thus adding one more method to those described earlier. Let X be a set, $(Y, \mathcal{T}')$ a topological space, and $f : X \to Y$ a map. We can endow X with a topology with respect to which f becomes continuous. Consider the collection $\mathcal{T} = \{f^{-1}(P) : P \in \mathcal{T}'\}$ of subsets of X in which $f^{-1}(P) = \{x \in X : f(x) \in P\}$ for an open set P in Y. It is routine to verify that this collection defines a topology $\mathcal{T}$ on X and f becomes continuous.

If $\{f_\alpha\}_{\alpha \in \Gamma}$ is a family of maps from a set X into a topological space $(X, \mathcal{T}')$, then the collection $\mathcal{T} = \{f_\alpha^{-1}(P) : P \in \mathcal{T}', \alpha \in \Gamma\}$ defines a topology $\mathcal{T}$ on X with respect to which each f_α becomes continuous.

B. Generalizations of Continuous Maps

If one cannot have the most desirable basic property of continuity of a map between topological spaces, one studies weaker notions to acquire more information about the maps. For this reason, a number of concepts weaker than continuity have been introduced and studied in the mathematical literature. Some of these are given here.

1. Almost Continuity

A map $f : (X, \mathcal{T}) \to (Y, \mathcal{T}')$ is said to be almost continuous at $x_0 \in X$ if for each neighborhood V of $f(x_0)$, in Y, $\overline{f^{-1}(V)}$ is a neighborhood of x_0 in X, f is called almost continuous if it is so at each point of X. (Elsewhere in the mathematical literature, what I have called an almost-continuous map is also called nearly continuous.)

Comparing the notion of almost continuity with that of continuity, we discover easily that each continuous map is almost continuous. The converse is not true, however, [e.g., $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 1$ or 0 according to whether x is a rational or irrational number.]

2. Closed Graphs

Let $(X, \mathcal{T}), (Y, \mathcal{T}')$ be two topological spaces. We put $X \times Y = \{(x, y) : x \in X, y \in Y\}$, which is the collection of all pairs (x, y) with first coordinate x and second coordinate y. $X \times Y$ is called the *Cartesian product* of X and Y. We endow $X \times Y$ with a topology $\mathcal{P}$, which has for a base the collection $\{U \times V\}$, in which U comes from $\mathcal{T}$ and V from $\mathcal{T}'$. The topology so defined is called the *product topology* on $X \times Y$ (see Section III.B).

If $f : X \to Y$ is a map, the subset $\{(x, y) : y = f(x), x \in X\}$ of $X \times Y$ is called the *graph of* f and is sometimes denoted G_f. If G_f is a closed subset of $X \times Y$ when the Cartesian product is endowed with the product topology, then f is said to have a *closed graph*.

Each continuous function $f : \mathbb{R} \to \mathbb{R}$ has a closed graph. However, the converse is not true [e.g., $f(x) = 1/x$ $(x \neq 0)$ and $= 0$ $(x = 0)$ has a closed graph but it is not continuous]. On the other hand, the identity map of any indiscrete space onto itself is continuous but does not have a closed graph. Thus, in general, the closed-graph property is not an exact generalization of continuity. It is true, however, if the continuous map is from a topological space into a Hausdorff space, which is introduced in Section IV.

Any result in which the closed-graph property of a map implies its continuity is called a closed-graph theorem.

3. Upper (or Lower) Semicontinuous Functions

A real-valued function f on a topological space $(X, \mathcal{T})$ is called upper (or lower) semicontinuous if for each real number α, the set $\{x \in X : f(x) < \alpha\}$ [respectively, $\{x \in X : f(x) > \alpha\}$] is an open subset of X. If we put $(-f)(x) = -f(x)$ for all $x \in X$, then f is lower semicontinuous iff $-f$ is upper semicontinuous. Indeed, each

continuous real-valued function on a topological space is upper and lower semicontinuous.

4. Right- and Left-Continuous Functions

Let X be an interval of $\mathbb{R}$ and $f: X \to \mathbb{R}$ a function. f is said to be right (or left) continuous at $x_0 \in X$ if for each arbitrary real number $\varepsilon > 0$, there is a real number $\delta > 0$ such that $|f(x) - f(x_0)| < \varepsilon$ whenever $x_0 < x < x_0 + \delta$ (respectively, $x_0 - \delta < x < x_0$). Clearly, f is continuous at x_0 iff it is right and left continuous at x_0. However, a right-continuous or left-continuous function may fail to be continuous [e.g., $f(x) = [x]$, the largest integer less than or equal to $x \in \mathbb{R}$; $[x]$ is called the *step function*, which is right continuous but not continuous at integers].

III. METHODS OF CONSTRUCTING NEW TOPOLOGICAL SPACES

There are a number of methods of constructing new topological spaces from old. In this section we present some of them.

A. Subspaces and Induced Topology

Let $(X, \mathcal{T})$ be a topological space. Any subset Y of X can be made into a topological space as follows. Consider the collection $\mathcal{T}' = \{P \cap Y : P \in \mathcal{T}\}$ of subsets of Y. Then it is easy to see that this collection defines a topology of Y called the *induced* or *relative topology*. In this topology, a subset Q of Y is *relatively open* iff there is an open set P in X such that $Q = P \cap Y$. If a subset E of Y is open in X, clearly E is relatively open. However, a relatively open subset of Y need not be open in X unless Y is itself an open subset of X. A subset Y of X endowed with the induced topology $\mathcal{T}'$ is called a *subspace* of X.

A topological space $(X, \mathcal{T})$ is said to have the *hereditary property* if each subspace of X satisfies the same property as does $(X, \mathcal{T})$ (see Table III for more information).

B. Products and Product Topology

There are two types of products: finite products and infinite products. Infinite products are of two kinds: countable products and arbitrary infinite products. By forming products of topological spaces we produce new topological spaces.

1. Finite Products

Let $(X_k, \mathcal{T}_k)$, $k = 1, 2, \ldots, n$, be n topological spaces with their respective topologies $\mathcal{T}_k$. By $\prod_{k=1}^{n} X_k = X_1 \times X_2 \times \cdots \times X_n$, we mean the set of all ordered n-tuples $(x_1, x_2, \ldots, x_n)$ in which $x_k \in X_k$, $k = 1, 2, \ldots, n$. Here, $\prod_{k=1}^{n} X_k$ is called the *finite Cartesian product* of $X_1, X_2, \ldots, X_n$. We topologize $\prod_{k=1}^{n} X_k$ as follows. The family $\mathcal{B} = \{U_1 \times U_2 \times \cdots \times U_n : U_k \text{ is open in } (X_k, \mathcal{T}_k), k = 1, \ldots, n\}$ forms a base of a topology, called the *product topology* on $\prod_{k=1}^{n} X_k$. In other words, a subset P of $\prod_{k=1}^{n} X_k$ is open in the product topology iff for each point $(x_1, \ldots, x_n) \in P$ there are open sets U_k in X_k for $k = 1, 2, \ldots, n$ with $x_k \in U_k$ such that $U_1 \times U_2 \times \cdots \times U_n \subset P$. We have already considered the special case when $n = 2$.

If $X = \prod_{k=1}^{n} X_k$ is a finite product with the product topology, the map $p_k(x_1, \ldots, x_n) = x_k$ is called the *kth projection* of X into X_k. Each p_k is continuous, open, and surjective. A finite product $\prod_{k=1}^{n} X_k$ is said to have the *productive property* if X satisfies the same property as does each X_k (see Table III).

2. Infinite Products

Let A be a nonempty indexing set. For each $\alpha \in A$, let $(X_\alpha, \mathcal{T}_\alpha)$ be a topological space with the topology $\mathcal{T}_\alpha$. By $\prod_{\alpha \in A} X_\alpha$, we mean the set of all maps $x: A \to \bigcup_{\alpha \in A} X_\alpha$ such that $x(\alpha) = x_\alpha \in X_\alpha$ for all $\alpha \in A$. The $\prod_{\alpha \in A} X_\alpha$ is called a *product*. If A is a finite set, then indeed $\prod_{\alpha \in A} X_\alpha$ is a finite product, as seen above. If A is countable (i.e., there is a one-to-one map between the elements of A and those of positive integers), then $\prod_{\alpha \in A} X_\alpha$ is called the *countable product* and is often written $\prod_{k=1}^{\infty} X_k$. If A is arbitrarily infinite and not countable, the product $\prod_{\alpha \in A} X_\alpha$ is called the *arbitrary infinite product* or simply *arbitrary product*. Here, again, the map $p_\alpha: \prod_{\alpha \in A} X_\alpha \to X_\alpha$ defined by $p_\alpha(x) = x_\alpha \in X_\alpha$ is called the αth projection.

To define a topology on $\prod_{\alpha \in A} X_\alpha$, we consider the family $\mathcal{B}$ of all subsets $P = \prod_{\alpha \in A} P_\alpha$ of $\prod_{\alpha \in A} X_\alpha$ in which P_α is an open subset of X_α for a finite number of α's, say, α_1, $\alpha_2, \ldots, \alpha_n$ and $P_\alpha = X_\alpha$ for all $\alpha \neq \alpha_1, \alpha_2, \ldots, \alpha_n$. Then $\mathcal{B}$ forms a base of a topology of $\prod_{\alpha \in A} X_\alpha$, called the *product topology*. With the product topology on $\prod_{\alpha \in A} X_\alpha$, each αth projection becomes continuous and open.

The product topology on $\prod_{\alpha \in A} X_\alpha$ is very useful. For example, a map f of any topological space $(X, \mathcal{T})$ into $\prod_{\alpha \in A} X_\alpha$ is continuous iff each composition map $p_\alpha \circ f: X \to X_\alpha$ is continuous.

As for finite products, one can define productive property for infinite products (see Table III).

C. Quotient Topology

Let $(X, \mathcal{T})$ be a topological space, Y an arbitrary set, and $f: X \to Y$ a surjective map. We endow Y with a topology that will make f continuous. For this, we designate a

set Q of Y open whenever $f^{-1}(Q) = \{x \in X : f(x) \in Q\}$ is open in X. The collection of all such Q's defines a topology on Y, called the *quotient topology*. Clearly, if Y is endowed with the quotient topology, then $f : X \to Y$ becomes continuous.

Note that the quotient topology on Y depends on the topology of X and the map f. Indeed, one can always endow Y with the indiscrete topology to make f continuous. Among the topologies on Y that make f continuous, the quotient topology is the finest topology.

Like the product topology, the quotient topology is also very useful. For instance, suppose that Y is endowed with the quotient topology defined by $(X, \mathcal{T})$ and f. Let $(Z, \mathcal{T}')$ be a topological space and $g : Y \to Z$ a map. Then g is continuous iff $g \circ f : X \to Z$ is continuous.

IV. SEPARATION AXIOMS

The properties that appear to be obvious for real numbers are not easily available in general topological spaces. For instance, if a, b are two distinct real numbers, then for $\varepsilon = |a - b| > 0$, open intervals $A = \{x \in \mathbb{R} : |x - a| < \frac{1}{2}\varepsilon\}$ and $B = \{x \in \mathbb{R} : |x - b| < \frac{1}{2}\varepsilon\}$ are disjoint with $a \in A$ and $b \in B$. It is this kind of property that may not hold in general topological spaces. Such properties are termed *separation axioms*, also known as Alexandroff–Hopf Trennungaxioms.

One of the reasons separation axioms are necessary in the study of topological spaces is that they enable us to decide about the uniqueness of limits of convergent sequences or nets. This is not a trivial reason, for in an indiscrete space, no sequence or net has a unique limit. As a matter of fact, every sequence or net converges to every point of the indiscrete space. This abnormality makes the indiscrete spaces useless.

Now we consider the separation axioms.

A. T_0-Axiom

A topological space is said to satisfy the T_0-axiom or is called a *T_0-space* if at least one member of any pair of distinct points has an open neighborhood that does not contain the other. This axiom is not strong enough to yield any worthwhile results. Note that an indiscrete space containing more than one point is not a T_0-space.

B. T_1-Axiom

A topological space is said to satisfy the T_1-axiom or is called a T_1-space if each member of any pair of distinct points has an open neighborhood that does not contain the other. Although obviously better than the T_0-axiom, this axiom is not strong enough to guarantee the uniqueness of the limit of convergent sequences or nets. However, it does give us that each singleton (the set consisting of a single point) is a closed set and conversely (i.e., if each singleton is a closed set, the topological space must be a T_1-space). The class of T_1-spaces is hereditary and productive (see Table III).

C. T_2-Axiom

This is the most desirable separation axiom for the above-mentioned purpose and other reasons. A topological space is said to satisfy the T_2-axiom or is called a *T_2-space* or *Hausdorff space* (after F. Hausdorff) if both members of any pair of distinct points have disjoint open neighborhoods. This is indeed the weakest separation axiom that guarantees the uniqueness of the limit of a convergent sequence or net. As a matter of fact, a topological space is a Hausdorff space iff each convergent net or filter has the unique limit. The T_2-axiom is also equivalent to having the graph of the identity map closed. This class of spaces is also hereditary and productive.

There is another group of separation axioms that separate points from closed sets or closed sets from closed sets, as follows.

D. Regularity

A topological space is called regular (R) if for any closed set C and any point x not in C, there are disjoint open sets U and V such that $C \subset U$ and $x \in V$. Regularity is equivalent to the criterion that each point of the space have a base of closed neighborhoods. In general, regularity and the T_2-axiom have nothing in common. Therefore, to make them comparable, one defines *T_3-spaces* as those that are regular and satisfy the T_1-axiom. Now T_3-spaces are properly contained in the class of T_2-spaces and therefore satisfy the above-mentioned desirable property of uniqueness of limits of convergent nets in topological spaces as well. Moreover, they are hereditary and productive.

E. Complete Regularity

A topological space $(X, \mathcal{T})$ is said to be completely regular (CR) if for each closed set C of X and any point x not in C there is a continuous function $f : X \to [0, 1]$ such that $f(C) = 0$ and $f(x) = 1$. Complete regularity implies regularity but may fail to satisfy the T_2-axiom. However, if one joins the T_1-axiom with complete regularity, the resulting space does satisfy the T_2-axiom. A completely

regular T_1-space is generally known as a $T_{3\,1/2}$-space or *Tychonoff space* (after the Russian mathematician A. Tychonoff, who contributed a great deal to topology). Thus, each Tychonoff space is a Hausdorff space. The significance of Tychonoff spaces among others lies in the fact that on such spaces exist nonconstant continuous functions. Even on regular spaces there may not exist any nonconstant continuous function. Both classes of completely regular and Tychonoff spaces satisfy hereditary and productive properties.

F. Normality

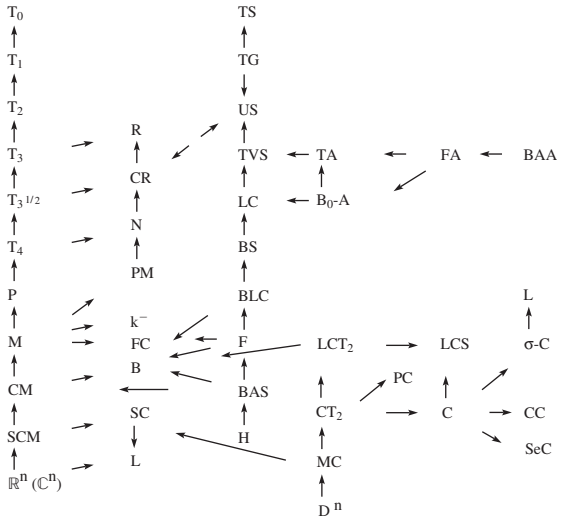
A topological space is said to be normal (N) if for any pair of disjoint closed subsets C_1, C_2 there are disjoint open sets U_1, U_2 such that $C_1 \subset U_1$ and $C_2 \subset U_2$. A useful characterization of normality is as follows. For each closed set C and open set U with $C \subset U$, there is an open set V such that $C \subset V \subset \bar{V} \subset U$. Even this apparently stronger separation axiom does not imply a Hausdorff space or the T_2-axiom, since singletons in a normal space need not be closed sets. Thus, to make it comparable with regularity and the T_2-axiom, one attaches the T_1-axiom to normality. A normal T_1-space is generally called a *T_4-space*. Now it is clear that each T_4-space is a T_3-space and hence Hausdorff, as shown earlier.

The axiom of complete regularity mentioned between those of regularity and normality apparently looks like an odd person out, but this is not the case, thanks to Urysohn's lemma: Let C_1, C_2 be any two disjoint closed subsets of a normal space $(X, \mathcal{T})$. Then there exists a continuous function $f : X \to [0, 1]$ such that $f(C_1) = 0$ and $f(C_2) = 1$. From this it follows that a T_4-space is a Tychonoff space, which in turn implies that it is a T_3-space and hence Hausdorff.

Extensions of functions from a subspace to the whole space play an important role in almost every aspect of mathematics. Thus, the question is which of the above axioms, if any, ensures extensions. We are indebted to H. Tietze for the so-called Tietze's extension theorem: If Y is a closed subspace of a T_4-space $(X, \mathcal{T})$ and $f : Y \to \mathbb{R}$ a continuous map, there exists a continuous map $\bar{f} : X \to \mathbb{R}$ such that $\bar{f}(x) = f(x)$ for all $x \in Y$.

There are other separation axioms finer or stronger than normality for which the reader is referred to texts on topology. However, normality is sufficient enough to carry out major parts of mathematical analysis. As noted in the preceding paragraphs, of the six separation axioms, T_0 is the weakest and T_4 the strongest. Consult Table II for a complete picture of these implications. For information about whether these six classes of spaces have hereditary or productive properties (or both), as mentioned in the preceding section, see Table III.

TABLE II Interconnection between Topological Spaces[a]



[a] A → B means that A is contained in B.

V. SPECIAL TOPOLOGICAL SPACES

In this section we consider some of the most useful topological spaces that are frequently mentioned and studied in topology and functional analysis (a modern branch of mathematics that deals simultaneously with abstract and concrete analysis).

A. Metric Spaces and Metrization

An important class of topological spaces that subsumes the classical spaces $\mathbb{R}^n$ or $\mathbb{C}^n$ is that of metric spaces.

Let X be a set. A function $d : X \times X \to \mathbb{R}$ is called a *metric* of X if d satisfies the following axioms:

(i) $d(x, y) \geq 0$ for all $x, y \in X$.
(ii) $d(x, y) = 0$ if $x = y$.
(iii) $d(x, y) = 0$ implies that $x = y$.
(iv) $d(x, y) = d(y, x)$.
(v) $d(x, y) \leq d(x, z) + d(z, y)$ (called the *triangle inequality*).

If d satisfies only (i), (ii), (iv), and (v), then d is called a *pseudometric* on X. A set X with a metric (or pseudometric) d is called a *metric* (or *pseudometric*) (M or PM, respectively) *space* and is sometimes denoted (X, d).

Each pseudometric (in particular, metric) space is indeed a topological space, for a set P in X is open if for each $x \in P$ there is a real number $r > 0$ such that the *open ball* $B_r(x) = \{y \in X : d(y, x) < r\}$ of radius r, centered at x,

TABLE III Hereditary and Productive Properties of Some Important Spaces[a]

Name	Hereditary		Productive properties		
	Arbitrary	Closed	Finite	Countable	Arbitrary
Topological spaces					
C (compact)	N	Y	Y	Y	Y
CM (compact metric)	N	Y	Y	Y	N
FC (first countable)	Y	Y	Y	Y	N
LCS (locally compact space)	N	Y	Y	N	N
M (metric)	Y	Y	Y	Y	N
P (paracompact)	N	Y	N	N	N
T_0	Y	Y	Y	Y	Y
T_1	Y	Y	Y	Y	Y
T_2	Y	Y	Y	Y	Y
T_3	Y	Y	Y	Y	Y
$T_{3 1/2}$	Y	Y	Y	Y	Y
T_4	N	Y	N	N	N
Topological vector spaces (vector subspaces)					
TVS (topological vector space)	Y	Y	Y	Y	Y
BAS (Banach space)	N	Y	Y	N	N
F (Fréchet)	N	Y	Y	Y	N
H (Hilbert)	N	Y	Y	N	N
Topological algebras (subalgebras)					
TA (topological algebra)	Y	Y	Y	Y	Y
BAA (Banach algebra)	N	Y	Y	N	N
B_0-A (B_0-algebra)	N	Y	Y	Y	N
FA (Fréchet algebra)	N	Y	Y	Y	N

[a] N (no) indicates that the property does not hold; Y (yes), that it does hold.

is completely contained in P; in symbols $x \in B_r(x) \subset P$. With this definition of open sets in a metric space (X, d), it is easily verified that the collection of all such open sets in X defines a topology called the *metric topology* of X and the collection of all open balls forms a base of the metric topology.

To see pseudometric or metric spaces from the separation-axiom point of view, note that each pseudometric (metric) space is normal (T_4-space), thus forming a telescopic chain of spaces from metric spaces up to T_0-spaces (see Table II).

If we consider the countable collection $\{B_{1/n}(x)\}$ of open balls of radius $1/n$ ($n \geq 1$, integer), centered at x, it follows that a metric space is first countable. A metric space is second countable iff it is separable.

Metric spaces enjoy an important place among topological spaces because, unlike general topological spaces, in metric spaces one can assign distances $d(x, y)$ between two points—a feature very noticeable in Euclidean spaces $\mathbb{R}^n$. Thus, one can study the notions of Cauchy sequences and uniformly continuous functions in metric spaces, which cannot be defined in general topological spaces.

A sequence $\{x_n\}$ in a metric space (X, d) is said to *converge* to $x \in X$ if for each arbitrary real number $\varepsilon > 0$ there is a positive integer n_0 such that $d(x_n, x) < \varepsilon$ whenever $n \geq n_0$. $\{x_n\}$ is called a *Cauchy sequence* if for each real number $\varepsilon > 0$ there is a positive integer n_0 such that $d(x_n, x_m) < \varepsilon$ whenever $n, m \geq n_0$.

Every convergent sequence is a Cauchy sequence, but the converse is not true. [For example, the sequence of rational numbers ≥ 1 and $< \sqrt{2}$, approximating $\sqrt{2}$, converges to $\sqrt{2}$ (which is not a rational number) and is a Cauchy sequence in the metric space of rational numbers.] A metric space (X, d) is called *complete* if every Cauchy sequence in it is convergent. (The set $\mathbb{R}$ of all real numbers is complete by the so-called Cauchy criterion for convergence of a real sequence.) We denote complete metric spaces CM and separable complete metric spaces SCM.

One can obtain topological information about closed sets by convergent sequences in a metric space (X, d) as follows. A subset C of X is closed iff for each sequence $\{x_n\}$ in C converging to x, it follows that $x \in C$. Since metric spaces, being T_4-spaces, are Hausdorff, the limits of convergent sequences are unique. Furthermore, metric spaces clearly enjoy the hereditary property but not the general productive property (see Table III). However, if (X_k, d_k), $k = 1, 2, \ldots$, is a countable family of metric

spaces, then the countable product $X = \prod_{k=1}^{\infty} X_k$ is a metric space with the metric

$$d(x, y) = \sum_{k=1}^{\infty} \frac{1}{2^k} \frac{d_k(x_k, y_k)}{1 + d_k(x_k, y_k)},$$

in which both $x = \{x_k\}$ and $y = \{y_k\}$ belong to X. In particular, the space ω of all real sequences is a metric space.

As pointed out above by the example of rational numbers, not every metric space is complete. However, there is a process, called the completion, by which an incomplete metric space can be completed. More precisely, if (X, d) is an incomplete metric space, there exists a unique complete metric space $(\hat{X}, \hat{d})$ containing a dense copy of $(X, d) \cdot (\hat{X}, \hat{d})$, called the *completion* of (X, d). For example, the set of all real numbers is the completion of the set of all rational numbers.

For metric spaces, the continuity of a map can be described as follows. Let (X, d) and (Y, d') be two metric spaces. A map $f : X \to Y$ is continuous at $x_0 \in X$ if and only if for each $\varepsilon > 0$ there is $\delta > 0$ such that for all $x \in X$, $d(x, x_0) < \delta$ implies that $d'(f(x), f(x_0)) < \varepsilon$. This is also equivalent to the so-called *sequential continuity* at x_0 : for all sequences $\{x_n\}$ in X converging to x_0, the sequence $\{f(x_n)\}$ converges to $f(x_0)$. A map $f : X \to Y$ is called *uniformly continuous* if for each $\varepsilon > 0$ there is $\delta > 0$ such that for all $x, y \in X$, $d(x, y) < \delta$ implies that $d'(f(x), f(y)) < \varepsilon$. If for all $x, y \in X$, $d'(f(x), f(y)) = d(x, y)$, then $f : X \to Y$ is called an *isometry* and thus f embeds X into Y in such a way that the metric induced from Y on $f(X)$ coincides with the metric of X. That is why it is sometimes called an *isometric embedding* of X into Y. An embedding need not be surjective. However, if $f(X) = Y$, then f is called an *isometric homeomorphism*. In this case, X and Y are set-theoretically and topologically the same.

A subset of a topological space is called an F_σ-set (respectively, G_δ-set) if it can be written as the countable union (respectively, countable intersection) of closed (respectively, open) sets. A subset A of a topological space is an F_σ-set (G_δ-set) iff its complement A^c is a G_δ-set (respectively, F_σ-set). Moreover, a countable union (countable intersection) of F_σ-sets (respectively, G_δ-sets) is an F_σ-set (respectively, G_δ-set). Also, a finite intersection (finite union) of F_σ-sets (respectively, G_δ-sets) is an F_σ-set (respectively, G_δ-set). Every closed (open) set of a metric space is a G_δ-set (respectively, F_σ-set). In particular, each closed (open) subset of $\mathbb{R}$ is a G_δ-set (respectively, an F_σ-set).

Since it is easier to grasp the topology of a metric space than that of general topological spaces, it is important to know when a topological space can be endowed with a metric topology. Indeed, indiscrete topology can never be defined by a metric, since a metric space is always Hausdorff, whereas the indiscrete space does not even satisfy the T_0-axiom. On the other hand, the discrete topology can be given by the trivial metric $d(x, y) = 1$ or 0 according to whether $x \neq y$ or $x = y$.

Whenever the topology of a topological space can be defined by a metric, the space is called *metrizable*. A nontrivial notable result in this regard is due to P. Urysohn, which goes by the name Urysohn's metrization theorem: Each second countable T_3-space is metrizable.

A separable complete metric space is called a *Polish space* and a metric space which is a continuous image of a Polish space is called a *Souslin space*. A closed subspace of a Polish (respectively, Souslin) space is also a Polish (respectively, Souslin) space. The same is true for countable products. Moreover, an open subset of a Polish space is also a Polish space. For example, $\mathbb{R}$ and the set of all irrational numbers endowed with the metric topology induced from $\mathbb{R}$ are Polish spaces. We give further properties of these spaces in the sequel.

B. Compactness and Compactifications

Another class of very important and useful topological spaces in topology as well as in analysis is that of compact spaces. Let $(X, \mathcal{T})$ be a topological space. A collection $\{P_\alpha\}$ of subsets of X is said to be a *covering* of a set $E \subset X$ if $E \subset \bigcup_\alpha P_\alpha$; in other words, each element of E belongs to some P_α. If $E = X$, then $\{P_\alpha\}$ is a covering of X if $X = \bigcup_\alpha P_\alpha$. If each member of the covering $\{P_\alpha\}$ of E is open, then $\{P_\alpha\}$ is called an *open covering* of E. If the number of sets in a covering is finite, it is called a *finite covering*.

1. Compact Spaces

A topological space $(X, \mathcal{T})$ is said to be compact (C) if from each open covering of X it is possible to extract a finite open covering of X. $(X, \mathcal{T})$ is called a *Lindelöf space* (L) if from each open covering $\{P_\alpha\}$ of X it is possible to extract a countable subcollection $\{P_{\alpha_i}, i \geq 1\}$ such that $X = \bigcup_{i=1}^{\infty} P_{\alpha_i}$. It is clear that each compact space is a Lindelöf space. However, the converse is not true (e.g., $\mathbb{R}$, with the natural topology). One of the celebrated theorems in analysis, called the Heine–Borel theorem, tells us that each closed bounded interval $[a, b]$, $-\infty < a \leq b < \infty$, of $\mathbb{R}$ is compact, but $\mathbb{R}$ itself is not. Similarly, each closed disk D^n in $\mathbb{R}^n$ is compact.

Compactness has several characterizations. Some of them are as follows. $(X, \mathcal{T})$ is compact $\Leftrightarrow$ for each family $\{C_\alpha\}$ of closed sets in X with the finite-intersection

property (i.e., for any finite subcollection $\{C_{\alpha_i}\}_{i=1}^n$ of $\{C_\alpha\}$, $\bigcap_{i=1}^n C_{\alpha_i} \neq \varnothing$), we have $\bigcap_a C_\alpha \neq \varnothing \Leftrightarrow$ each net or filter of X has a cluster point $\Leftrightarrow$ each ultrafilter in X is convergent.

Since the indiscrete space has only two open sets (actually only one nonempty open set), it is always compact, although not Hausdorff. On the other hand, a discrete space is compact iff it is a finite set. A Hausdorff compact space, however, is a T_4-space and so is normal.

With regard to the hereditary property, compact spaces do not have it (see Table III), but a closed subset of a compact space is compact. A far-reaching result for application purposes is concerned with the product of compact spaces and goes by the name of the *Tychonoff theorem*: If $(X_\alpha, \mathcal{T}_\alpha)$ $(\alpha \in A)$ is a family of compact spaces, the arbitrary product $\prod_{\alpha\in A} X_\alpha$ is also compact under the product topology. In addition, compact spaces are rich in other properties. For instance, the continuous image of a compact space is compact. Moreover, a continuous bijective map of a compact space onto a Hausdorff space is a homeomorphism. Every continuous real-valued function on a compact space is bounded; that is, there is $M > 0$ such that $|f(x)| \leq M$ for all $x \in X$.

Although a subspace of a compact space need not be compact, every subspace of a compact Hausdorff space is a Tychonoff space. This, among other properties, highlights the significance of a compact Hausdorff space. That is why it is sometimes called a *compactum*. Two esotoric examples of compacta are as follows.

a. Hilbert cube. Let $H^\omega = \{\{x_n\}, x_n \in \mathbb{R} : 0 \leq x_n \leq (1/n)\}$ with the metric $d(\{x_n\}, \{y_n\}) = \sqrt{\sum_{n=1}^\infty (x_n - y_n)^2}$. Then H^ω is a separable compact metric space, hence a second countable compactum and a Polish space. H^ω is called the Hilbert cube. It is a compact subset of the Hilbert space ℓ_2 (see Section VIII.B).

b. Cantor set. Let $A_0 = [0, 1]$, the closed unit interval. By removing the middle third of A_0 we obtain $A_1 = [0, 1/3] \cup [2/3, 1]$. Continuing this process, we obtain $A_n = [0, 1/3^n] \cup \ldots \cup [((3^n - 1)/3^n) - 1], n = 0, 1, 2, \ldots$. Each A_n being a closed subset of the compact space A_0, we obtain a nonempty compact set $\mathfrak{C} = \bigcap_{n=0}^\infty A_n$, called the *Cantor set.* Alternatively, one can describe $\mathfrak{C}$ as follows: Let $E = \{0, 2\}$ be the two-point set and let E^ω be the countable product of E by itself, which is the set of all sequences with entries 0 or 2. For each $x = \{x_i\} \in E^\omega$, put $\varphi(x) = \sum_{i=1}^\infty (x_i/3^i)$. φ is an injective map of E^ω into $[0, 1]$. It can be shown that $\mathfrak{C} = \varphi(E^\omega)$, the Cantor set. $\mathfrak{C}$ is an uncountable compactum, has no isolated points, contains no interval of $\mathbb{R}$, and is perfect (i.e., closed and dense in itself).

2. Countably and Sequentially Compact Spaces

A topological space $(X, \mathcal{T})$ is called countably compact (CC) if from each countable open covering of X it is possible to extract a finite open covering. $(X, \mathcal{T})$ is called sequentially compact (SeC) if every sequence in it contains a convergent subsequence.

It is easily seen that each compact space is countably compact as well as sequentially compact, but a countably compact or a sequentially compact space may fail to be compact. In general, there is no connection between countable compactness and sequential compactness, although the two notions coincide on T_1-spaces. A metric space is compact iff countably compact iff sequentially compact. For hereditary and productive properties, see Table III.

3. σ-Compact Spaces

A topological space is called σ-compact if it is the countable union of compact spaces. Indeed, each compact space is σ-compact, but the converse is not true (e.g., $\mathbb{R}$).

Another frequently used notion of compactness in topology is as follows.

4. Locally Compact Spaces

A topological space $(X, \mathcal{T})$ is said to be locally compact (LCS) if each point of X has a neighborhood whose closure is compact.

Since a compact space is always a closed and compact neighborhood of each of its points, it follows that each compact space is locally compact, but the converse is not true (e.g., $\mathbb{R}$). In general, a locally compact space does not satisfy any separation axiom, but a Hausdorff locally compact space is a Tychonoff space and hence a T_3-space. Note that a Hausdorff locally compact space need not be normal.

Although locally compact spaces need not have the hereditary property, a closed subspace of a locally compact space is locally compact (see Table II). Moreover, a continuous, open, and surjective image of a locally compact space is locally, compact. Locally compact spaces are good for one-point compactifications (see below).

5. Hemicompact Spaces

A topological space $(X, \mathcal{T})$ is called hemicompact if there exists a countable family $\{K_n\}$ of compact subsets K_n of X such that $X = \bigcup_{n=1}^\infty K_n$ and each compact subset of X is contained in some K_n. ($K_n^>$s are said to form a fundamental system of compact subsets of X.) Each locally compact hemicompact space is σ-compact. Each locally compact

hemicompact (hence σ-compact) metrizable space is a Polish space.

6. MB Spaces

A metric space (X, d) is said to be an MB space if each closed bounded subset of X is compact. [Note that a subset B of X is bounded if the diameter $\delta(B) = \sup\{d(x, y) : x, y \in B\}$ is finite.] Clearly, each compact metric space is an MB space and each MB space is complete, separable, locally compact and σ-compact (hence hemicompact), and also a Lindelöf, Polish space. As examples, $\mathbb{R}^n$ (n-dimensional euclidean space) is an MB space, but the set of rational numbers with the induced metric from $\mathbb{R}$ is not.

Each continuous map of an MB space into a metric space carries bounded sets into bounded sets. A metric space (F, d') which is the continuous image of an MB space (E, d) such that the inverse image of a bounded set of F is bounded in E is an MB space. Indeed, a closed subset of an MB space is an MB space and, so, is the countable product of compact metric spaces.

7. Pseudocompact Spaces

A Hausdorff topological space $(X, \mathcal{T})$ is called pseudocompact (PC) if every continuous real-valued function on X is bounded (see Section V.B.1). Every Hausdorff compact or countably compact space is pseudocompact. The converse is not true, however. In particular, each closed bounded interval $[a, b]$ is pseudocompact. See Table III for other implications.

8. Paracompact Spaces

A Hausdorff topological space $(X, \mathcal{T})$ is called paracompact (P) if each open covering $\{P_\alpha\}$ of X has a *refinement* $\{Q_\beta\}$ (i.e., there exists another covering $\{Q_\beta\}$ of X such that each P_α contains some Q_β) such that $\{Q_\beta\}$ is *neighborhood finite* (i.e., for each $x \in X$ there is an open neighborhood U of x such that U intersects with only a finite number of Q_β's).

Since a finite open subcovering of an open covering is a refinement, each Hausdorff compact space is paracompact, but the converse is not true. Each paracompact space is a T_4-space (hence normal) and each metric space is paracompact. Now the significance of paracompact spaces becomes very notable. These spaces are useful for the extension of continuous functions from subspaces to the whole spaces, as well as for metrization. Paracompact spaces do not have the hereditary or productive properties, but a closed subspace of a paracompact space is paracompact.

9. Compactifications

A topological space $(\hat{X}, \hat{\mathcal{T}})$ is said to be a compactification of a topological space $(X, \mathcal{T})$ if $(\hat{X}, \hat{\mathcal{T}})$ is compact and contains a homeomorphic image of $(\hat{X}, \mathcal{T})$ that is dense in $(X, \hat{\mathcal{T}})$.

Among compactifications, two are significant for applications: one-point compactification and Stone-Čech compactification. The former is the smallest, whereas the latter is the largest in a certain sense. The former requires local compactness, whereas the latter needs complete regularity.

Let $(X, \mathcal{T})$ be a Hausdorff locally compact space. Let ω be any point outside X. Put $\hat{X} = X \cup \{\omega\}$ and endow $\hat{X}$ with the topology $\hat{\mathcal{T}}$ defined as follows. Each open subset of X is open in $\hat{X}$ (i.e., each $U \in \mathcal{T}$ is also in $\hat{\mathcal{T}}$) and any subset $\hat{U}$ of $\hat{X}$ that contains ω is in $\hat{\mathcal{T}}$ provided that $\hat{X} \setminus \hat{U}$ is compact in X. With this prescription of open sets, it is easy to verify that $\hat{\mathcal{T}}$ defines a topology such that $(\hat{X}, \hat{\mathcal{T}})$ is a Hausdorff compact space in which $(X, \mathcal{T})$ is dense. Moreover, $\hat{X} \setminus X = \{\omega\}$ consists of a single point. $(\hat{X}, \hat{\mathcal{T}})$ is called the *one-point compactification of* $(X, \mathcal{T})$.

As an example, $\mathbb{R}$ is a Hausdorff locally compact space. If we put $\hat{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, then $\hat{\mathbb{R}}$ is the one-point compactification of $\mathbb{R}$. It is easy to see that $\hat{\mathbb{R}}$ is homeomorphic with the circle: $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$.

For the Stone-Čech compactification, let $(X, \mathcal{T})$ be a Tychonoff space. Then the set $C(X)$ of all continuous functions from X into $[0, 1]$ contains nonconstant continuous functions, provided that X is not trivial (i.e., it consists of more than one point). If we put $I = [0, 1]$, then $I^{C(X)} = \prod_{f \in C(X)} I_f$, $(I_f = I)$, the infinite product is compact by the Tychonoff theorem (see Section V.B.1). If to each $x \in X$, we assign $(f(x))_{f \in C(X)} \in I^{C(X)}$, we get an injective map. The closure of this injective image, denoted βX, in the compact space $I^{C(X)}$ is also compact. βX is called the *Stone-Čech compactification* of X.

10. Perfect and Proper Maps

Let $(X, \mathcal{T})$ and $(Y, \mathcal{T}')$ be two topological spaces. A map $f : X \to Y$ is called *inversely compact* (or simply *compact*) if the inverse image $f^{-1}(\{y\})$ of each singleton $\{y\}$ in Y is compact in X. Note that a continuous map need not be compact, nor need a compact map be continuous, or closed, or open. A continuous, closed, compact, surjective map is called *perfect.*

Perfect maps transport a number of properties of their domains to their ranges. For instance, if $f : (X, \mathcal{T}) \to (Y, \mathcal{T}')$ is perfect and if X is Hausdorff or regular or completely regular or metrizable or second countable or compact or countably compact or Lindelöf or locally compact or paracompact, so is Y, respectively. Perfect maps are often used in category theory.

A map $f:(X,\mathcal{T})\to(Y,\mathcal{T}')$ is called *proper* (à la Bourbaki) if for each compact subset B of X, $f(B)$ is compact in Y, and for each compact subset C of Y, $f^{-1}(C)$ is compact in X. Each continuous map between compacta is proper and each continuous closed compact map of a T_2-space to a T_2-space is proper.

C. Baire Spaces and *k*-Spaces

Apart from those spaces dealt with in the preceding section, there are two more classes that are frequently used in topology and functional analysis. They are Baire spaces and k-spaces.

1. Baire Spaces

A subset A of a topological space $(X,\mathcal{T})$ is called *nondense* or *nowhere dense* if $(\bar{A})^0=\varnothing$ (i.e., the interior of the closure of A is empty). A countable union of nondense sets is called a *set of the first category* or *a meager* set. A set that is not of the first category is called a *set of the second category*. The complement of a set of the first category is called a *residual set*.

A topological space $(X,\mathcal{T})$ is called a *Baire space* (B) if for any countable collection of closed nondense sets $\{A_n\}$ such that $X=\bigcup_{n=1}^{\infty}A_n$, there is at least one n for which $A_n^0\neq\varnothing$. The class of all Baire spaces includes two of the important classes of topological spaces studied above, namely, the class of Hausdorff locally compact (hence, compact) spaces and the class of all complete metric spaces. The fact that each complete metric space is a Baire space is known as the *Baire category theorem.* Moreover, each Baire space is of the second category, and each open subset of a Baire space is a Baire space.

Let τ denote a cardinal number and let $D(\tau)$ be a set of real numbers of cardinality τ. Let $B_\tau=\{\{x_i\}:x_i\in D(\tau)\}$. For $x=\{x_i\}$, $y=\{y_i\}$ in B_τ, put $d(x,y)=1/k$, where k is the least integer i such that $x_i\neq y_i$. Then d is a metric on B_τ and the metric space (B_τ,d) is called a *Baire space* of *weight* τ. The metric topology of B_τ has a base of closed–open sets.

2. *k*-Spaces

A topological space $(X,\mathcal{T})$ is called a k-space if each subset C of X is closed whenever for each compact subset K of X, $C\cap K$ is closed.

All Hausdorff locally compact (hence, compact) spaces as well as metric spaces are included in the class of all k-spaces. However, a k-space need not be a Baire space. Although the class of k-spaces does not have the hereditary property, it is true that a closed subspace of a Hausdorff k-space is a k-space. Nor does the productive property hold in the class of k-spaces. A notable property for k-spaces is as follows. Each real-valued function on a k-space is continuous iff its restriction on each compact subset is continuous. Clearly, this generalizes the known fact about real-valued functions defined on the set of real numbers.

A completely regular T_2-space (i.e., a Tychonoff space) $(X,\mathcal{T})$ is called a k_r-space if each real-valued function on X whose restriction to each compact subset of X is continuous is indeed continuous on X. Each k-space is a k_r-space, but the converse is not true. For instance, the uncountable product $\mathbb{R}^\nu$ of $\mathbb{R}$, endowed with the product topology, is a k_r-space but not a k-space. One of the very useful properties of k-spaces is that each proper map between two k-spaces is continuous and closed.

D. Uniform Spaces and Uniformization

There are two important reasons for studying uniform spaces. One is to define uniform continuity, and the other is to define completeness for spaces more general than metric spaces. After a little thought, it becomes obvious to the reader that these two important concepts from classical analysis cannot be expressed in general topological spaces.

Let X be any set and $X^2=X\times X$, the Cartesian product of X by itself. Let $\mathcal{U}=\{U\}$ be a filter in X^2 satisfying the following:

(i) Each U in $\mathcal{U}$ contains the diagonal set $\{(x,y)\in X^2:x=y\}$.
(ii) For each U in $\mathcal{U}$, $U^{-1}=\{(y,x):(x,y)\in U\}$ is also in $\mathcal{U}$.
(iii) For each U and $\mathcal{U}$ there is a V in $\mathcal{U}$ such that $V\circ V\subset U$, in which $V\circ V=\{(x,y)\in X^2$: for some z both (x,z) and (z,y) are in $V\}$.

Then the filter $\mathcal{U}=\{U\}$ is called a *uniformity* and the pair $(X,\mathcal{U})$ is called a *uniform space* (US).

Each metric space (X,d) becomes a uniform space with the collection $\{U_\varepsilon\}$ of sets in X^2 as its uniformity, where $U_\varepsilon=\{(x,y)\in X^2:d(x,y)<\varepsilon\}$ for all real numbers $\varepsilon>0$.

Each uniform space $(X,\mathcal{U})$ can be endowed with a topology called the *uniform topology* defined as follows. A subset P of X is called open in the uniform topology if for each $x\in P$ there is U in $\mathcal{U}$ such that $\{y\in X:(x,y)\in U\}\subset P$. Each uniform space with the uniform topology is completely regular. Thus, if the uniform topology is Hausdorff, the uniform space becomes a Tychonoff space.

A map f of a uniform space $(X,\mathcal{U})$ into another uniform space $(Y,\mathcal{V})$ is called *uniformly continuous* if for each

V in $\mathcal{V}$ there exists U in $\mathcal{U}$ such that $(f(x), f(y)) \in V$ whenever $(x, y) \in U$. In particular, for metric spaces $(X, d), (Y, d')$, the uniform continuity of f has the following formulation. For each $\varepsilon > 0$ there exists $\delta > 0$ such that $d'(f(x), f(y)) < \varepsilon$ whenever $d(x, y) < \delta$.

Indeed, each uniformly continuous map is continuous. However, the converse is not true [e.g., $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = x^2$]. However, if $(X, \mathcal{T})$ is a compact uniform space, each real-valued continuous function on X is uniformly continuous. In particular, each continuous map $f : [a, b] \to \mathbb{R}$ is uniformly continuous.

A filter $\{F_\alpha\}$ (or net $\{x_\alpha\}$) in a uniform space $(X, \mathcal{U})$ is called a *Cauchy filter* (or *Cauchy net*) if for each U in $\mathcal{U}$ there exists F_α such that the set $\{(x, y) : x, y \in F_\alpha\} \subset U$ [or there exists α_0 such that for all $\alpha, \beta \geq \alpha_0$, $(x_0, x_\beta) \in U$]. Indeed, each convergent filter (or net) is a Cauchy filter (or net), but not conversely. Whenever the converse holds, the uniform space is said to be *complete*. As for metric spaces, each incomplete Hausdorff uniform space $(X, \mathcal{T})$ can be completed to a complete Hausdorff uniform space $(\tilde{X}, \tilde{\mathcal{T}})$. $(\tilde{X}, \tilde{\mathcal{T}})$ is called the *completion of* $(X, \mathcal{T})$.

The class of all complete Hausdorff uniform spaces includes all compact Hausdorff uniform spaces and is productive but not hereditary. However, a closed subset of a complete Hausdorff uniform space is complete.

A topological space $(X, \mathcal{T})$ is said to be *uniformizable* if there exists a uniformity $\mathcal{U}$ with respect to which $\mathcal{T}$ becomes the uniform topology. Indeed, each metric space is uniformizable, as shown above. An important characterization of uniformization is the following. A topological space $(X, \mathcal{T})$ is uniformizable iff it is completely regular iff $\mathcal{T}$ is defined by a family $\{d_\alpha\}$ of pseudometrics (see Sections II.A and V.A). A topological space whose topology is defined by a family of pseudometrics is called a *gauge space*.

E. Sequential Spaces

First, we note that a sequence $\{x_n\}$ in a nonmetric topological space $(X, \mathcal{T})$ is said to converge to a point $x \in X$ if for each neighborhood U of x there is a positive integer n_0, depending upon U, such that for all $n \geq n_0$, $x_n \in U$. Indeed, this definition is equivalent to the convergence of a sequence in a metric space given in Section V.A.

A topological space $(X, \mathcal{T})$ is called a *sequential space* if for any subset A of x, $A \neq \bar{A}$, there is a sequence $\{x_n\}$ in A converging to a point of $\bar{A} \backslash A$. Each metrizable space (in particular, $\mathbb{R}^n$) is a sequential space, but $\mathbb{R}^{\mathbb{R}}$ (see Section VII), the set of all real functions on $\mathbb{R}$ endowed with the product topology, is not.

For a subset A of a sequential space, the set of all points $x \in X$ for which there exists a sequence in A converging to x, is called the *sequential closure* of A and is denoted $\bar{A}^s$. It is easily seen that $\bar{\varnothing}^s = \varnothing$, $A \subset \bar{A}^s \subset \bar{A}$, and $\overline{A \cup B}^s = \bar{A}^s \cup \bar{B}^s$ for all subsets A, B of X. In general, however, $\overline{\bar{A}^s}^s \neq \bar{A}^s$, i.e., the sequential closure operation is not idempotent, unlike the topological closure, where we have $\bar{\bar{A}} = \bar{A}$.

A topological space $(X, \mathcal{T})$ is called a Fréchet–Urysohn space if $\bar{A}^s = \bar{A}$ for all subsets A of X. Clearly, each Fréchet–Urysohn space is a sequential space but the converse is not true. A useful characterization of a Fréchet–Urysohn space is: A topological space $(X, \mathcal{T})$ is a Fréchet–Urysohn space iff each subspace of $(X, \mathcal{T})$ is a sequential space. A first countable (hence metrizable) space is a Fréchet–Urysohn space. Each continuous map $f : (X, T) \to (Y, T')$ is sequentially continuous, but the converse is not true. However, if $(X, \mathcal{T})$ is a sequential space, then each sequential map of $(X, \mathcal{T})$ into $(Y, \mathcal{T}')$ is continuous. Indeed, on metric spaces, sequential continuity coincides with the usual concept of continuity as noted above (see Section V.A).

F. Proximity Spaces

Let $\mathcal{P}(X)$ denote the class of all subsets of a set X. Let δ define a binary relation on $\mathcal{P}(X)$. If $A \delta B$, then A is said to be *near* B for A, B in $\mathcal{P}(X)$ and $A \not\delta B$ ($\not\delta$, the negation of δ) means that A is *distant* from B. The pair (X, δ) is called a *proximity space* if the following axioms hold:

(i) $A \delta B \Leftrightarrow B \delta A$.
(ii) $\{x\} \delta \{x\}$.
(iii) $A \delta (B \cup C) \Leftrightarrow A \delta B$ or $A \delta C$.
(iv) $A \not\delta \varnothing$ for all $A \subseteq X$.
(v) $A \not\delta B \Rightarrow$ there is $C \in \mathcal{P}(X)$ such that $A \not\delta C$ and $B \not\delta (X \backslash C)$, where A, B, C are subsets of X and $x \in X$. δ is called a *proximity* for X.

A proximity δ for (X, δ) induces a topology, called the *proximity topology*, on X as follows. Let $\mathcal{T}_\delta$ denote the collection of all subsets P of X such that for each $x \in P$, $\{x\} \not\delta P^c$. The collection $\mathcal{T}_\delta$ satisfies all the properties of a topology and $(X, \mathcal{T}_\delta)$ becomes a completely regular topological space.

For a proximity space (X, δ), $d(A, B) = 0 \Leftrightarrow A \delta B$ gives a metric, where $d(A, B) = \inf\{d(a, b) : a \in A, b \in B\}$. On the other hand, if (X, d) is a metric space, then the relation defined by $A \delta_d B \Leftrightarrow d(A, B) = 0$ gives a proximity δ_d and (X, δ_d) becomes a proximity space. Moreover, the proximity topology induced by δ_d coincides with the original metric topology of (X, d). Thus each metric space is a proximity space.

Similarly, if $(X, \mathcal{U})$ is a Hausdorff uniform space, then the relation δ_u, defined by $A\delta_u B \Leftrightarrow$ for each $U \in \mathcal{U}$ there exists an element $a \in A$ and an element $b \in B$ such that $(a, b) \in U$ gives a proximity relation on X. On the other hand, for each proximity space (X, δ), let $U_{A,B} = [(A \times B) \cup (B \times A)]^c$. Then the collection $\mathcal{U}_\delta = \{U_{A,B} : A \not\delta B\}$ forms a subbase of a uniformity for X. Moreover, the proximity derived from this uniformity coincides with the given proximity δ. Thus each Hausdorff uniform space admits a proximity called the *uniform proximity*.

The collection of all proximities on a set X can be ordered by inclusion. A proximity δ_1 is larger than a proximity δ_2 if $A\delta_1 B \Rightarrow A\delta_2 B$ for all $A, B \in \mathcal{P}(X)$. Under this ordering, there is a largest proximity which can be described as follows: $A \not\delta B \Leftrightarrow A \cap B = \emptyset$. The largest proximity induces a discrete topology.

For any proximity space (X, δ), $\mathrm{Cl}_\delta A = \{x \in X : \{x\}\delta A\}$ defines a closure operation on X which gives the same proximity topology. Moreover, $A\delta B \Leftrightarrow \mathrm{Cl}_\delta A \delta \mathrm{Cl}_\delta B$. If (X, δ) is a compact proximity space, then $A\delta B \Leftrightarrow \mathrm{Cl}_\delta A \cap \mathrm{Cl}_\delta B \neq \emptyset$. Every Hausdorff proximity space (X, δ) is a dense subspace of a unique compact Hausdorff space αX such that $A\delta B$ in X iff $\mathrm{Cl}_\delta A \cap \mathrm{Cl}_\delta B \neq \emptyset$ where $\mathrm{Cl}_\delta A$ is taken in αX. αX is called the *Smirnov compactification* of X.

G. Connected and Locally Connected Spaces

For the definition of these spaces, see Section I.A. In addition to the properties of these spaces given in that section, we have the following.

A continuous map of a topological space into another carries connected sets into connected sets. A product of connected spaces is connected. If A is a connected subset of a topological space and B is any set such that $A \subset B \subset \bar{A}$, then B is connected. Any interval of $\mathbb{R}$, including $\mathbb{R}$ itself, is connected. Each convex subset of a real topological vector space (see Section VIII.B) is connected. An interesting characterization of connected spaces is: A topological space $(X, \mathcal{T})$ is connected if and only if every continuous map f of X into a discrete space Y is constant, i.e., $f(X) = \{y\}$, $y \in Y$.

A characterization of locally connected spaces is: A topological space is locally connected if and only if the components of its open sets are open. Local connectedness is preserved under closed maps.

A connected, metrizable compactum $(X, \mathcal{T})$ is sometimes called a *continuum.* Let $(X, \mathcal{T})$ be a continuum and $(Y, \mathcal{T}')$ a Hausdorff space. If $f : (X, \mathcal{T}) \to (Y, \mathcal{T}')$ is a continuous surjective map, then $(Y, \mathcal{T}')$ is also a continuum. Further, if $(Y, \mathcal{T}')$ is locally connected, so is $(Y, \mathcal{T}')$.

A topological space each of whose components consists of a singleton is called *totally disconnected.* For example, the set of integers is totally disconnected.

H. Extensions and Embeddings

Recall Tietz's extension theorem (Section IV), which states that each continuous function from a closed subset Y of a normal space $(X, \mathcal{T})$ into $[0, 1]$ can be extended to a continuous function of X into $[0, 1]$. It can be shown that $[0, 1]$ can be replaced by any interval $[a, b]$, $-\infty < a < b < \infty$. Actually, $[a, b]$ can be replaced by unit ball $B_n = \{(x_1, \ldots, x_n), x_i \in \mathbb{R} : \sum_{i=1}^n |x_i^2| \leq 1\}$ of $\mathbb{R}^n$. Furthermore, similarly, a continuous map $f : Y \to S^n = \{(x_1, \ldots, x_n), x_i \in \mathbb{R} : \sum_{i=1}^n |x_i^2| = 1\}$ (unit sphere of $\mathbb{R}^n$) extends to a continuous map on a neighborhood of Y.

We note that the given function $f : Y \to [0, 1]$ in Tietz's extension theorem is bounded. It is possible to replace $[0, 1]$ by $\mathbb{R}$, i.e., the theorem is true even for unbounded functions. Furthermore, normality of $(X, \mathcal{T})$ can be relaxed under certain conditions.

Let Y be a compact subset of a Tychonoff space $(X, \mathcal{T})$ (note that each normal T_1-space is a Tychonoff space) and $f : Y \to [0, 1]$ a continuous function, then there exists a continuous function $\tilde{f} : (X, \mathcal{T}) \to [0, 1]$ such that $\tilde{f}(x) = f(x)$ for all $x \in Y$. Further, each continuous bounded function on a Tychonoff space $(X, \mathcal{T})$ extends to its Stone-Čech compactification βX. In this result, the implicit assumption of the range of f being in $[a, b]$ can be weakened under suitable conditions. For instance, a continuous map f of a Tychonoff space $(X, \mathcal{T})$ to a compactum Y can be extended to a continuous map $\tilde{f} : \beta X \to Y$ such that $\tilde{f}(x) = f(x)$ for $x \in X$.

Let $(X, \mathcal{U})$ be a uniform space with its uniform topology induced by $\mathcal{U}$, and Y a dense subspace of X. Then each uniformly continuous map f of Y into a complete Hausdorff uniform space Z extends to a uniformly continuous map $\tilde{f} : X \to Z$ such that $\tilde{f}(x) = f(x)$ for all $x \in Y$.

Whenever there is a bijective map between two sets, they are deemed to be set-theoretically the same. If there is a surjective homeomorphism between two topological spaces, they are regarded to be the same set-theoretically as well as topologically. However, if a homeomorphism is into, i.e., if there is a continuous open injective map f of a topological space $(X, \mathcal{T})$ into $(Y, \mathcal{T}')$, then f is called an *embedding* of X into Y. In this case, $f(X)$ need not be equal to Y but X is homeomorphic to $f(X)$ (a subset of Y) with the topology induced from Y. Embeddings of abstract topological spaces into given or known topological spaces are of paramount importance and interests.

Simple examples of embeddings are as follows: Each metric space can be embedded into a complete metric space (its completion). In particular, the set of rational numbers is embedded into reals. Each Hausdorff uniform space can be embedded into a complete Hausdorff uniform space. Somewhat more sophisticated embeddings are as follows. Each second countable T_3-space can be embedded into the Hilbert cube. Furthermore, a Hausdorff space embeds into the Cantor perfect set if and only if it has a countable base of closed–open sets. A metric space (X, d) can be embedded into a Baire space $B(\tau)$ of weight τ if and only if X has a base $\mathcal{B}$ of closed–open sets such that the cardinality of $\mathcal{B}$ is less than or equal to τ. Each uniform space embeds in a product of pseudometric spaces and each Hausdorff uniform space embeds into a product of metric spaces.

Topological spaces having a base of closed–open sets are called *0-dimensional.* Every 0-dimensional T_0-space is completely regular. In a compact 0-dimensional space, two disjoint closed subsets can be separated by two disjoint open sets whose union is the whole space. Every locally compact totally disconnected space is 0-dimensional.

I. Topologies Defined by an Order

Let $(X, \leq)$ be a linearly ordered set, i.e., for $x, y \in X$, "$x \leq y$" means that "x is less than or equal to y," and for all $x, y \in X, x \neq y$, either $x < y$ or $y < x$. The notation $(x, y) = \{z \in X : x < z < y\}$ is called an "interval" in X. The collection of all such intervals forms a base of a topology on X, called the *order topology*, and $(X, \leq)$ endowed with the order topology is called an *ordered space.* $\mathbb{R}$ is an ordered space. An ordered space which is sequential is first countable. A discrete space is an ordered space. Unlike metric spaces, a separable ordered space need not be second countable.

$\mathbb{R}^2$ with the usual metric topology is not an ordered space. Let $I^2 = [0, 1] \times [0, 1]$ be the unit square in $\mathbb{R}^2$. We define the ordering in I^2 as follows: $(a_1, b_1) \leq (a_2, b_2)$ if and only if $a_1 < a_2$ or $a_1 = a_2$ and $b_1 < b_2$. This is called *dexicographic ordering* on I^2. Thus one obtains an order topology on I^2. It is possible to show from this example that a subspace of an ordered space need not be an ordered space.

We end this section with two diagrams indicating the interconnection of some spaces studied above. (As usual, $A \to B$ means that A is contained in B.)

$$\begin{array}{ccccc} & & \text{locally compact } T_2 & & \\ & \nearrow & & \searrow & \\ \text{Compact } T_2 & & & & \text{Baire} \to k\text{-space} \to k_r\text{-space.} \\ & \searrow & & \nearrow & \\ & & \text{complete metric} & & \end{array} \tag{1}$$

$$\begin{array}{ccccccc} \text{Metric compact} & \to \text{compact} \to & \left\{\begin{array}{l}\text{locally compact}\\ \sigma\text{-compact}\end{array}\right\} & \to & \text{paracompact} \\ \downarrow & & & & \downarrow \\ \begin{array}{l}\text{locally compact}\\ \text{2nd countable}\end{array} & \to \text{Polish} \to \text{Souslin} & & & \to \text{metric} \to \text{normal.} \end{array} \tag{2}$$

As pointed out above, a continuous function may not have a fixed point in general. However, if the space is restricted, one has a different situation. Specifically, if $D^n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1^2 + \cdots + x_n^2 \leq 1\}$ is the closed n-dimensional disk, each continuous map $f : D^n \leftarrow D^n$ has a fixed point but not necessarily unique. This is known as the Brouwer fixed point theorem in the mathematical literature. This result has many extensions in functional analysis.

VI. FIXED POINTS

The information regarding those points that remain fixed under certain transformations is very useful. It can be put to work in many fields (e.g., differential equations and algebraic topology).

Let X be a set and $f : X \to X$ a self-map. $x \in X$ is called a *fixed point* of f if $f(x) = x$. Indeed, the identity map $i : X \to X, i(x) = x$, keeps every point of X fixed. On the other hand, $f : \mathbb{R} \to \mathbb{R}$ when $f(x) = e^x + 1$ has no fixed point. Neither of these extreme cases has useful applications. The case where there exists a unique fixed point is of utmost significance. The result that guarantees the existence and uniqueness of a fixed point is known as the contraction principle.

Let (X, d) be a metric space. A map $f : X \to X$ is called a *contraction* if there exists α, $0 \leq \alpha < 1$, such that for all $x, y \in X$ we have

$$d(f(x), f(y)) \leq \alpha d(x, y).$$

It is obvious that a contraction map is uniformly continuous (hence, continuous), but not conversely.

A. Contraction Principle

Each contraction map $f:(x,d)\to(X,d)$ of a complete metric space (X,d) has a unique fixed point.

Some important extensions are as follows: If C is a convex compact subset of a normed space (see Section VIII.B for the definitions), then a continuous map $f:C\to C$ has a fixed point (Banach–Schauder theorem). This is generalized by Tychonoff: Let C be a convex compact subset of a locally convex space (see Section VIII.B); then a continuous map $f:C\to C$ has a fixed point. An important contribution to fixed point theorems is due to Markov and Kakutani: Let C be a convex compact subset of a locally convex space. Let $\{f_\alpha:\ \alpha\in\Gamma\}$ be a family of continuous maps from C to C such that each f_α is *affine* [i.e., for all $x, y\in C$ and $0<t<1$, $f_\alpha(tx+(1-t)y)=tf_\alpha(x)+(1-t)f_\alpha(y)$] and the family is commuting [i.e., $f_\alpha\circ f_\beta(x)=f_\beta\circ f_\alpha(x)$ for all $x\in C$]. Then there is $x_0\in C$ such that $f_\alpha(x_0)=x_0$ for all $\alpha\in\Gamma$, i.e., the family $\{f_\alpha\}$ has a common fixed point.

VII. FUNCTION SPACES

So far we have been concerned with the properties of topological spaces and maps. In this section we want to topologize certain sets of functions. One might say that so far we have been working on the ground floor and now we want to move up to the first floor.

Let X, Y be two nonempty sets. Y^X denotes the set of *all* maps $f:X\to Y$. If X, Y are topological spaces, then $C(X,Y)$ denotes the set of all continuous maps of X into Y. Clearly, $C(X,Y)$ is a subset of Y^X. Both Y^X and $C(X,Y)$ with some appropriate topologies are called function spaces. Several topologies are very useful. Two of them are point-open topology and compact-open topology.

For each $x\in X$ and each open set V in Y, let $T(x,V)=\{f\in Y^X: f(x)\in V\}$. As x runs over X and V over all open subsets of Y, the collection $\{T(x,V)\}$ forms a subbase of a topology called the *point-open topology* and is denoted $\mathcal{T}_p$. Similarly, if K is a compact subset of X and V an open subset of Y, then

$$T(K,V)=\{f\in Y^X: f(x)\in V \text{ for all } x\in K\}.$$

The topology (denoted $\mathcal{T}_c$) having the collection $\{T(K,V)\}$ as a subbase is called the *compact-open topology*. Since each singleton is compact, it follows that $\mathcal{T}_c$ is finer than $\mathcal{T}_p$. Hence, the induced topology $\mathcal{T}_c$ on $C(X,Y)$ is finer than the induced topology $\mathcal{T}_p$. If Y is a Hausdorff space, both $(Y^X;\mathcal{T}_p/\mathcal{T}_c)$ and $(C(X,Y),\mathcal{T}_p/\mathcal{T}_c)$ are also Hausdorff spaces.

The following three cases are noteworthy. If (X,d) is a metric locally compact space and (Y,d') a complete metric space, then $(C(X,Y),\mathcal{T}_c)$ is a complete uniform space. If (X,d) is a compact metric space and (Y,d') a complete metric, then $(C(X,Y),\mathcal{T}_c)$ is a complete metric space with metric $d^+(f,g)=$ the least upper bound of $d'(f(x),g(x))$, $x\in X$ for $f, g\in C(X,Y)$. Thus, for a compact metric space (X,d), $C(X)=C(X,\mathbb{R})$ is a complete metric space. The same is true for $C[a,b]$.

Hereafter, both $C(X,\mathbb{R})$ and $C(X,\mathbb{C})$ are denoted $C(X)$, the space of all real- or complex-valued continuous functions on the topological space $(X,\mathcal{T})$. Indeed, if X is compact, then each $f\in C(X)$ is bounded and $\|f\|=$ the least upper bound of $\{|f(x)|, x\in X\}$ is finite. $\|f\|$ is called the *norm* of f and $d(f,g)=\|f-g\|$ gives a metric on $C(X)$ which is complete. For each fixed $x\in X$, $\varphi_x(f)=f(x)$, $f\in C(X)$, defines a map of $C(X)$ into $\mathbb{R}$ or $\mathbb{C}$. φ_x is called an *evaluation map*. The map φ_x is useful. We come back to it later.

In some particular cases, the topological space X can be embedded in $C(X)$. For instance, let (X,d) be a metric space and $C_b(X)$, the space of all continuous bounded functions on X. $C_b(X)$ is a metric space with the metric $d^+(f,g)=$ the least upper bound of $\{|f(x)-g(x)|: x\in X\}$, $f, g\in C_b(X)$. Now let $x_0\in X$ be a fixed element, and for each $y\in X$, set $f_y(x)=d(x,y)-d(x,x_0)$, $x\in X$. Then for each $y\in X$, $f_y\in C_b(X)$ and the map $y\to f_y$ of X into $C_b(X)$ is an isometry. Thus X can be isometrically embedded in $C_b(X)$. If X is compact, then $C_b(X)=C(X)$ and so each compact metric space can be isometrically embedded in $C(X)$.

If X, Y are topological spaces, then a continuous map $f:X\to Y$ induces a map $f^*:C(Y)\to C(X)$ by $f^*(g)=g\circ f$ (composition map). It is interesting to know what properties of f induce similar properties for f^*. An important scenario occurs when both X and Y are compact, viz., X, Y are homeomorphic if and only if $C(X), C(Y)$ are isometrically homeomorphic.

VIII. MARRIAGE OF TOPOLOGY AND ALGEBRA

Algebraic operations such as addition, subtraction, multiplication, and division play an important role in algebra. Topology, however, is not primarily concerned with them. When algebra and topology are married in the sense that a set with algebraic operations is endowed with a topology, new and interesting theories emerge. For instance, in the field of functional analysis (a major branch of mathematics), the topics of topological groups, topological vector

spaces, and topological algebras, among others, are the outcome of this marriage.

Among many topics in algebra, one specifically comes across the following algebraic structures: semigroups, groups, vector spaces, and algebras. By endowing them with topologies so that the underlying algebraic operations are continuous, one obtains interesting areas of mathematics.

A. Topological Semigroups and Groups

A set S with an associative binary operation [i.e., for all $x, y \in S, x \circ y \in S$ and $x \circ y \circ z = (x \circ y) \circ z = x \circ (y \circ z)$, in which "$\circ$" denotes the binary operation] is called a *semigroup*. A semigroup S with a topology $\mathcal{T}$ is called a *topological semigroup* (TS) if the map $(x, y) \to x \circ y$ of $S \times S$ into S is continuous.

A semigroup G is called a *group* if G has an *identity* $e (e \circ x = x \circ e = x$ for $x \in G)$ and each $x \in G$ has an *inverse* x^{-1} $(x^{-1} \circ x = x \circ x^{-1} = e)$. According to whether "$\circ$" is "$+$" or "$\times$," the group is called *additive* or *multiplicative*. If $x \circ y = y \circ x$, for all $x, y \in G$, the group G is called *Abelian* or *commutative*. A map f of a group G into a group H is called a *homomorphism* if $f(xy) = f(x) f(y)$. A bijective homomorphism of a group G into a group H is called an *isomorphism*.

A group G with a topology is called a *topological group* (TG) if the map $(x, y) \to xy^{-1}$ of $G \times G$ into G is continuous.

Each topological group is a uniform space, hence completely regular. If the topology satisfies the T_0-axiom, the topological group becomes a Tychonoff space, hence Hausdorff. A Hausdorff topological group is metrizable iff the neighborhood system at its identity has a countable base.

The class of all topological groups satisfies arbitrary productive properties. Moreover, the closure of each subgroup (i.e., a subset H of G such that H by its own right is a group under induced algebraic operations) is a topological group and each open subgroup is closed.

Let H be an invariant ($xH = Hx$ for all $x \in G$) subgroup of a group G. To each $x \in G$ we associate xH, called the *coset*. Let G/H denote the set of all cosets. With multiplication $(xH)(yH) = xyH$ and identity H, G/H becomes a group called the *quotient* or *factor* group. The map $\phi : x \to xH$ of G onto G/H is called the *quotient* or *canonical* map; ϕ is a homomorphism because $\phi(xy) = \phi(x)\phi(y)$. If G is a topological group, we can endow G/H with the quotient topology (see Section III), making ϕ continuous. It turns out that ϕ is a continuous, open, and surjective homomorphism of the topological group G into the topological group G/H. The quotient topology on G/H is Hausdorff iff H is closed.

1. Examples

$\mathbb{R}^n$ $(n \geq 1)$ with coordinatewise addition $(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n)$ is an additive Abelian group and has identity $(0, \ldots, 0)$. With the euclidean topology defined in Section 1.A, $\mathbb{R}^n$ is an Abelian additive topological group. In particular, the set $\mathbb{R}$ of all real numbers is an Abelian additive topological group in which $\mathbb{Z}$, the subgroup of all integers, is a closed invariant subgroup. Thus, the quotient group $\mathbb{R}/\mathbb{Z}$ is a topological group. We denote $\mathbb{R}/\mathbb{Z}$ by $\mathbb{T}$. $\mathbb{T}$ is actually isomorphic with the circle group $\{e^{it} : 0 \leq t \leq 2\pi, i = \sqrt{-1}\}$. Hence, $\mathbb{T}$ is a compact Abelian topological group.

If G is an Abelian topological group, the set G' of all continuous homomorphisms $\phi : G \to \mathbb{T}$ is called the *dual group* of G. We can endow G' with the compact-open topology $\mathcal{T}_c$ (see Section VII). If G is a Hausdorff locally compact Abelian topological group, then $(G', \mathcal{T}_c)$ is also a Hausdorff locally compact Abelian topological group. Repeating the process, we see that G'' (the dual group of G') with the compact-open topology is also a locally compact Abelian topological group. A celebrated result called the Pontrjagin duality theorem tells us that G and G'' are isomorphic and homeomorphic. Furthermore, if G is compact (discrete), G' is discrete (compact).

If algebraic operations in a topological group with some additional structure are differentiable instead of being only continuous, it leads to the theory of Lie groups, which form an important branch of mathematics.

B. Topological Vector Spaces, Banach Spaces, and Hilbert Spaces

An algebraic system that is richer than groups is vector space. All scalars involved in this section are either real or complex numbers.

An Abelian additive group E is called a *real* or *complex* (depending on which scalars are used throughout) *linear* or *vector space* if for all $x \in E$ and scalar λ, $\lambda x \in E$ satisfies the following: $\lambda(x + y) = \lambda x + \lambda y$; $(\lambda + \mu)x = \lambda x + \mu x$; $\lambda(\mu x) = (\lambda\mu)x$ for all scalars λ, μ and all $x, y \in E$; $1x = x$ and $0x = 0$, the identity of the additive group E. If $F \subset E$ and F is a vector space in its own right over the same scalars as those of E, then F is called a *linear subspace* of E. A subset C of E is called *circled* if for scalars λ, $|\lambda| \leq 1$ and $x \in C$, $\lambda x \in C$. A subset C of a vector space E is called *convex* if for all $x, y \in C$ and $0 \leq \lambda \leq 1$, $\lambda x + (1 - \lambda)y \in C$. C is *absorbing* if for all $x \in E$ there is $\alpha_0 > 0$ such that for λ, $|\lambda| \geq \alpha_0$, $x \in \lambda C$.

A linear or vector space E endowed with a topology $\mathcal{T}$ is called a *topological linear* or *vector space* (TVS) if the maps $(x, y) \to x + y$ and $(\lambda, x) \to \lambda x$ of $E \times E$ into E and of $K \times E$ into E, respectively, are continuous, in

which the field K of real numbers or complex numbers is endowed with its natural topology. If there exists a base of convex neighborhoods of identity 0 in E, then E is called a *locally convex* (LC) space.

Clearly, each topological vector space is an Abelian additive topological group, hence it is a uniform space and, therefore, completely regular. Hence, a Hausdorff topological vector space is a Tychonoff space. As for topological groups, a Hausdorff topological vector space is metrizable iff the neighborhood system at 0 has a countable base. Among the metrizable topological vector spaces, there is a very distinguished and useful subclass of spaces called the normable spaces defined below.

Let E be a real (or complex) vector space. A map $p : E \to K$ is called a *functional*. If $p(x) \geq 0$ for all $x \in E$ such that $p(\lambda x) = |\lambda| p(x)$ and $p(x + y) \leq p(x) + p(y)$, then p is called a *seminorm*. If $p(x) = 0 \Leftrightarrow x = 0$, then p is called a *norm*, denoted $p(x) = \|x\|$. A seminorm p on a vector space E defines a pseudometric $d(x, y) = p(x - y)$. If p is a norm, d becomes a metric. If the topology of a topological vector space is induced by a *norm*, it is called a *normable* topological vector space. If a normable topological vector space is complete in its induced metric topology, it is called a *Banach space* (BAS). Each Banach space is a complete metric locally convex space called a *Fréchet space* (F). A Hausdorff locally convex space is called a *barreled space* (BS) if, in it, each *barrel* (convex, circled, absorbing, and closed set) is a neighborhood of 0. Each Baire locally convex (BLC) space is a barreled space, and each Fréchet space is a Baire space.

A map f of a vector space E into another vector space F is called *linear* if $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$ for all $x, y \in E$ and scalars α, β. A functional f on a normed space E is called *bounded* if there is a real number $M > 0$ such that $|f(x)| \leq M\|x\|$ for all $x \in E$. An interesting but simple fact is that a linear functional on a normed space is bounded iff it is continuous.

To exhibit the richness that results from the marriage of topology and vector spaces, we cite a few prototypical results.

The first is the so-called Hahn–Banach extension theorem: Let F be a linear subspace of a vector space E. Let p be a seminorm on E, and f a linear functional on F such that $|f(x)| \leq p(x)$ for all $x \in F$. Then there exists a linear functional $\tilde{f}$ on E such that $\tilde{f}(x) = f(x)$ for all $x \in F$ and $|\tilde{f}(x)| \leq p(x)$ for all $x \in E$. It is worth comparing this result with Tietze's extension theorem given in Section IV.

The heart of functional analysis, especially that of topological vector spaces, lies in the so-called *twins* of functional analysis, popularly known as the open-mapping and closed-graph theorems.

1. Open-Mapping Theorem

Each continuous linear surjective map of a Fréchet (in particular, Banach) space onto a barreled (in particular, Fréchet or Banach) space is open.

2. Closed-Graph Theorem

Each linear map of a barreled (in particular, Fréchet or Banach) space into a Fréchet (or Banach) space with closed graph is continuous.

Another important result in topological vector spaces is as follows. If $\{f_n\}$ is a sequence of continuous linear functionals on a barreled (or Fréchet or Banach) space E such that $f(x) = \lim_n f_n(x)$, $x \in E$, then f is also linear and continuous. (This is called the Banach–Steinhaus theorem.)

If E is a real or complex topological vector space, the set E' of all continuous linear functionals on E is called the *dual* of E. If E is a normed space, so is E' with the norm $\|f\| =$ the least upper bound of $\{|f(x)| : \|x\| \leq 1\}$. Actually E' is a Banach space regardless of E being a complete or incomplete normed space. In special cases E' can be determined by the points of E, as shown below.

A normed space E is called a *pre-Hilbert* or *an inner product space* if the norm $\|\cdot\|$ satisfies the so-called *parallelogram law*: $\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2)$. The parallelogram law implies the existence of a bilinear functional $\langle , \rangle : E \times E \to K$ satisfying the following properties: $\langle x, x \rangle \geq 0$; $\langle x, x \rangle = 0 \Leftrightarrow x = 0$; $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$; and $\langle x, y \rangle = \langle y, x \rangle$ (or $\overline{\langle y, x \rangle}$ for the complex scalars). We see that $\|x\| = +\sqrt{\langle x, x \rangle}$ gives a norm. If a pre-Hilbert space is complete, it is called a *Hilbert space* (H). For a Hilbert space H, its dual H' can be identified with the points of H as demonstrated by the so-called Riesz representation theorem: If H is a Hilbert space, then for each $f \in H'$ there exists a unique element $y_f \in H$ such that $f(x) = \langle x, y_f \rangle$, $x \in H$ with $\|y_f\| = \|f\|$. (See Table II for the interrelation of various topological spaces.)

Hereafter, we assume all topological vector spaces to be Hausdorff. For any vector space E over $\mathbb{R}$ or $\mathbb{C}$, E^* denotes the set of all real or complex linear functionals on E, called the *algebraic dual* of E. If E is a TVS, E' denotes the set of all continuous linear functionals on E, called the *topological dual* or simply the *dual* of E as mentioned above. If E is LC and $E \neq \{0\}$, then $E' \neq \{0\}$ and clearly $E' \subset E^* \subset \mathbb{R}^E$ (or $\mathbb{C}^E$), where $\mathbb{R}^E$ carries the pointwise convergence topology. The topology induced from $\mathbb{R}^E$ to E' is called the weak-star or *w^*-topology*. Under this topology all maps $f \to f(x)$ (for each fixed x, i.e., evaluation maps) are continuous.

Similarly, the coarsest topology on E which makes all the maps $x \to f(x)$ (for any fixed $f \in E'$) continuous

is called the *weak topology* $\sigma(E, E')$ on E, which is coarser than the initial topology of E. The space E with the weak topology $\sigma(E, E')$ becomes an *LC* space. The finest locally convex topology on E that gives the same dual E' of E is called the *Mackey topology* $\tau(E, E')$, which is finer than the initial topology $\mathcal{T}$ of E. Thus we have $\sigma(E, E') \subset \mathcal{T} \subset \tau(E, E')$. An *LC* space with the Mackey topology [i.e., $\mathcal{T} = \tau(E, E')$] is called a *Mackey space.* Every barreled (hence Fréchet, Banach, or Hilbert) space is a Mackey space. However, there exist Mackey spaces which are not Fréchet spaces. The weak, weak*, and Mackey topologies are extensively used in functional analysis. Here we list only a few samples of their usage.

Let E be a Banach space; then for each weakly compact subset A of E, the convex closure $\bar{c}_o(A)$ of A is also weakly compact, where $\bar{c}_o(A)$ is the intersection of all closed convex subsets of E containing A. Further, a weakly closed subset A of E is weakly sequentially compact if and only if A is weakly compact. This is known as the Eberlein theorem. Note that in general topological spaces sequential compactness is not equivalent to compactness. Furthermore, if A is a closed convex subset of E such that for each $f \in E'$ there is $x_f \in A$ with $|f(x_f)| =$ the least upper bound of $\{|f(x)| : x \in A\}$, then A is weakly compact. This is known as the James theorem.

As pointed out earlier, the dual E' of a normed space E is a Banach space with the norm $\|f\|$. The w^*-topology on E' is coarser than this norm topology. The unit ball $\{f \in E' : \|f\| \leq 1\}$ of E' is w^*-compact (Alaoglu's theorem) but not norm compact in general. *Actually*, the unit ball of a normed space E is norm compact iff E is finite-dimensional, i.e., E is homeomorphic to $\mathbb{R}^n$ (or $\mathbb{C}^n$) for some finite positive integer n.

Since the dual E' of a normed space E is a normed (actually Banach) space, we can consider the dual E'' of E'. E'' is called the *bidual* of E. Clearly E'' is also a Banach space. There is a natural embedding of E in E''. Put $x''(f) = f(x)$, $f \in E'$, for each $x \in E$. Then it can be verified that $x'' \in E''$ for each $x \in E$ and so $x \to x''$ gives a mapping of E into E'' such that $\|x''\| = \|x\|$. Thus E is embedded into E'' isometrically. In general this embedding is not surjective, i.e., $E \subset E''$, $E \neq E''$. Whenever $E = E''$, E is called *reflexive.* Since E'' is a Banach space, it follows that a reflexive normed space must be a Banach space. By virtue of the Riesz representation theorem, each Hilbert space is reflexive. Here are some examples of reflexive and nonreflexive spaces. Let $1 \leq p < \infty$ be a given real number. Let $\ell_p = \{\{a_i\} : \sum_{i=1}^{\infty} |a_i|^p < \infty\}$. Then ℓ_p is a Banach space with the norm $\|\{a_i\}\|_p = \{\sum_{i=1}^{\infty} |a_i|^p\}^{\frac{1}{p}}$. For all $p's$, $1 < p < \infty$, ℓ_p is a reflexive Banach space; note that ℓ_2, being a Hilbert space, is reflexive, but ℓ_1 is not reflexive. Note that for $1 < p < \infty$, $\ell_p' = \ell_q$, where $(1/p) + (1/q) = 1$ and $\ell_1' = \ell_\infty$, the space of all bounded sequences.

For a Tychonoff space $(X, \mathcal{T})$, $C(X) = (C(X), \mathcal{T}_c)$ denotes the set of all real or complex continuous functions on X, endowed with the compact-open topology $\mathcal{T}_c$ (see Section VII). One sees that $C(X)$ is an *LC* space. It is not, in general, a metrizable, or a complete, or a barreled or a Fréchet or a Banach space. As shown above, if X is compact, then $C(X)$ is a Banach space. Conversely, if $C(X)$ is a Banach space, then indeed X is compact. This immediately suggests an interplay of topological properties of X and $C(X)$.

To display this duet between X and $C(X)$ briefly, we need the following concept. Let $\mathcal{L} \in C'(X)$, the dual of $C(X)$. The smallest compact subset A of X such that for all $f \in C(X)$ with $f(A) = 0$ implies $\mathcal{L}(f) = 0$ is called the *support* of $\mathcal{L}$, denoted $\text{supp}\mathcal{L}$. If B' is a subset of $C'(X)$, then $\text{supp}B' = \text{Cl}\{\cup(\text{supp}\mathcal{L} : \mathcal{L} \in B')\}$ is the support of B'. Now we have the following:

- $C(X)$ is a Banach space iff X is compact.
- $C(X)$ is metrizable iff X is hemicompact.
- $C(X)$ is complete iff X is a k_r-space.
- $C(X)$ is a Fréchet space iff X is a hemicompact, k_r-space.
- $C(X)$ is a Mackey space iff $\text{supp}B'$ is compact for each convex circled w^*-compact subset B' of $C'(X)$.
- $C(X)$ is a barreled space iff X is a μ-space [i.e., for each w^*-bounded subset B' of $C'(X)$, $\text{supp}B'$ is compact].

C. Topological Algebras

An algebraic structure richer than vector spaces is what is known as algebra. A real or complex vector space A is called an *algebra* if there is some multiplication defined on A, that is, for all $x, y \in A$, $xy \in A$ satisfying the following axioms: $x(y + z) = xy + xz$; $(x + y)z = xz + yz$; $\lambda(xy) = (\lambda x)y = x(\lambda y)$; $x(yz) = (xy)z = xyz$ for all $x, y, z \in A$ and scalar λ. An algebra A is called *commutative* if $xy = yx$ for all $x, y \in A$. A has an identity e if $ex = xe = x$ for all $x \in A$. An element $x \in A$ is called *invertible* if there is $y \in A$ such that $xy = yx = e$. y is called the *inverse* of x and written $y = x^{-1}$. An algebra in which each nonzero element has an inverse is called a *division* algebra. An operation $^* : x \to x^*$ of A onto A is called an *involution* if $(x + y)^* = x^* + y^*$, $(\lambda x)^* = \bar{\lambda} x^*$, $(xy)^* = y^* x^*$, and $x^{**} = x$.

An algebra A with a topology $\mathcal{T}$ is called a *topological algebra* (TA) if the maps $(x, y) \to x + y$, $(\lambda, x) \to \lambda x$, and $(x, y) \to xy$ are continuous. It is clear that each topological algebra is a topological vector space. Hence, the results pertaining to topological vector spaces can be

used for topological algebras. If the topology is given by a norm on an algebra, it is called a *normed algebra*, provided that $\|xy\| \leq \|x\|\|y\|$. A complete normed algebra is called a *Banach algebra* (BAA). A Banach algebra with an involution * is called a B*-algebra if $\|xx^*\| = \|x\|^2$ for all x. We deal with these algebras in the next section.

If the topology of a topological algebra is locally convex, it is called a *locally convex algebra*. A complete metric locally convex algebra is called a *B_0-algebra* (B_0-A). Clearly, each Banach algebra is a B_0-algebra.

Examples of nonnormed algebras also abound in the literature. For example, for any Tychonoff space $(X, \mathcal{T})$, the set $C(X)$ of all continuous real or complex functions forms an algebra with pointwise operations, i.e.,

$$(f+g)(x) = f(x) + g(x), \qquad (\lambda f)(x) = \lambda f(x),$$
$$(fg)(x) = f(x)g(x),$$

where λ is a real or complex scalar. With the compact-open topology T_c, $C(X)$ is actually a locally convex algebra. Among many applications associated with $C(X)$, there is a celebrated result called the Stone–Weierstrass theorem: Let A be a subalgebra of $C(X)$ such that

(i) for all $x, y \in X, x \neq y$, there is $f \in A$ with $f(x) \neq f(y)$, (i.e., A separates points of X),
(ii) for each $x \in X$, there is $f \in A$ with $f(x) \neq 0$, and
(iii) for each $f \in A$, the complex conjugate $\bar{f}$ [i.e., $\bar{f}(x) = \overline{f(x)}$] is in A.

Then A is dense in $C(X)$ [i.e., $\bar{A} = C(X)$]. If, in addition, A is a closed subalgebra of $C(X)$, then $A = C(X)$. Note that condition (iii) is redundant for real algebras. From this theorem, one derives the classical Weierstrass theorem: Each continuous real function on $[a, b]$, $-\infty < a < b < \infty$, can be approximated by polynomials.

IX. BANACH ALGEBRAS

Banach algebras form an important subclass of topological algebras because they have a richer structure owing to the norm. Hence, they have more useful applications. The theory of Banach algebras can be used to advantage in mathematical analysis, Fourier series, representation theory, and other significant areas of mathematics.

An important result revealing a basic fact about Banach division algebra is the so-called Gelfand–Mazur theorem: A complex Banach division algebra is isomorphically homeomorphic with the algebra of complex numbers.

A functional f on an algebra A is called *multiplicative* if $f(xy) = f(x)f(y)$. A simple but beautiful result regarding multiplicative functionals is as follows. Each multiplicative linear functional on a Banach algebra is continuous. It is beautiful because an algebraic hypothesis (namely, linearity and multiplicativity) gives a topological conclusion (namely, the continuity). It is not known if this beautiful result is true for Fréchet algebras [i.e., those B_0-algebras whose topology is defined by a sequence of seminorms $\{p_n\}$ such that $p_n(xy) \leq p_n(x)p_n(y)$ for all $x, y \in A$].

If $\Delta(A)$ denotes the set of all nonzero multiplicative linear functionals on a commutative Banach algebra A, the above result says that $\Delta(A) \subset A'$, the dual of A. The set $\Delta(A)$ endowed with the $\mathcal{T}_p$ topology is called the *maximal ideal space* of A. $\Delta(A)$ is locally compact, but if the Banach algebra has identity, $\Delta(A)$ is a compact Hausdorff space.

Banach algebras are often grouped into three classes: function algebras, group algebras, and operator algebras. Function algebras are the targets of study in topology, whereas group algebras are studied in harmonic analysis and operator algebras in operator theory. The latter two constitute large fields of mathematics.

Here we consider only an instance of function algebras. Let $(X, \mathcal{T})$ be a compact Hausdorff space and $C(X)$ the set of all continuous complex-valued functions on X. With algebraic operations $(f+g)(x) = f(x) + g(x)$, $(\lambda f)(x) = \lambda f(x)$, $(fg)(x) = f(x)g(x)$, $C(X)$ becomes a complex algebra. If we put $\|f\| =$ the least upper bound of $\{|f(x)|, x \in X\}$, we have a norm on $C(X)$ with $\|fg\| \leq \|f\|\|g\|$. Using the last result in Section VII, $C(X)$ is complete in the norm topology. In other words, $C(X)$ is a Banach algebra. It is commutative and has identity $1(1(x) = 1)$. Moreover, there is an involution $f^*(x) = \overline{f(x)}$, complex conjugate. One can verify that $\|ff^*\| = \|f\|^2$. Thus, $C(X)$ is a B*-algebra. It was a crowning achievement of I. Gelfand and M. Naimark (Russian mathematicians) to show that each commutative B^*-algebra with identity is isometric (i.e., norm goes into norm) to $C(X)$ for some compact Hausdorff space X. Actually, it turns out that $X = \Delta(A)$, the maximal ideal space of A.

X. ALGEBRAIC TOPOLOGY

Certain properties of topological spaces, such as "invariance" and "congruence," can, under suitable conditions, be expressed in terms of groups—specifically homotopy and homology groups—associated with spaces. The study of this phenomenon constitutes part of an important field of mathematics called algebraic topology. Among many fascinating topics in this area, there are two noteworthy theories: homotopy and homology.

A. Homotopy

Let X, Y be two topological spaces and $I = [0, 1]$. Two maps $f, g: X \to Y$ are said to be *homotopic* (written $f \sim g$) if there exists a continuous map $\phi: X \times I \to Y$ such that $\phi(x, 0) = f(x)$ and $\phi(x, 1) = g(x)$ for $x \in X$. If g is a constant map [i.e., $g(x) = y_0 \in Y$ for all $x \in X$) and $f \sim g$, then f is called *null homotopic* and written $f \sim 0$.

In some cases each map is null homotopic. For instance, if $f : X \to \mathbb{R}^n$ or $g : \mathbb{R}^n \to Y$, then $f \sim 0$ and $g \sim 0$. However, it fails if we replace $\mathbb{R}^n$ by $S^{n-1} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1^2 + \cdots + x_n^2 = 1\}$, the n-dimensional sphere.

The relation $f \sim g$ is an equivalence relation ($f \sim f$; $f \sim g \Rightarrow g \sim f$; $f \sim g$ and $g \sim h \Rightarrow f - h$). Thus, the homotopic relation decomposes $C(X, Y)$ into disjoint equivalence classes $\dot{f}$ ($g \in \dot{f} \Leftrightarrow f \sim g$) called the *homotopic classes*, which are denoted $[X, Y]$. If $\Psi : X \to Y$ is continuous, then for any topological space Z, there exists a map $\Psi^* : [Y, Z] \to [X, Z]$ defined by $\Psi^*(f) = f \circ \Psi$, $f \in [Y, Z]$. Thus, the topological properties of $\Psi : X \to Y$ can be studied by means of the properties of Ψ^*, and vice versa.

Homotopy is actually the study of the extension of maps, to which the reader has been introduced in previous sections. For $f \sim g$ iff the map $\phi : (X \times \{0\}) \cup (X \times \{1\}) \to Y$ defined by $\phi(x, 0) = f(x), \phi(x, 1) = g(x), x \in X$ can be extended to a continuous map $\tilde{\phi}$: $X \times I \to Y$. Such an extension in a particular case has a special name.

Two topological spaces X, Y are said to be *homotopic* if there exist continuous maps $f : X \to Y$ and $g : Y \to X$ such that the composition maps $f \circ g$ and $g \circ f$ are homotopic to the identity maps of Y and X, respectively. The symbol "$X \approx Y$" means that X is homotopic to Y.

It is easy to see that the relation of being homotopic is an equivalence relation on the class of all topological spaces. It follows easily that if X is homeomorphic with Y, then $X \approx Y$. But the converse need not be true. For example, $\mathbb{R}^n$ is homotopic to $\{0\}$ but is not homeomorphic.

The equivalence relation of being "homotopic" decomposes the class of all topological spaces into disjoint equivalence classes of homotopic spaces.

A property of a topological space is called a *topological* (respectively, *homotopic*) *invariant* if it remains *unaltered* under homeomorphisms (respectively, homotopic maps). Most topological invariants are not *homotopic invariants*. But some of them are.

A topological space (X, T) is called *pathwise connected* if for each pair $a, b \in X$, $a \neq b$, there is a path $f: I \to X$ such that $f(0) = a$, $f(1) = b$, $\mathbb{R}^n$, S^n (for $n \geq 1$) are pathwise connected. Pathwise connectedness is a homotopic invariant.

1. Retracts

A subset A of a topological space $(X, \mathcal{T})$ is said to be a retract of X if the identity map $i: A \to A(i(x) = x)$ can be extended to a continuous map $f: X \to A$ so that $f(x) = x$ for all $x \in A$. In this case f is called a *retract*.

Each retract of a Hausdorff topological space is closed. Each closed unit disk $D^n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1^2 + \cdots + x_n^2 \leq 1\}$ is a retract of $\mathbb{R}^n$ and each sphere S^{n-1} is a retract of $\mathbb{R}^n \backslash \{0\}$.

A topological space Y is called an AR for *absolute retract* [or *AR* (*normal*)] if for any topological (respectively, normal) space X and a closed subjset $A \subset X$, each continuous map $A \to Y$ has a continuous extension to a map $X \to Y$. As shown in Tietze's extension theorem (Section IV), $\mathbb{R}$ is an AR (normal).

A subset E of a topological space $(X, \mathcal{T})$ is said to be *deformable* into a subset $F \subset X$ if the identity map $i: E \to E$ is homotopic in X to a map of E into F. Deformability is *topologically invariant*; that is, if $\phi: X \to Y$ is a homeomorphism, then $\phi(E)$ is deformable to $\phi(F)$ whenever E is deformable to F. For example, $\mathbb{R}^n \backslash \{0\}$ is deformable to S^{n-1}.

2. Degree

Another property of a map that is invariant under homotopy is the degree of a self-map of S^n. If $f: S^n \to S^n$ is a continuous map, there exists an integer $D(f)$ (positive, negative, or zero) called the *degree* of f. In the simple case of $n = 1$, the degree of $f: S^1 \to S^1$ (note that S^1 is the circle) represents the number of time and sense that the image point $f(x)$ rotates around S^1 when x rotates in one oriented direction of S^1. For example, the degree of $f(z) = z^n$ when f maps S^1 into S^1 is n. It follows that if $f, g: S^n \to S^n$, then $f \sim g \Leftrightarrow D(f) = D(g)$. From this or otherwise. Brouwer derived the fact that the identity map of S^n is not null homotopic. However, if $m < n$ and $f: S^m \to S^n$, then $f \sim 0$. If $m > n$, the situation is in general uncertain. However, if $n > 1$, then $f : S^n \to S^1$ is null homotopic. If $f: S^n \to S^n$ is *antipodal* [$f(-x) = -f(x)$], $n \geq 0$, then $D(f)$ is an odd number and hence f cannot be null homotopic.

3. Jordan Curves

To study further properties of $\mathbb{R}^n$, we say that a subset A of $\mathbb{R}^n$ *separates* $\mathbb{R}^n$ if $\mathbb{R}^n \backslash A$ is not connected (see Section I). For example, the circle S^1 separates the plane $\mathbb{R}^2$. Actually the *Jordan separation theorem* states that every homeomorphic image of S^{n-1} into $\mathbb{R}^n$ separates $\mathbb{R}^n$. Furthermore, it can be derived that if $m \neq n$, then $\mathbb{R}^n$ is not

homeomorphic with $\mathbb{R}^m$. (This is called the *invariance of dimension theorem.*) In the particular case of $\mathbb{R}^2$, every homeomorphic image of S^1 separates $\mathbb{R}^2$ in exactly two components. (This is the *Jordan curve theorem.*)

4. Paths and Loops

A continuous map p of $I = [0, 1]$ into a topological space $(X, \mathcal{T})$ is called a *path* in X. $p(0) = a$ is called the *starting point* of p, and $p(1) = b$ the end point of p. If $p(0) = a = p(1) = b$, then p is called a *loop*. The set of all loops having their starting and end points at a is denoted $\mathcal{P}(X, a)$ and is called the *loop space*. Clearly, $\mathcal{P}(X, a)$ is a subset of $C(I, X)$ and so it can be given the compact-open topology $\mathcal{T}_c$, (see Section VIII). The study of loop spaces can be subsumed under a general structure called the H-structure.

5. H-Structures

A topological space X with a continuous map $\eta : X \times X \to X$ is said to have an H-structure if there exists a fixed point $a \in X$ making the maps $x \to \eta(x, a)$ and $x \to \eta(a, x)$ homotopic to the identity map of X. A topological space that carries an H-structure is called an *H-space*. For example, each topological group (see Section VIII.A) carries an H-structure, and so does a loop space $\mathcal{P}(X, a)$.

In general the point a in the H-structure is not unique. Let $P = \{a$: the maps $x \to \eta(x, a)$ and $x \to \eta(a, x)$ are homotopic to the identity map of $X\}$. P is called a *path component* of X and is said to be the *principal component* of the H-space. If Comp X denotes the discrete space of all path components, it forms a group, called the *fundamental group*, of X at a. In general, the fundamental group is not Abelian. However, if X is *path-connected* [i.e., for all pairs $a, b, a \neq b$, in X, there is a path $p: I \to X$ with $p(0) = a$, $p(1) = b$], then the fundamental group is Abelian.

In modern algebraic topology, the study of H-structures can be subsumed under *fiber structure*.

6. Fiber Spaces

A triple (X, p, B), consisting of two topological spaces, X and B, and a continuous surjective map $p: X \to B$, is called a *fiber structure*. X is called a *total* (or *fibered*) space, B a *base* space, and p a projection. We say that (X, p, B) is a fiber structure over B, and for each $b \in B$, the set $p^{-1}(b)$ is called the *fiber* over b.

Let Y be a topological space and $f: Y \to B$ a continuous function. If there exists a continuous function $\tilde{f}: X \to B$ such that $p \circ \tilde{f} = f$, then $\tilde{f}$ is called the *lifting* or *covering* of f. In particular, a lifting of the identity $i : B \to B$ is called a *cross section*.

Let $\mathcal{A}$ be a class of topological spaces. A fiber structure (X, p, b) is called a *fiber space* (or *fibration*) for the class $\mathcal{A}$ if for each Y in $\mathcal{A}$, each continuous map $f: Y \times \{0\} \to X$, and each homotopy $\varphi: Y \times I \to B$ of $p \circ f$, there exists a homotopy $\tilde{\varphi}$ of f covering φ. A fiber space for the class of all topological spaces is called a Hurewicz *fibration*.

It can be shown that if both X and B are metric spaces and if (X, p, B) is a fibration for the class of all metric spaces, then (X, p, B) is a Hurewicz fibration.

For further details and application of these ideas, the interested reader can consult any appropriate book on algebraic topology; some relevant texts are listed in the Bibliography.

B. Homology

Homology can be studied in a more general way than we do here. We are concerned with the homology of polyhedra in $\mathbb{R}^n$. Usually such a study is part of the so-called combinatorial topology, which is a part of algebraic topology.

Let $A = \{a_0, a_1, \ldots, a_m\}$ be a set of $m + 1$ $(m \leq n)$ points in $\mathbb{R}^n$ such that vectors $\{a_i - a_0 : 1 \leq i \leq m\}$ are linearly independent. ($\{x_0, \ldots, x_m\}$ are linearly independent if $\sum_{i=0}^m \lambda_i x_i = 0 \Rightarrow \lambda_i = 0$, for $i = 0, \ldots, m$.) The convex hull $\Delta_m = \{\sum_{i=0}^m \lambda_i a_i : \lambda_i \geq 0, \sum_{i=0}^m \lambda_i = 1\}$ of A is called an *m-dimensional simplex*. The points in A are called the *vertices* of Δ_m. If $m = 1$, then Δ_1 is a *straight line segment*. If $m = 2$, Δ_2 is a *triangle*, Δ_3 a *tetrahedron*, and so on.

If $r < m$, the simplex Δ_r with vertices $\{a_0, \ldots, a_{r-1}, a_{r+1}, \ldots, a_m\}$ is called the *r-face* of Δ_m.

A finite collection K of simplexes in $\mathbb{R}^n$ is called a *geometric complex* or just a *complex* if every face of each simplex is also in K and every two simplexes, if they intersect, intersect in a common face. If a complex K contains at least one m-dimensional simplex but not an $(m + 1)$-dimensional one, then K is said to be an *m-complex*. The point set of a complex is called a *polyhedron*.

If one gives an ordering to the vertices of a simplex, the simplex is said to have an *orientation* and is called an *oriented simplex*.

Let $\Delta_m^1, \Delta_m^2, \ldots, \Delta_m^{k(m)}$ denote a set of arbitrarily oriented m-simplexes in a complex K. Then $C_m = \sum_{i=1}^{k(m)} g_i \Delta_m^i$ (where g_i are integers) is called an *m-chain*. The set of all m-chains forms an Abelian additive group. If Δ_{m+1} is an $(m + 1)$-dimensional simplex, $\partial \Delta_{m+1}$ denotes the sum of all m-dimensional faces. The ∂ is called the boundary of Δ_{m+1}. One can verify that $\partial^2 = 0$. An m-chain C_m is called an *m-cycle* if $\partial C_m = 0$. Let Z_m denote the set of all m-cycles. Z_m is a subgroup of the group of m-chains. An m-cycle is called *homologous* to zero if it is the boundary of an $(m + 1)$-chain. The set of all

m-cycles that are homologous to zero forms a subgroup of Z_m. The quotient or factor group $Z_m/H_m = B_m$ is called the *m-dimensional homology group or Betti group* of the complex K.

Betti groups are topologically invariant that is, if the two polyhedra K and K' are homeomorphic, their respective Betti groups are isomorphic.

Using group theory from modern algebra, it can be claimed that each Betti group has a finite number of generators. These generators are of two kinds: One kind gives an infinite cyclic group, and the other kind gives groups of finite order. The number of generators of the first kind is called the *rank* of the Betti group, whereas the order of each generator of the second kind is called the *torsion coefficient* of the Betti group. These numbers are related by the so-called *Euler–Poincaré formula*:

Let K be an n-complex. Let $k(m)$ denote the number of m-simplexes of K, and $p(m)$ the m-dimensional Betti number for $m = 0, \ldots, n$. Then $\sum_{m=0}^{n}(-1)^m k(m) = \sum_{m=0}^{n}(-1)^m p(m)$. The number $\chi(K) = \sum_{m=0}^{n}(-1)^m p(m)$ is called the *Euler characteristic* of the complex K.

1. Example

For Δ_2, the triangle joining the vertices $\{a_0, a_1, a_2\}$ in $\mathbb{R}^2$, $k(0) = 3$, $k(1) = 3$, $k(2) = 1$, and so $\chi(\Delta_2) = 3-3+1 = 1$.

There are a number of fascinating topics in algebraic topology that have not been touched here, for example, dimension theory, fiber bundles, and manifolds. The interested reader can consult the Bibliography.

SEE ALSO THE FOLLOWING ARTICLES

ALGEBRA, ABSTRACT • COMPLEX ANALYSIS • CONVEX SETS • DISTRIBUTED PARAMETER SYSTEMS • KNOTS • MANIFOLD GEOMETRY • SET THEORY

BIBLIOGRAPHY

Armstrong, M. A. (1983). "Basic Topology," Springer-Verlag, New York.

Gamelin, T. W., and Greene, R. E. (1983). "Introduction to Topology," Saunders, Philadelphia.

Hochschild, G. P. (1981). "Basic Theory of Algebraic Groups and Lie Algebras," Springer-Verlag, New York.

Husain, T. (1977). "Topology and Maps," Plenum, New York.

Husain, T. (1981). "Introduction to Topological Groups" (reprint), Kreiger, Huntington, NY.

Husain, T. (1983). "Multiplicative Functionals on Topological Algebras," Pitman, Boston.

Husemoeller, D. (1982). "Fibre Bundles," Springer-Verlag, New York.

Rourke, C. P., and Sanderson, B. J. (1982). "Introduction to Piecewise Linear Topology," Springer-Verlag, New York.



Variational Calculus

John Troutman
Syracuse University

I. Historical Perspective
II. Analytic Foundations
III. Constraints
IV. Optimal Control Problems
V. Minimization Theory
VI. Hamiltonian Contributions
VII. Multidimensional Integral Problems
VIII. Variational Calculus in the Large
IX. Conclusion

GLOSSARY

Absolutely continuous Designates a continuous function on an interval which can be represented as the indefinite Lebesgue integral of its a.e. defined derivative.

a.e. Abbreviates "almost everywhere," meaning except for a set of Lebesgue measure zero.

Banach space A real linear space $\mathcal{Y}$ with a norm $\| \ \|$ that is complete in that if $y_n \in \mathcal{Y}$, $n = 1, 2, \ldots,$ then $\sum_{n=1}^{\infty} \|y_n\| < +\infty$ implies existence in $\mathcal{Y}$ of $\lim_{N\to\infty} \sum_{n=1}^{N} y_n$. Its *dual* $\mathcal{Y}^*$ is the linear space of all real valued norm continuous linear functions on $\mathcal{Y}$.

$\hat{\boldsymbol{C}}[\boldsymbol{a}, \boldsymbol{b}]$ Denotes the set of functions $\mathbf{y}$ on the compact interval $[a, b]$ that are $\hat{C}$ or *piecewise continuous* in that they have one-sided limits everywhere which agree except at a finite number of points.

$\hat{\boldsymbol{C}}^{\mathbf{1}}[\boldsymbol{a}, \boldsymbol{b}]$ Denotes the set of functions $\mathbf{y}$ on the compact interval $[a, b]$ that are $\hat{C}^1$ or *piecewise* C^1 in that they are continuous and have first derivatives $\mathbf{y}'$ from the left and from the right everywhere which agree except at a finite set of points called corner points.

$\boldsymbol{C^k}$ Abbreviates "k times continuously differentiable" for $k = 1, 2, \ldots$.

$\boldsymbol{C^k(S)}$ Denotes the class of functions that are C^k on a set S when S is either an open set or the closure of an open set in $\mathbb{R}^d$. (In the latter case, functions and their designated derivatives are assumed to have continuous extensions to ∂S, the boundary of S.)

div u Denotes the divergence of a vector valued function $\mathbf{u}$ with differentiable components $u_1, u_2, \ldots, u_d$ defined on a set in $\mathbb{R}^d$.

$\boldsymbol{f(x, y, y')}$ Indicates a real valued function of $2d + 1$ real variables. Insofar as they exist, the partial derivatives of f are denoted by subscripts f_x, $f_{\mathbf{y}}$, $f_{\mathbf{y}'}$, where the latter pair are vector valued.

inf Abbreviates "infimum," a synonym for "greatest lower bound."

Hilbert space A Banach space $\mathcal{Y}$ whose norm at y is given by $\|y\| = |(y, y)|^{1/2}$, where $(\cdot, \cdot)$ is a fixed inner product on $\mathcal{Y}$.

Homology The branch of algebraic topology which treats invariants that arise when objects of one class bound those of another in the sense that circles bound disks or other surfaces.

$L_p(S)$ The set of functions u on a set $S \subseteq \mathbb{R}^d$ for which $|u|^p$ is Lebesgue integrable over S $(p > 0)$.

Norm A real-valued function $\|\cdot\|$ on a linear space $\mathcal{Y}$ such that (i) $\|y\| \geq 0$ with equality *iff* $y = 0$; (ii) $\|cy\| = |c|\,\|y\|$ for all scalars c; (iii) $\|y + v\| \leq \|y\| + \|v\|$, for all $y, v \in \mathcal{Y}$. It topologizes $\mathcal{Y}$ through the distance assigning metric $\rho(y, v) = \|y - v\|$, $y, v \in \mathcal{Y}$.

sup Abbreviates "supremum," a synonym for "least upper bound."

∇u Denotes the gradient of a differentiable function u defined on a set in $\mathbb{R}^d$.

Vectors They are in $\mathbb{R}^d$ for some $d = 1, 2, \ldots$ and they are denoted by bold-faced letters, unless $d = 1$. $\mathbf{u} \cdot \mathbf{v}$ denotes the scalar product between vectors $\mathbf{u}$ and $\mathbf{v}$ so that $|\mathbf{u}| = (\mathbf{u} \cdot \mathbf{u})^{\frac{1}{2}}$ gives the length of $\mathbf{u}$.

VARIATIONAL CALCULUS is the branch of mathematics that treats optimization of functions of infinitely many variables through extensions of the concepts and techniques of ordinary calculus. Nearly as old as calculus, it too represents a continuation of ancient Greek efforts to provide idealized models for our universe. Already implicit in Greek geometry is the timeless conviction that a straight line supplies the shortest path between points. Moreover, by the first century, Zenodorus had proved geometrically that a circle encloses more area than isoperimetric polygons and conjectured that a sphere has maximal volume among solids with boundary surfaces of equal area. The belief that light travels optimal paths was incorporated by Heron four centuries later in his deduction of the principle of reflection.

I. HISTORICAL PERSPECTIVE

The modern chapter opened in 1686 with Newton's investigation of optimal design for projectiles of revolution. And a decade later, Johan Bernoulli challenged his fellow mathematicians, including Newton and Leibniz, to solve a problem in Galilean dynamics by finding the *brachistochrone*. This was the name he proposed to give the curve joining fixed points displaced horizontally and vertically along which a bead would fall from rest without friction in least time. (Fig. 1).

Some sixty years earlier, Galileo had conjectured that the optimal curve would be a circular arc, but Bernoulli and all fellow respondents discovered that it must lie on a cycloid, the curve traced by a point on the rim of a wheel as it rolls along a straight horizontal path. These early mathematicians used reasoning based on physical analogy and geometry that would be questioned today and all assumed tacitly that a minimizing curve exists. (However, Bernoulli subsequently tried to prove geometrically that the cycloidal arc does minimize the time of fall.) Most arguments employed were ad hoc but that of Bernoulli's older brother Jakob offered greater scope. In his solution, Jakob computed the time of fall along a curve which varied only slightly from an hypothesized brachistochrone. Then requiring that this exceed the time along the brachistochrone itself, he introduced a limiting argument to obtain parametric equations for the optimal curve. This variational approach was used to good effect on related problems by the Bernoullis and by Johan's outstanding disciple, Euler. As subsequently developed by Lagrange and his successors, the limiting argument became that of differentiating a parameter-dependent integral, and the method became known as the *calculus of variations*, or more succinctly, *variational calculus*.

This branch of analysis has remained active since its inception partly because the desire for optimal behavior continues to spur Western development. In particular, it provides the framework for extracting governing equations for a process, physical or mathematical, that

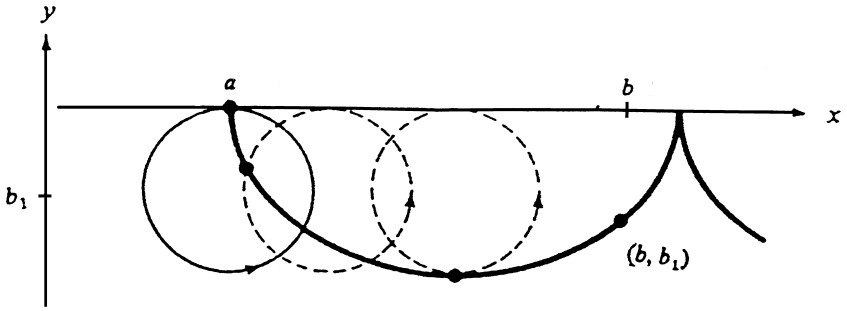


FIGURE 1 Bernoulli's brachistochrone problem.

is supposedly optimal relative to a class of competitors. Moreover, these equations can have nonoptimal solutions which suggests that other systems of equations may admit variational derivation. In previous centuries it was natural to employ variational methods when formulating and exploring principles of virtual work, least action, minimum potential energy, and under Hamilton, stationary action leading to Hamilton's equations of motion. There are also modern variational derivations of the equations of electrodynamics, relativity, and quantum mechanics.

Variational calculus together with the rest of mathematics experienced the 19th century transition from accepting intuitive conviction to demanding rigorous deduction in proofs. In lectures commencing around 1870, Weierstrass presented counterexamples to cherished claims of Gauss, Dirichlet, and Riemann concerning existence of minima in variational calculus. In modern terms, these minima were being confused with infima (greatest lower bounds), but the debacle of Dirichlet's Principle as it was called, generated issues that were only fully resolved a half century later through methods of functional analysis involving the Lebesgue integral. Even the subsequently developed tools of nonlinear functional analysis were tested and refined through application to problems in variational calculus. Consequently, it is now possible to give descriptive conditions that guarantee existence of an extremal solution to a problem in variational calculus. These conditions usually involve convexity which, expressed in various forms, lies at the heart of this subject.

Much recent activity concerns problems in optimal control that gained prominence during World War II from the need to steer a vehicle or projectile along a path deemed optimal relative to some targeted objective. Restricted versions of these problems are covered by standard variational methods and analysis of general problems led Pontrjagin and others to uncover a new principle of optimality. This principle is now invoked in disciplines as widely ranging as submarine navigation and economics, and it too has been incorporated in modern variational calculus.

Regardless of theoretical assurances, most problems in variational calculus or optimal control cannot be solved explicitly or exactly because of the nonlinearities involved. Consequently, the need for reliable methods of approximation has been apparent from the outset. In fact, Euler used piecewise linear approximations of an unknown extremal function to derive its governing equation. But serious consideration of approximation methods had to await improved understanding of convergence. Then in 1908, to approximate mode shapes of vibrating elastic plates, Ritz successfully modified earlier methods of Rayleigh and placed them in a Hilbert space setting. This approach, further modified by Galerkin and others, has been widely extended in a host of schemes formulated to take advantage of rapidly increasing computational strength.

II. ANALYTIC FOUNDATIONS

The basic problem of the variational calculus is to find or characterize a function $\mathbf{y}_0$ which optimizes (maximizes or minimizes) an integral of the form

$$F(\mathbf{y}) = \int_a^b f(x, \mathbf{y}(x), \mathbf{y}'(x))\, dx \tag{1}$$

on a set $\mathcal{D}$ of differentiable vector-valued functions $\mathbf{y}$ defined on the compact interval $[a, b]$. Here, f is a given real-valued function on a set $[a, b] \times D$ for a given domain D of $\mathbb{R}^{2d}$, and we assume that the composite $f[\mathbf{y}(x)] \stackrel{\text{def}}{=} f(x, \mathbf{y}(x), \mathbf{y}'(x))$ is Riemann integrable over $[a, b]$, for all $\mathbf{y}$ in $\mathcal{D}$. We also assume initially that the functions $\mathbf{y}$ in $\mathcal{D}$ meet fixed endpoint conditions of the form

$$\mathbf{y}(a) = \mathbf{a}, \qquad \mathbf{y}(b) = \mathbf{b}. \tag{2}$$

Since the maximum value of F on $\mathcal{D}$ would be given by a function $\mathbf{y}_0$ that minimizes $-F$ on $\mathcal{D}$, we can restrict attention to minimization issues.

For example, the minimum value of

$$L(\mathbf{y}) = \int_a^b |\mathbf{y}'(x)|\, dx \tag{3}$$

on

$$\mathcal{D} = \{\mathbf{y} \in C^1[a, b]\colon \mathbf{y}(a) = \mathbf{a}, \mathbf{y}(b) = \mathbf{b}\} \tag{4}$$

should provide the shortest length of a smooth curve joining points $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^d$, when we suppose that a typical curve is parametrized by a continuously differentiable function $\mathbf{y}$ on the common interval $[a, b]$. We expect the minimum value to be $|\mathbf{a} - \mathbf{b}|$ which is given by the straight line path $\mathbf{y}_0$ joining the points.

On the other hand, in a formulation consistent with Fig. 1, Bernoulli's brachistochrone would be given by a function $\mathbf{y}_0$ that minimizes the integral

$$F(y) = \int_a^b y^{-\frac{1}{2}}(x)\sqrt{1 + y'(x)^2}\, dx \tag{5}$$

on

$$\mathcal{D} = \{y \in C^1[a, b]\colon y(a) = 0,\ y(b) = b_1 > 0,\ y(x) \geq 0\} \tag{6}$$

Here, $f(x, y, y') = y^{-\frac{1}{2}}\sqrt{1 + y'^2}$, where y' designates a real variable, but it is not evident that a minimizing function exists, nor how to find one, if it does.

When a C^1 function $\mathbf{y}_0$ minimizes F on $\mathcal{D}$ of Eq. (4) it usually minimizes F on the larger class $\hat{\mathcal{D}}$ of piecewise C^1 functions with the same endpoint values, since each such function is almost C^1. However, it is possible that

only a $\hat{C}^1$ function can minimize F on $\hat{\mathcal{D}}$. This is illustrated by an integral of Weierstrass

$$F(y) = \int_{-1}^{1} y^2(1 - y')^2 \, dx \tag{7}$$

which is zero for the piecewise C^1 function $y_0(x) = \max(x, 0)$, but is positive for any other function in

$$\hat{\mathcal{D}} = \{y \in \hat{C}^1[-1, 1]: y(-1) = 0, \quad y(1) = 1\}. \tag{8}$$

Thus, in variational calculus even a simple problem need not have a solution in the class of functions of interest. As classical theory developed, functions were routinely assigned as much differentiability as desired, and smooth solutions to many problems were obtained through methods to be discussed presently. Modern theory has focused on eliminating as much differentiability/continuity as possible in order to extend the search to a class of objects large enough to guarantee existence of a minimizer. Whether the resulting object itself has practical utility is a matter of debate, but it can often be usefully approximated by one that does.

A. Necessary Conditions

If $\mathbf{y}_0$ minimizes

$$F(\mathbf{y}) = \int_a^b f[\mathbf{y}(x)] \, dx = \int_a^b f(x, \mathbf{y}(x), \mathbf{y}'(x)) \, dx \tag{9}$$

on

$$\mathcal{D} = \{\mathbf{y} \in C^1[a, b]: y(a) = \mathbf{a}, \quad y(b) = \mathbf{b}\} \tag{10}$$

then the *Gâteaux variation*

$$\delta F(\mathbf{y}_0; \mathbf{v}) \stackrel{\text{def}}{=} \lim_{\varepsilon \to 0} \varepsilon^{-1}\big[F(\mathbf{y}_0 + \varepsilon \mathbf{v}) - F(\mathbf{y}_0)\big] = 0 \tag{11}$$

in each "direction" $\mathbf{v}$ in $C^1[a, b]$ for which $\mathbf{y}_0 + \varepsilon \mathbf{v}$ is in $\mathcal{D}$ when ε is a sufficiently small real number, and the limit exists.

When f is continuous together with each component of its partial gradients $f_{\mathbf{y}}$ and $f_{\mathbf{y}'}$, then in general

$$\delta F(\mathbf{y}_0; \mathbf{v}) = \int_a^b \{f_{\mathbf{y}}[\mathbf{y}_0(x)] \cdot \mathbf{v}(x) + f_{\mathbf{y}'}[\mathbf{y}_0(x)] \cdot \mathbf{v}'(x)\} \, dx \tag{12}$$

and $\mathbf{y}_0 + \varepsilon \mathbf{v}$ is in $\mathcal{D}$ for small ε when $\mathbf{v}(a) = \mathbf{v}(b) = \mathbf{0}$.

Finally, $\delta F(\mathbf{y}_0\,; \mathbf{v}) = 0$ for all such $\mathbf{v}$ if and only if $\mathbf{y}_0$ is a solution of the first E-L (Euler-Lagrange) equation

$$\frac{d}{dx} f_{\mathbf{y}'}[\mathbf{y}_0(x)] = f_{\mathbf{y}}[\mathbf{y}_0(x)] \tag{13}$$

(interpreted componentwise).

[If Eq. (13) holds, then indeed from Eq. (12):

$$\delta F(\mathbf{y}_0; \mathbf{v}) = f_{\mathbf{y}'}[\mathbf{y}_0(x)] \cdot \mathbf{v}(x) \,|_a^b = 0 \text{ when } \mathbf{v}(a) = \mathbf{v}(b) = \mathbf{0}, \tag{14}$$

and the converse assertion follows by showing that the continuous function

$$\mathbf{g}(x) \stackrel{\text{def}}{=} f_{\mathbf{y}'}[\mathbf{y}_0(x)] - \int_a^x f_{\mathbf{y}}[\mathbf{y}_0(t)] \, dt \tag{15}$$

is constant.]

Now Eq. (13) is usually a second-order system of nonlinear differential equations for an unknown minimizing function $\mathbf{y}_0$. This system need not have any solutions on a given interval $[a, b]$ with prescribed endpoint values, and even if it does, finding an explicit solution is quite challenging. However, when f is continuously differentiable the system for a minimizing $\mathbf{y}_0$ does have a useful first integral. Then

$$\begin{aligned} h(x) &\stackrel{\text{def}}{=} f[\mathbf{y}_0(x)] - f_{\mathbf{y}'}[\mathbf{y}_0(x)] \cdot \mathbf{y}'_0(x) \\ &= \int_a^x f_x[\mathbf{y}_0(t)] \, dt + \text{const.} \end{aligned} \tag{16}$$

which gives the second E-L equation,

$$\frac{d}{dx} h(x) = f_x[\mathbf{y}_0(x)], \quad x \in (a, b). \tag{17}$$

Note that unlike Eq. (13), this second equation is real-valued; indeed, it represents a residual effect of variations with respect to the independent variable x.

When a piecewise C^1 function $\mathbf{y}_0$ minimizes F on the corresponding set

$$\hat{\mathcal{D}} = \{\mathbf{y} \in \hat{C}^1[a, b]: \mathbf{y}(a) = \mathbf{a}, \mathbf{y}(b) = \mathbf{b}\} \tag{18}$$

then although $\mathbf{g}$ of Eq. (15) remains constant, we cannot conclude that Eq. (13) follows. Instead, as was shown independently by Weierstrass and Erdmann, we can say that Eq. (13) holds on each interval that excludes corner points of $\mathbf{y}_0$ which can occur only at x values c where $\mathbf{y}'_0(c+) \neq \mathbf{y}'_0(c-)$ and

$$f_{\mathbf{y}'}[\mathbf{y}_0(c+)] = f_{\mathbf{y}'}[\mathbf{y}_0(c-)]. \tag{19}$$

Similarly, when f is C^1, Eq. (17) holds on each interval that excludes corner points of $\mathbf{y}_0$ which can occur only at values c where in addition

$$(f - \mathbf{y}' \cdot f_{\mathbf{y}'})[\mathbf{y}_0(c+)] = (f - \mathbf{y}' \cdot f_{\mathbf{y}'})[\mathbf{y}_0(c-)]. \tag{20}$$

Equations (19) and (20) are known, respectively, as the first and second W-E (Weierstrass-Erdmann) corner conditions. [A $\mathbf{y}_0$ that minimizes F only on a sufficiently restricted subset of $\hat{\mathcal{D}}$ in Eq. (18) can have a corner point where Eq. (19) holds but Eq. (20) does not.]

For the Weierstrass example of Eq. (7), these conditions limit the location of a possible corner point $(c, y_0(c))$ of a minimizing y_0 to values where $y_0^2(c)(1 - y'_0(c+)) = y_0^2(c)(1 - y'_0(c-))$, and since $y'_0(c+) \neq y'_0(c-)$: $y_0(c) = 0$. Moreover from Eq. (17), we see that $h(x) = y_0^2(1 - y_0'^2)(x)$ is constant so that it too is

zero. Thus, on an interval excluding corner points, either $y_0(x) = 0$ or $y_0'(x) = \pm 1$. Assembling these facts we arrive at $y_0(x) = \max(x, 0)$ as the only candidate in $\hat{\mathcal{D}}$ of Eq. (8).

If we replace f by $-f$ in Eq. (9), then all necessary conditions obtained so far are unaffected; thus they would also hold for a $\mathbf{y}_0$ that maximizes F on $\mathcal{D}$ (or on $\hat{\mathcal{D}}$). To derive a condition that arises only for a minimizer, Weierstrass employed a small variation whose derivative at a single point remains constant as the variation diminishes in magnitude to zero. He found that if $\mathbf{y}_0$ minimizes F on $\mathcal{D}$ then for each $x \in [a, b]$:

$$f(x, \mathbf{y}_0(x), \mathbf{y}') - f(x, \mathbf{y}_0(x), \mathbf{y}_0'(x)) \geq f_{\mathbf{y}'}[\mathbf{y}_0(x)] \cdot (\mathbf{y}' - \mathbf{y}_0'(x)) \tag{21}$$

for all vectors $\mathbf{y}'$ of interest. [If $\mathbf{y}_0$ minimizes F on $\hat{\mathcal{D}}$ then this inequality holds at a corner point $(c, \mathbf{y}_0(c))$ when $\mathbf{y}_0'(x)$ is replaced by either $\mathbf{y}_0'(c+)$ or $\mathbf{y}_0'(c-)$.] Assuming enough continuity it follows that the matrix having elements $f_{y_i' y_j'}(x, \mathbf{y}_0(x), \mathbf{y}_0'(x))$ arranged in natural order is positive semidefinite. When it is positive definite, Hilbert showed that a C^1 solution $\mathbf{y}_0$ of Eq. (13) is necessarily C^2.

B. Stationarity

Condition (11), $\delta F(\mathbf{y}_0; \mathbf{v}) = 0$ for all $\mathbf{v}$ with $\mathbf{v}(\boldsymbol{a}) = \mathbf{v}(\boldsymbol{b}) = \mathbf{0}$, is not sufficient to infer that $\mathbf{y}_0$ is an extremum for F on $\mathcal{D}$. It would also hold when $\mathbf{y}_0$ is only a local extremum, that is, an extremum in an appropriately defined neighborhood of $\mathbf{y}_0$. (In particular, $\mathbf{y}_0$ could be a *strong* local extremum relative to a neighborhood in which $|\mathbf{y} - \mathbf{y}_0|$ is uniformly small, or a *weak* local extremum relative to a neighborhood in which $|\mathbf{y}' - \mathbf{y}_0'|$ is also uniformly small.) But condition (11) alone cannot be used to infer extremal behavior in any direction $\mathbf{v}$. Instead, it characterizes F as having stationary behavior at $\mathbf{y}_0$ in that at $\mathbf{y}_0$, F exhibits no predisposition to either increase or decrease in any associated direction $\mathbf{v}$. For the same reason, solutions of the E-L system [Eq. (13)] are sometimes called *stationary functions*, although the misleading terms *extremal function* or *extremal* are common in the literature.

C. Sufficient Conditions; Convexity

It is obvious graphically that a real-valued function f of a single real variable t has a minimum value at a point where its derivative is zero provided that its graph resembles an arc of an upward directed parabola. Such functions are now said to be convex, a condition that can be inferred if f has a nonnegative second derivative. However, this implies a weaker property, that for all t, t_0 of interest:

$$f(t) - f(t_0) \geq f'(t_0)(t - t_0), \tag{22}$$

which indicates that the graph of f lies above (or on) each of its tangent lines. Finally, (22) implies that on each interval $[t_0, t_1]$ of definition, the graph of f lies below the chordal line joining the endpoints, so that

$$f((1-s)t_0 + st_1) \leq (1-s)f(t_0) + sf(t_1), \quad 0 \leq s \leq 1. \tag{23}$$

The last condition is purely geometrical and does not require differentiability or even continuity of f, although the latter can be deduced when f is bounded. But for our immediate purposes, (22) is more natural; indeed, we already encountered one version of it in (21).

Accordingly, we say that the function $f(x, \mathbf{y}, \mathbf{y}')$ is *convex in* $\mathbf{y}$ *and* $\mathbf{y}'$ when for each fixed x,

$$f(x, \mathbf{y}, \mathbf{y}') - f(x, \mathbf{y}_0, \mathbf{y}_0') \geq f_{\mathbf{y}}(x, \mathbf{y}_0, \mathbf{y}_0') \cdot (\mathbf{y} - \mathbf{y}_0) + f_{\mathbf{y}'}(x, \mathbf{y}_0, \mathbf{y}_0') \cdot (\mathbf{y}' - \mathbf{y}_0') \tag{24}$$

for all vectors $\mathbf{y}$, $\mathbf{y}'$, $\mathbf{y}_0$, and $\mathbf{y}_0'$ of interest. In this case, from Eqs. (9) and (12), it follows that

$$F(\mathbf{y}) - F(\mathbf{y}_0) \geq \delta F(\mathbf{y}_0; \mathbf{v}) \qquad \text{for} \quad \mathbf{v} = \mathbf{y} - \mathbf{y}_0,$$

which ascribes a directional convexity to F.

In particular, when $\mathbf{y_0}$ in $\mathcal{D}$ of (10) is a solution of Eq. (13), then from Eq. (14), $\delta F(\mathbf{y}_0; \mathbf{v})$ is zero for $\mathbf{y}$ in $\mathcal{D}$, so that $\mathbf{y}_0$ minimizes F on $\mathcal{D}$. If $\mathbf{y}_0$ is in $\hat{\mathcal{D}}$ of (18) and satisfies the E-L system [Eq. (13)] on intervals between corner points, at each of which it meets the first W-E condition [or equivalently, $\mathbf{g}(x)$ of Eq. (15) is constant] then under (24), $\mathbf{y}_0$ minimizes F on $\hat{\mathcal{D}}$. Moreover, under a slightly stronger version of (24) there is at most one such $\mathbf{y}_0$ in $\hat{\mathcal{D}}$.

Most functions f do not exhibit the full convexity of (24) but many do, including the one-dimensional examples, $f(x, y, y') = \sqrt{1 + y'^2}$, $\sqrt{y'^2 + y^{-2}}$ and $e^x\sqrt{1 + y'^2} + (\sin x)y$. Even when the integrand f is not sufficiently convex, it is possible that an invertible change of dependent variables of the form $\mathbf{y} = \boldsymbol{\varphi}(\boldsymbol{\eta})$ results in a new integrand that is suitably convex. Although the brachistochrone integrand $f(x, y, y') = \sqrt{(1 + y'^2)y^{-1}}$ of Eq. (5) does not satisfy (24), the transformation $y = \eta^2/2$ (so $y' = \eta\eta'$) results in the new integrand $\tilde{f}(x, \eta, \eta') = \sqrt{2(\eta'^2 + \eta^{-2})}$ which does, and this gives an elementary means of proving that the cycloidal arc is the brachistochrone. If usable convexifying transformations in $\mathbf{y}$ cannot be found, it may be possible to recast a problem so that a new integrand is convex, but it is an open question whether every problem with a known minimum can be transformed into one that has an integrand satisfying (24).

In fact, explorations of (24) are of surprisingly recent origin. Sufficiency was first considered by Legendre who in 1786 attempted to establish local minimization of F at $\mathbf{y}_0$ by requiring in effect, the positivity of $\delta^2 F(\mathbf{y}_0; \mathbf{v})$, the second variation of F. In the one-dimensional case

when $\mathbf{y} = y$ and f are C^2, using Eqs. (11) and (12) it is straightforward to show that

$$\delta^2 F(y_0; v) = \int_a^b \left(f_{yy} v^2 + 2 f_{yy'} v v' + f_{y'y'} v'^2 \right) dx \quad (25)$$

where the partial derivatives of f are evaluated at $(x, y_0(x), y_0'(x))$. Assuming that $\delta F(y_0; v) = 0$, it is true that having $\delta^2 F(y_0; v) > 0$ ensures local minimization of F at y_0 *in the direction* v, but flaws in Legendre's efforts to infer positivity of $\delta^2 F(y_0; v)$ from that of $f_{y'y'}(x, y_0(x), y_0'(x))$ impeded progress with this approach. Then, in 1836 Jacobi carefully examined Legendre's analysis and showed that local minimization can be established when f is C^4, provided that a positive solution u is available for the linear differential equation

$$(f_{yy'} u + f_{y'y'} u')' = f_{yy} u + f_{yy'} u' \quad (26)$$

now known as *Jacobi's equation*. Moreover, Jacobi proved that this new condition is essential for minimization, even when $f_{y'y'}$ is positive.

In these early works, convexity manifested itself through positivity of second derivatives, but the next wave of investigations, launched by Weierstrass in lectures around 1879, brought a new formulation. After recognizing that (21) was a necessary condition for minimization, Weierstrass began to study integrand functions $f(x, \mathbf{y}, \mathbf{y}')$ that satisfy the related inequality

$$f(x, \mathbf{y}_0, \mathbf{y}') - f(x, \mathbf{y}_0, \mathbf{y}_0') \geq f_{\mathbf{y}'}(x, \mathbf{y}_0, \mathbf{y}_0') \cdot (\mathbf{y}' - \mathbf{y}_0') \quad (27)$$

for all vectors $\mathbf{y}_0'$ and $\mathbf{y}'$ of interest. (In modern terminology, such functions are convex in $\mathbf{y}'$ only.) In subsequent work extended substantially by Hilbert (1900), it was shown that under (27), a solution $\mathbf{y}_0$ of E-L (13) in $\mathcal{D}$ of (10) does minimize F on $\hat{\mathcal{D}}$ of (18) when $\mathbf{y}_0$ can be embedded in a suitable family (a field) of other solutions to Eq. (13). The resulting Weierstrass-Hilbert field theory of sufficiency dominated subsequent development, although it can be rather difficult to find a usable field. Indeed, in order to guarantee theoretically that a given solution $\mathbf{y}_0$ can be embedded in a field large enough to ensure local minimization for a C^3 f, one must prove that an auxiliary linear system has solutions that remain linearly independent on the interval (a, b). In the simple case where $\mathbf{y}$ is real-valued, this condition requires a positive solution u of the Jacobi equation (26). On the other hand, condition (24) is effective only on a fixed interval while field theory methods are in principle more adaptable.

III. CONSTRAINTS

The fixed endpoint conditions $\mathbf{y}(a) = \mathbf{a}$, $\mathbf{y}(b) = \mathbf{b}$ considered previously represent constraints on the admissible functions. When other constraints are imposed the results must be modified.

For example, if we require only $\mathbf{y}(a) = \mathbf{a}$, then at b a minimizing function $\mathbf{y}_0$ must necessarily satisfy the natural boundary conditions

$$f_{\mathbf{y}'}[\mathbf{y}_0(b)] = \mathbf{0} \quad (28)$$

whose relevance is visible in Eq. (14). [Conversely, under the convexity of (24), any solution $\mathbf{y}_0$ of E-L on (a, b) with $\mathbf{y}_0(a) = \mathbf{a}$, that satisfies Eq. (28) must minimize.] Requiring only $\mathbf{y}(a) = \mathbf{a}$ constrains the right endpoint of the graphs of the admissible functions to lie on the (hyper)plane $x = b$ in $\mathbb{R}^{d+1}$. When this plane is replaced by a smooth transversal surface T with equation $\tau(x, \mathbf{y}) = 0$, then b is not fixed and the corresponding natural boundary condition at T is

$$[h\tau_{\mathbf{y}} - \tau_x f_{\mathbf{y}'}]_T = \mathbf{0}, \quad (29)$$

where h is given in Eq.(16). In the so-called infinite horizon case where T reduces to the line $\mathbf{y} = \mathbf{b}$ so that $\tau(x, \mathbf{y}) = \mathbf{y} - \mathbf{b}$, then for minimization b must be selected to make $h(b, \mathbf{b}) = 0$. Of course, similar constraints at the left endpoint result in corresponding natural boundary conditions, and under (27) Weierstrass-Hilbert field theory of sufficiency can in principle be modified to accommodate any of these endpoint conditions. For other types of constraints, it is usually necessary to modify the integrand.

A. Isoperimetric Constraints

Variational methods can be used to examine the Zenodorus conjecture concerning the maximal planar area enclosable by isoperimetric curves (those having the same length). The most natural approach involves optimizing one integral function (the area) while holding constant the values of another (the length). Constraints of this type, first investigated by Euler, are termed *isoperimetric* as are optimization problems involving only a finite number of such constraints. Suppose we wish to minimize F of Eq. (9) on $\mathcal{D}$ of Eq. (10) under the *isoperimetric* constraints

$$\mathbf{G}(\mathbf{y}) \stackrel{\text{def}}{=} \int_a^b \mathbf{g}(x, \mathbf{y}, \mathbf{y}')\, dx = \boldsymbol{\ell}, \text{ a constant vector}, \quad (30)$$

where $\mathbf{g}$ is a given vector-valued function. To do so, we introduce a Lagrangian multiplier vector $\boldsymbol{\lambda}$ and seek instead a $\mathbf{y}_0$ in $\mathcal{D}$ that minimizes

$$\tilde{F}(\mathbf{y}) = F(\mathbf{y}) + \boldsymbol{\lambda} \cdot \mathbf{G}(\mathbf{y}) \quad (31)$$

on $\mathcal{D}$. Clearly, any such $\boldsymbol{\lambda}$-dependent $\mathbf{y}_0$ also minimizes F on the subset of $\mathcal{D}$ where $\mathbf{G}(y) = \mathbf{G}(\mathbf{y}_0)$, so that if now $\boldsymbol{\lambda}$ can be selected to make $\mathbf{G}(\mathbf{y}_0) = \boldsymbol{\ell}$, an associated $\mathbf{y}_0$ would supply a minimizer. It can also be shown that if a

minimizing $\mathbf{y}_0$ for this isoperimetric problem exists, then in principle so must an associated $\boldsymbol{\lambda}$ unless the determinant with elements $\delta G_i(\mathbf{y}_0\,;\mathbf{v}_j)$ vanishes for all choices of $\mathbf{v}_j$ with $\mathbf{v}_j(a)=\mathbf{v}_j(b)=\mathbf{0}$. Here G_i is the ith component of $\mathbf{G}$.

Note that if $\mathbf{y}_0$ minimizes $\tilde{F}$ of Eq. (31) on $\mathcal{D}$, then it also minimizes F on $\mathcal{D}$ under the corresponding inequality constraint

$$\boldsymbol{\lambda}\cdot\mathbf{G}(\mathbf{y})\le\boldsymbol{\lambda}\cdot\mathbf{G}(\mathbf{y}_0)=\boldsymbol{\lambda}\cdot\boldsymbol{\ell} \tag{32}$$

which can be realized through inequalities involving the signs of the components of $\boldsymbol{\lambda}$ and $\mathbf{G}$.

In practice, we seek a minimizing $\mathbf{y}_0$ for $\tilde{F}$ on $\mathcal{D}$ among solutions of the E-L equation (13) for $\tilde{f}=f+\boldsymbol{\lambda}\cdot\mathbf{g}$, the integrand of $\tilde{F}$. Then we select those $\boldsymbol{\lambda}$ for which $G(\mathbf{y}_0)=\boldsymbol{\ell}$, and try to establish minimization perhaps by appeal to convexity of the resulting $\tilde{f}$. To illustrate this procedure, consider the simple but nontrivial problem of minimizing

$$F(y)\overset{\text{def}}{=}\int_0^1 4y(x)\,dx \qquad \text{on}$$

$$\mathcal{D}=\{y\in C^1[0,1]\colon y(0)=0,\ y(1)=1\} \tag{33}$$

under the isoperimetric constraint

$$G(y)\overset{\text{def}}{=}\int_0^1 y'(x)^2\,dx=4/3. \tag{34}$$

Here, $\tilde{f}(x,y,y')=4y+\lambda y'^2$ provides the E-L equation $\frac{d}{dx}(2\lambda y')=4$ with solution $y_0(x)=(x^2+(\lambda-1)x)/\lambda$ in $\mathcal{D}$. Moreover $G(y_0)=1+(3\lambda^2)^{-1}$, and this equals $4/3$ when $\lambda=\pm1$. Now $\tilde{f}=4y+(1)y'^2$ is convex in the sense of (24) so that $y_0(x)=x^2$ solves our problem. [The other choice, $\lambda=-1$, gives a y_0 which in fact *maximizes* F on $\mathcal{D}$ under (34) since with it, $-\tilde{f}=-4y+y'^2$ is again convex in the sense of (24).]

If we replace $\mathcal{D}$ of Eq. (33) by $\mathcal{D}_1=\{y\in C^1[0,1]\colon y(0)=0\}$ then minimization of $\tilde{F}$ on $\mathcal{D}_1$ requires a y_1 that satisfies the natural boundary condition of Eq. (28), $\tilde{f}_{y'}[y_1(1)]=0$ or $y_1'(1)=0$; this gives the minimizing function $y_1(x)=x^2-2x$. Finally we see that the same y_0 or y_1 minimizes F on $\mathcal{D}$ or $\mathcal{D}_1$, respectively, under the inequality constraint $G(y)\le 4/3$.

B. Eigenvalue Problems

A nonuniform string of linear density $\rho=\rho(x)$ and local tensile resistance $\tau=\tau(x)$ stretched between fixed supports a distance l apart can be shown to vibrate freely with small deflection in mode shapes $y_n=y_n(x)$ at natural frequency ω_n for $n=1,2,\ldots$. Furthermore $\lambda_n\overset{\text{def}}{=}\omega_n^2=\min\{\int_0^1\tau y'^2\,dx\colon y\in\mathcal{D}_n\}$ where $\mathcal{D}_1=\{y\in C^1[0,l]\colon y(0)=y(l)=0;\ \int_0^1\rho y^2dx=1\}$ and for $n>1$: $\mathcal{D}_n=\{y\in\mathcal{D}_1\colon\int_0^1\rho y y_k\,dx=0,k=1,2\ldots n-1\}$ assuming successive identification of modal functions y_k is possible. (For a uniform string, τ and ρ are positive constants; then $y_n(x)=\sin(n\pi x/l)$ and $\omega_n=n\omega_1$ for $n=1,2\ldots$, where $\omega_1^2=\tau\pi^2/\rho l^2$.) In 1873 Lord Rayleigh showed how to estimate a few lower frequencies ω_n using approximations of associated modal functions y_n. Then Weyl, Courant, and others found alternate characterizations for eigenvalues λ_n that replace use of eigenfunctions y_n with more general assessments involving maxima of minima (or minima of maxima) of values of F on subfamilies of $\mathcal{D}$. For example, when $\rho=1$, λ_n is the *maximum* (over all sets of continuous functions $\varphi_1,\varphi_2,\ldots,\varphi_{n-1}$) of the *minimum* value of $F(y)$ for all y in $\mathcal{D}_1$ for which $\int_0^1 y\varphi_j\,dx=0$, $j<n$.

Here, the λ_n are the eigenvalues of an associated Sturm-Liouville system, and this variational approach extends to characterizing eigenvalues of other linear operators on Hilbert spaces.

C. Lagrangian Constraints

Certain applications require minimizing F of Eq. (9) on $\mathcal{D}$ of Eq. (10) under interval constraints of the form

$$\mathbf{g}[\mathbf{y}(x)]=\mathbf{g}(x,\mathbf{y}(x),\mathbf{y}'(x))=\mathbf{0},\quad a\le x\le b \tag{35}$$

first considered by Lagrange. Again $\mathbf{g}$ is a given vector-valued function. Upon introducing a corresponding multiplier function $\boldsymbol{\lambda}(x)$, we see that a $\mathbf{y}_0$ in $\mathcal{D}$ for which $g[\mathbf{y}_0(x)]\equiv 0$ will solve this problem provided that for some $\boldsymbol{\lambda}$, $\mathbf{y}_0$ minimizes

$$\tilde{F}(\mathbf{y})=F(\mathbf{y})+\int_a^b\boldsymbol{\lambda}(x)\cdot\mathbf{g}[\mathbf{y}(x)]\,dx \tag{36}$$

on $\mathcal{D}$. Moreover this $\mathbf{y}_0$ will also minimize F on $\mathcal{D}$ under the Lagrangian inequality constraint

$$\boldsymbol{\lambda}(x)\cdot\mathbf{g}[\mathbf{y}(x)]\le 0 \tag{37}$$

which can be realized through inequalities involving the components of $\boldsymbol{\lambda}$ and $\mathbf{g}$. Thus we would seek $\mathbf{y}_0$ in $\mathcal{D}$ and a function $\boldsymbol{\lambda}$ on $[a,b]$ that satisfy the differential system

$$\frac{d}{dx}\tilde{f}_{\mathbf{y}'}[\mathbf{y}_0(x)]=\tilde{f}_{\mathbf{y}}[\mathbf{y}_0(x)] \tag{38}$$

$$\mathbf{g}[\mathbf{y}_0(x)]=\mathbf{0} \tag{39}$$

where $\tilde{f}=f+\boldsymbol{\lambda}\cdot\mathbf{g}$. If for some resulting $\boldsymbol{\lambda}$, $\tilde{f}$ is, say, convex in the sense of (24), then an associated $\mathbf{y}_0$ would solve the problem. Note that Eq. (38) usually involves the derivatives of the multiplier functions $\boldsymbol{\lambda}$.

For problems with Lagrangian constraints, the presence of multiplier functions is unavoidable. If $\mathbf{y}_0$ minimizes F on $\mathcal{D}$ under (35), then in general it can be shown that there exists a $\boldsymbol{\lambda}_0$ (either 0 or 1) and a multiplier function $\boldsymbol{\lambda}$ satisfying Eq. (38) when

$$\tilde{f} = \lambda_0 f + \boldsymbol{\lambda} \cdot \mathbf{g}.$$

Moreover, if $\mathbf{y}_0$ minimizes F on $\mathcal{D}$ under the corresponding inequality constraint

$$\mathbf{g}[\mathbf{y}(x)] \leq \mathbf{0} \quad \text{(interpreted componentwise)} \tag{40}$$

then in the same sense, the multiplier functions meet the Karush-Kuhn-Tucker conditions $\boldsymbol{\lambda}(x) \geq \mathbf{0}$. A component constraint g_i is said to be active at points x where its associated multiplier $\lambda_i(x)$ is nonzero and inactive otherwise.

D. Problem Conversions; Higher Derivatives

Lagrangian constraints can be introduced to replace constraints of other types and to unify the theoretical appearance of problems in variational calculus. For example, the isoperimetric problem of Eqs. (33) and (34) can be replaced by one of finding a vector function $\mathbf{y} = (y, y_1, y_2)$ with components in $C^1[0, 1]$ that has minimum last coordinate value $y_2(1)$ among those that satisfy the Lagrangian system

$$y_1' = g(x, y, y') = y'^2$$
$$y_2' = f(x, y, y') = 4y$$

with $\mathbf{y}(0) = \mathbf{0}$ and $y(1) = 1$, $y_1(1) = 4/3$. This approach, attributed to Maier, has proved useful in the development of optimal control theory.

Lagrangian constraints also provide a straightforward means of extending the basic problems in the variational calculus to functions involving higher derivatives. For example, to minimize

$$F^*(y) = \int_a^b f^*(x, y, y', y'')\, dx$$

on $\mathcal{D}^* = \{y \in C^2[a, b]: y(a) = 0, y'(a) = 1, y(b) = 2\}$ we could seek a pair $\mathbf{y} = (y, y_1)$ that minimizes the integral

$$F(\mathbf{y}) = \int_a^b f(x, \mathbf{y}, \mathbf{y}')\, dx$$

on $\mathcal{D} = \{\mathbf{y} \in C^1[a, b]: \mathbf{y}(a) = (0, 1), y(b) = 2\}$ under the Lagrangian constraint

$$y' - y_1 = 0.$$

Here $f(x, \mathbf{y}, \mathbf{y}') = f^*(x, y, y_1, y_1')$, and it is readily seen how to extend this approach to integrals involving even higher derivatives, or, as in the previous example, to replace the integral itself.

Since each new Lagrangian constraint introduces an additional unknown multiplier function, the advisability of using these transformations in a specific application is a matter of taste and experience. However, it is important to be aware that a given problem may be replaced by one more amenable to either theoretical or practical analysis.

IV. OPTIMAL CONTROL PROBLEMS

We operate a land vehicle by applying controls (steering wheel, throttle, brakes, etc.) to alter its state (position, temperature, etc.) over a time interval to achieve some overall purpose, often conceived of in optimal terms. Restrictions are imposed on state variables, controls, and goals.

Assume that the state $\mathbf{y}(t) \in \mathbb{R}^d$ of a controllable dynamical system at time t evolves in time under laws expressed through a given system of differential equations in the form

$$\dot{\mathbf{y}}(t) = \mathbf{G}(t, \mathbf{y}(t), \mathbf{u}(t)), \quad 0 \leq t \leq T \tag{41}$$

where $\mathbf{u}(t)$ represents the values of the control variables at time t assumed to lie in a fixed control region U of $\mathbb{R}^k$.

Suppose we wish to operate this system from a fixed initial state $\mathbf{y}(0) = \mathbf{a}$, over the time interval $[0, T]$ so as to optimize a performance criterion that can be assessed by an integral of the form

$$F(\mathbf{y}, \mathbf{u}, T) = \int_0^T f(t, \mathbf{y}(t), \mathbf{u}(t))\, dt. \tag{42}$$

For example, the choice of $f \equiv 1$, results in optimizing the time-of-travel, T. The problem is said to be autonomous when neither f nor $\mathbf{G}$ depend explicitly on t. It is usually assumed that the control function $\mathbf{u}$ is piecewise continuous.

When the target time T is fixed, we can use convexity to suggest a method of attack. For simplicity, consider only the problem of minimizing

$$F(y, u) = \int_0^T f(t, y(t), u(t))\, dt$$

on

$$\mathcal{D} = \{y \in \hat{C}^1[0, T], u \in \hat{C}[0, T]: y(0) = a_1; y(T) = b_1\}$$

under the state law

$$\dot{y}(t) = g(t, y(t), u(t)). \tag{43}$$

To do so, we introduce the auxiliary function

$$h = f + pg \tag{44}$$

expressed in terms of an unspecified multiplier function $p = p(t)$ (also called the costate or adjoint function). Then, following the analysis in Section III.C, we see that on $\mathcal{D}$, admissible state and control functions y_0 and u_0, respectively, that minimize

$$\tilde{F}(y, u) = \int_0^T [h(t, y(t), u(t)) - p(t)\dot{y}(t)]\, dt \tag{45}$$

must minimize F under Eq. (43).

Moreover, they will minimize $\tilde{F}$ on $\mathcal{D}$ provided that h is convex in y and u, and that they satisfy the appropriate E-L equations (13) together with the Weierstrass-Erdmann

conditions (19), (20) for $\tilde{f}$, the integrand function of $\tilde{F}$. These take the forms

$$-\dot{p} = h_y = f_y + pg_y \tag{46}$$

with continuity of p and $h_0(t) = h(t, y_0(t), u_0(t))$ and

$$h_u(t, y_0(t), u_0(t)) = 0, \tag{47}$$

the latter requiring that the control set U be open. Since at each t, $h(t, y_0(t), u)$ is supposed convex in u, we see that the optimal control $u_0(t)$ *minimizes* this function. In fact, when the control set U is itself convex (an interval), we can replace Eq. (47) by requiring that at each t

$$h(t, y_0(t), u_0(t)) = \min_{u \in U} h(t, y_0(t), u) \tag{48}$$

Thus, to attack our problem, we should seek solutions y_0, u_0, p to the *adjoint equation* (46) that have the minimizing property (48). The underlying principle, first announced in little-read work of Hestenes in 1950, gained wide acceptance only when it was rediscovered by Russian mathematicians. It is usually credited to their leader, and with the replacement of h by $-h$, is known as the "Pontrjagin Maximal Principle." These mathematicians also proved that related conditions are necessary for solution of optimal control problems under even more general target conditions and obtained corresponding natural boundary conditions that characterize optimal targets. Similar results can be obtained to characterize optimal controls in the presence of inequality restrictions on the state and control variables, expressed in the form of Karush-Kuhn-Tucker type conditions.

Although this approach to optimal control emphasizes its relationship to variational calculus, other treatments have proved beneficial, including those of Young, Bellman, and Leitmann. Each has advantages in certain applications, but none makes the solution of control problems elementary. Moreover, an optimal control need not be unique, and even if it is, it can change dramatically with slight changes in initial conditions. In particular, numerical approximations to solutions can behave quite erratically. Finally, as with any extremal problem, optimal controls need not exist—at least within the family of admissible competitors being considered. Some of these difficulties are ameliorated through the probabilistic view taken in the stochastic version of control theory.

V. MINIMIZATION THEORY

Solution of a problem in variational calculus or optimal control can usually be reduced to the search for an element y_0 of a set $\mathcal{S}$ that minimizes the value of a function F defined on $\mathcal{S}$ with values in $(-\infty, \infty]$. $\mathcal{S}$ is usually a subset of a linear space of functions, $\mathcal{Y}$, but F must exhibit some continuity with respect to a topology on $\mathcal{Y}$ (or at least a topology on $\mathcal{S}$) in order to develop a theory. In the century just concluded many candidates for spaces $\mathcal{Y}$ and related topologies have been examined, in particular, those made possible by Lebesgue's extension of the integral in 1902, together with the development of functional analysis to explore the consequences.

What would guarantee existence of a minimizer y_0 is assurance that from a sequence $y_n \in \mathcal{S}$, $n = 1, 2, \ldots$, with values $F(y_n) \searrow \mu = \inf_{y \in \mathcal{S}} F(y)$ we can extract a subsequence $\{y_{n_k}\}$ with a "limit" y_0 in $\mathcal{S}$, for which $F(y_0) \le \lim_{k \to \infty} F(y_{n_k}) = \mu$. Thus F must be *lower semicontinuous* on $\mathcal{S}$ with respect to a topology that gives the sequence $\{y_n\}$ *sequentially compact closure* in $\mathcal{S}$. In work begun in 1915 Tonelli and his successors generalized earlier results of Hilbert (from 1901). They showed that an integral function

$$F(y) = \int_a^b f(x, y, y')\, dx \tag{49}$$

would have these properties on the subset $\mathcal{S}$ of absolutely continuous functions y on $[a, b]$ with $y(a)$ bounded, when for some fixed $p \in (1, +\infty)$

(i) The almost everywhere defined derivatives of y lie in the Lebesgue space $L_p[a, b]$.
(ii) $f(x, y, y') \ge A|y'|^p$ for some positive constant A.
(iii) $f(x, y, y')$ is convex in y'. (50)

These conditions can be relaxed somewhat. For example, the classical convexity of (iii) can be replaced by geometrical convexity as in (23). Moreover, similar results hold when $p = 1$, provided that (ii) is replaced by the coercive requirement $f(x, y, y') \ge |y'|\varphi(|y'|) > 0$ where φ is a continuous function on $[0, +\infty)$ with infinite limit at infinity. Finally, there are corresponding vector-valued versions, even for certain nonconvex f. However, the innocent-looking problem of finding an absolutely continuous function on $[0, 1]$ that minimizes the *positive* integral

$$F(y) = \int_0^1 [(y'^2 - 1)^2 + y^2]\, dx \tag{51}$$

among those with zero endpoint values has no solution, because for small sawtooth functions y with slopes ± 1 except at a finite set of points, $F(y) = \int_0^1 y^2\, dx$ is correspondingly small. And there is a similar integral problem with an absolutely continuous minimizer y_0 that cannot be approximated by smoother functions that almost minimize (the Lavrientiev phenomenon).

Despite this, Ekeland has shown recently that when F is lower semicontinuous and bounded below on a closed set

$\mathcal{S}$ of a Banach space $\mathcal{Y}$, then $\mathcal{S}$ does contain a minimizer for the slightly modified function $\tilde{F}(y) = F(y) + \varepsilon\|y - y_1\|$ with an appropriate choice of $\varepsilon > 0$ and $y_1 \in \mathcal{S}$.

In specific cases, it may be feasible to approximate an unknown but theoretically existent minimizer y_0 through elements y_n of a computationally accessible sequence. Many schemes have been proposed to do so including those based on a Ritz-Galerkin type procedure for selecting a y_n that minimizes F on an n-dimensional closed subset $\mathcal{S}_n$ with known basis for which $\mathcal{S}_n \subseteq \mathcal{S}_{n+1}$ and $\mathcal{S} = \cup_n \mathcal{S}_n$. Then $F(y_n) \geq F(y_{n+1}) \geq F(y_0)$ and under reasonable conditions $F(y_0) = \lim_{n\to\infty} F(y_n)$. Thus $F(y_n)$ approximates $F(y_0)$ from above, but it is much more difficult to arrange that y_n approximates y_0 in any usable sense.

Now, in general, conditions which guarantee existence of a minimizer y_0 for a function F do not permit us to further characterize it. Indeed, Ball and Mizel have produced an integral function F with polynomial integrand similar to that in (51) with an absolutely continuous minimizer y_0 which does not satisfy Eq. (13) even in distributional form. To mirror the classical development we need some means of differentiating F. We have already seen the utility of the Gâteaux variations $\delta F(y_0\,; v)$ of Eq. (11) when they exist. They must exist if y_0 is in a Banach space $\mathcal{Y}$ where F has the *Fréchet derivative* $F'(y_0)$, the unique continuous linear functional L on $\mathcal{Y}$ such that for all $v \in \mathcal{Y}$ with small norm $\|v\|$: $F(y_0 + v) = F(y_0) + Lv + \|v\|\mathcal{Z}(v)$; here, $\mathcal{Z}$ denotes a function with zero limit as $\|v\| \to 0$. Indeed, then $\delta F(y_0; v) = F'(y_0)v$.

When this occurs at an interior point y_0 of $\mathcal{S}$ where F is minimized, then $F'(y_0) = 0$. Moreover, if F is convex on $\mathcal{S}$ in that

$$F(y) - F(y_0) \geq F'(y_0)(y - y_0); \quad y, y_0 \in \mathcal{S} \tag{52}$$

then having $F'(y_0) = 0$ guarantees minimization at y_0 even when y_0 is a boundary point of $\mathcal{S}$. However, in the latter case $F'(y_0)$ need not be zero as is evident for the convex function $F_1(y) = y^2$ on the interval $\mathcal{S} = [2, 3)$ where $y_0 = 2$. In this example we do see that the line through the point $(2, 2^2)$ with slope zero lies below the graph of F_1, and in 1954, Rockafellar proved more generally, that when y_0 is a minimizer for a convex function F on a closed convex set $\mathcal{S}$, then the zero functional is in the set $\partial(F + \delta_{\mathcal{S}})(y_0) = \partial F(y_0) + \partial\delta_{\mathcal{S}}(y_0)$. Here $\delta_{\mathcal{S}}$, the convex indicator function for $\mathcal{S}$, has values 1 on $\mathcal{S}$, $+\infty$ otherwise, and $\partial F(y_0) = \{y^* \in \mathcal{Y}^*\colon F(y) - F(y_0) \geq y^*(y - y_0)\}$ defines the *subdifferential* of F at y_0 as a subset of linear functionals in the dual space $\mathcal{Y}^*$. The subdifferential represents one of the more successful efforts to extend the concept of differentiability. It is the basis of nonsmooth analysis, where the derivative of a function is a multifunction, a set-valued mapping. The usual linearity properties of derivatives have set-valued correspondents, and in some cases, integrals of multifunctions can be defined. But it is less evident how to differentiate such objects, and at present several candidates are being considered. What is significant is that nonsmooth extensions characterizing existence, necessary conditions, and sufficient conditions for minimization have been obtained for a large class of problems in the variational calculus. For example, the nonsmooth version of the Euler-Lagrange equation (13) for an absolutely continuous y_0 takes the form of a *differential inclusion*: There exists an absolutely continuous function p on $[a, b]$ such that for a.e. x, the pair $(p'(x),\ p(x))$ in $\mathbb{R}^2$ lies in the subgradient of $f(x, y, z)$ with respect to y and z at the point $(x, y_0(x), y_0'(x))$. Were f continuously differentiable, this subgradient would reduce to the ordinary partial derivatives of f at the point, so that $p = f_{y'}$, and its derivative is f_y, recapturing Eq. (13). In this way, differential inclusions provide nonsmooth extensions of differential equations. There are also nonsmooth approaches to minimizing a globally defined F restricted to a subset S, including use of a (nonsmooth) penalty function P whose value at y assigns a distance from y to S so large that minima for $F + P$ can only be found in S (where P vanishes).

VI. HAMILTONIAN CONTRIBUTIONS

The universe should operate efficiently. That conviction part philosophical, part religious, has been expressed in various terminology for centuries and the assessment of efficiency has received many formulations. For scientific purposes, the most fruitful has proved to be that of Hamilton (1834) which with some refinements has survived transition in modern physics to both relativity and quantum mechanics. In this model, we envisage the universe as a large but finite number N of particles with individual masses, positions $\mathbf{y}$ and velocities $\dot{\mathbf{y}} = \frac{d}{dt}\mathbf{y}$ giving rise to system kinetic energy T and potential energy V, at each time t. Then we postulate that between fixed times a and b at which positions are specified, the universe moves in a manner that makes stationary the *action integral*

$$F(y) = \int_a^b L(t, \mathbf{y}, \dot{\mathbf{y}})\, dt \tag{53}$$

expressed in terms of the Lagrangian

$$L = T - V. \tag{54}$$

Assuming that no further constraints are present, this would result in the E-L system (13)

$$\frac{d}{dt} L_{\dot{\mathbf{y}}} = L_{\mathbf{y}} \tag{55}$$

which are in fact Newton's equations of motion in the form already obtained by Lagrange (1762). (To permit certain constraints, Lagrange replaced the natural variables $\mathbf{y}$ by

so-called generalized coordinates $\mathbf{q}$ which are independent and meet these constraints.)

In order to attack the second-order system (55) Hamilton and his successors (notably Jacobi) effectively assumed that for fixed t and $\mathbf{y}$, the momentum vector system $\mathbf{p} = L_{\dot{\mathbf{y}}}(t, \mathbf{y}, \dot{\mathbf{y}})$ could be solved to express the variable(s) $\dot{\mathbf{y}}$ in terms of t, $\mathbf{y}$ and $\mathbf{p}$. Then through the chain rule, equations (55) become the first-order system in normal form

$$\dot{\mathbf{p}} = -H_{\mathbf{y}} \qquad \dot{\mathbf{y}} = H_{\mathbf{p}} \tag{56}$$

where the Hamiltonian H is defined by

$$H = H(t, \mathbf{y}, \mathbf{p} = \mathbf{p} \cdot \dot{\mathbf{y}} - L(t, \mathbf{y}, \dot{\mathbf{y}}) \tag{57}$$

assuming $\dot{\mathbf{y}} = \dot{\mathbf{y}}(t, \mathbf{y}, \mathbf{p})$ in the last equation. (These substitutions are now referred to collectively as the Legendre transformation.) In the classical case considered by Hamilton, L is independent of t, and $H = T + V$ is constant so that particles move on paths which preserve the total energy of the system. In general, Hamilton's equations (56) are analytically preferable to those of Lagrange (55). For example, their solutions are amenable to numerical approximation because they depend stably and uniquely on initial data.

Of course, the Legendre transformation can, in principle, be employed to replace the general E-L system (13) by an equivalent first order system in Hamiltonian form

$$\mathbf{p}' = -H_{\mathbf{y}} \qquad \mathbf{y}' = H_{\mathbf{p}} \tag{58}$$

where now $\mathbf{p} = f_{\mathbf{y}'}(x, \mathbf{y}, \mathbf{y}')$ is solved for $\mathbf{y}'$ and

$$H = H(x, \mathbf{y}, \mathbf{p}) = \mathbf{p} \cdot \mathbf{y}' - f(x, \mathbf{y}, \mathbf{y}') \tag{59}$$

in the resulting variables.

It is not difficult to show that for fixed x, f will be convex in $\mathbf{y}$ and $\mathbf{y}'$ in the sense of (24) precisely when H is convex in $\mathbf{p}$ for fixed $\mathbf{y}$ and $-H$ is convex in $\mathbf{y}$ for fixed $\mathbf{p}$. Such functions, among them $H(x, \mathbf{y}, \mathbf{p}) = |\mathbf{p}|^2 - |\mathbf{y}|^2$, are said to have saddle behavior.

Now, when f is just convex in $\mathbf{y}'$, then for fixed x, $\mathbf{y}$, $\mathbf{p}$, the $\mathbf{z}$ which makes $\mathbf{p} = f_{\mathbf{y}'}(x, \mathbf{y}, \mathbf{z})$ is that which maximizes $\mathbf{p} \cdot \mathbf{z} - f(x, \mathbf{y}, \mathbf{z})$. This suggests that when $f > -\infty$, the Hamiltonian should be given by

$$H(x, \mathbf{y}, \mathbf{p}) = \sup_{\mathbf{z}} [\mathbf{p} \cdot \mathbf{z} - f(x, \mathbf{y}, \mathbf{z})], \tag{60}$$

a definition that presupposes neither differentiability of f nor solvability of an auxiliary system. Note that $\mathbf{p} \cdot \mathbf{z}$ can be regarded as the value assigned at $\mathbf{z}$ to the linear functional $\mathbf{p}$ in the euclidean space considered as dual to that of $\mathbf{z}$. Building on earlier work of Fenchel (in euclidean spaces) and Moreau (in locally convex spaces), Rockafellar has used duality to study this Hamiltonian in a nonsmooth analytic setting where Hamiltonian inclusions replace Eqs. (58). He has also investigated the more general expression where the supremum is taken on $\mathbf{q} \cdot \mathbf{y} + \mathbf{p} \cdot \mathbf{z} - f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ for fixed $\mathbf{x}, \mathbf{p}, \mathbf{q}$, which he conjectures may be the true Hamiltonian for mathematical purposes.

VII. MULTIDIMENSIONAL INTEGRAL PROBLEMS

Heretofore we have considered minimization of a real-valued function F on a set of functions defined over an interval $[a, b]$. When this interval is replaced by $\bar{D}$, the closure of a bounded domain D of $\mathbb{R}^d$ for $d > 1$, the analysis is more complicated, but basic ingredients remain in evidence.

In particular, when ∇u denotes the gradient of the real-valued C^1-function $u = u(\mathbf{x})$, the appropriate E-L equation for a C^2-function $f = f(\mathbf{x}, u, \nabla u)$, is

$$\mathbf{div}\, f_{\nabla u} = f_u, \quad \mathbf{x} \in D. \tag{61}$$

(This equation governs each component u of a vector-valued $\mathbf{u}$.)

Only solutions u_0 of this partial differential equation could minimize the integral function

$$F(u) = \int_{\bar{D}} f(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x}))\, d\mathbf{x} \tag{62}$$

on the set

$$\mathcal{D} = \{u \in C^2(\bar{D}) \colon u|_S = \gamma\} \tag{63}$$

where γ is a given continuous real-valued function on S, the boundary of a fixed domain D. Here $d\mathbf{x}$ denotes the volume element in $\mathbb{R}^d$, and we suppose initially that S is so regular that the Riemann integral over $\bar{D} = D \cup S$ is defined and Green's Theorem holds.

Conversely, when f is convex in the variables u and ∇u, then each $u_0 \in \mathcal{D}$ that satisfies Eq. (61) must minimize F on $\mathcal{D}$. [However, in the multidimensional case, minimization does not imply a Weierstrass convexity in ∇u analogous to that in (21).]

A. Dirichlet's Principle

For example, $f(\mathbf{x}, u, \nabla u) = f(\nabla u) = |\nabla u|^2$ is convex in this sense and $f_{\nabla_u} = 2\nabla u$, so that each solution $u_0 \in \mathcal{D}$ of *Laplace's equation* $\mathbf{div}\,(\nabla u_0) = 0$ in D minimizes the Dirichlet integral

$$F(u) = \int_{\bar{D}} |\nabla u|^2\, d\mathbf{x} \tag{64}$$

on $\mathcal{D}$. Solutions of Laplace's equation are called *harmonic functions*. It is not difficult to show that in this case there

is at most one such minimizing function, but even here, it is far from evident that a minimizing harmonic function exists. Indeed, the assertion that it must (Dirichlet's Principle) applied uncritically by Riemann (1851), then challenged by Weierstrass, spurred the development of much subsequent analysis including the so-called direct methods of variational calculus pioneered by Hilbert in 1901. Moreover, on the disk $D = \{(r, \theta): r < 1\}$, the harmonic function $u_1(r, \theta) = \sum_{n=1}^{\infty} n^{-2} r^{n!} \cos(n!\theta)$ (in polar coordinates) extends continuously to $\bar{D}$, but its Dirichlet integral is infinite. Here, the highly irregular boundary values $\gamma = \gamma_1$ of u_1 determine a $\mathcal{D}$ of Eq. (63) that is empty, but for less pathological γ, Dirichlet's Principle has been established through functional analytic methods on much more general domains.

Again the basic procedure is to use convexity in ∇u and coercivity [(50) (ii) and (iii)] to prove existence of a minimizer u_0 in a Sobolev space of functions whose derivatives are in a set of Lebesgue integrable functions. Then one tries to show that u_0 is regular (smooth) enough to satisfy Laplace's equation in some sense. In fact, satisfaction in the distributional sense guarantees that u_0 is equal to a C^2 harmonic function almost everywhere (Weyl's Lemma).

Finding harmonic functions in D that extend continuously to assume prescribed values γ on ∂D is known as the *Dirichlet problem* for D. It is comparatively easy to show that there is at most one solution, and this boundary value problem is of independent interest because its solution u_0, if any, describes various physical situations. For example, if D is a domain in $\mathbb{R}^3$, then u_0 represents the steady-state temperature distribution within a homogenous medium (D) resulting from prescribed surface temperatures γ. However, as first noted by Lebesgue, we cannot expect a solution even on physical grounds: If γ is zero except on the surface of a sufficiently sharp thorn directed inward, then the temperature in D need not rise continuously to assume a prescribed maximum value of 100° at the tip of the thorn. For such domains, Dirichlet's Principle cannot hold—at least in its classical formulation. There is a "minimizing" harmonic function, but it does not assume the boundary values. The full study of related questions especially by Kellog and F. Riesz, resulted in the development of modern potential theory and its extension to solutions of other elliptic partial differential equations.

B. Plateau's Problem

Success in understanding, legitimizing, and extending Dirichlet's Principle encouraged examination of the related but much more difficult minimal surface problem. Although first considered by Lagrange in 1760, this problem is now attributed to Plateau, the Belgian physicist, whose mid-nineteenth century experiments helped define

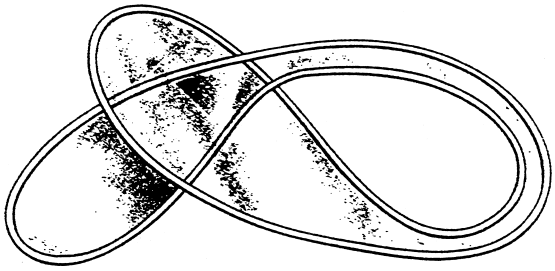


FIGURE 2 Soap film generated by a wire loop.

it. He conjectured that surface tension causes the actual shapes taken by soap films supported by wire loops withdrawn from a glycerine solution to be those that possess minimum area among competing surfaces with the same boundaries. However, Fig. 2 shows how difficult it can be to describe the possibly re-entrant competing surfaces in order to assign their areas, even for a single loop.

If we restrict attention to smooth surfaces that can be parametrized by a vector-valued mapping from a fixed planar region D, then their areas are given by an integral function similar to Eq. (62), and through Eq. (61) applied componentwise it can be shown that any smooth minimal surface must have zero mean curvature. Such surfaces can, in turn, be studied with the help of Dirichlet's Principle. Around 1931 this program led Douglas and Rado to produce independently the first major result: Among all continuous mappings of a closed base disk D into $\mathbb{R}^3$ whose boundary is mapped one-to-one onto a closed curve C (representing a single wire loop) there is one whose smooth image surface (the soap film) has minimum area. However, it was shown subsequently that some curves C admit spanning surfaces described by mappings from other base regions (disks with handles) that have less area. Moreover actual soap films which need not touch the entire bounding curve C could obviously have less area. Finally, multiple bounding curves C representing separate wire loops further complicate the descriptive picture, and their soap films can have smooth surface pieces that do not terminate on C. Indeed, the parallel circular rings of Fig. 3a can generate the soap film indicated having the shaded internal disk with a singular bounding circle in addition to those of the smooth surface of revolution and the pair of shaded disjoint circular disks of Fig. 3b and 3c.

What is needed is mathematics that can describe and assign size to such objects. For this purpose, Weierstrass introduced so-called *parametric* integrals whose values like those in Eq. (3) are unaffected by reparametrization and considered their minimization in his lectures. To handle other anomalies, several tools, both analytic and topological, were developed during the past century including

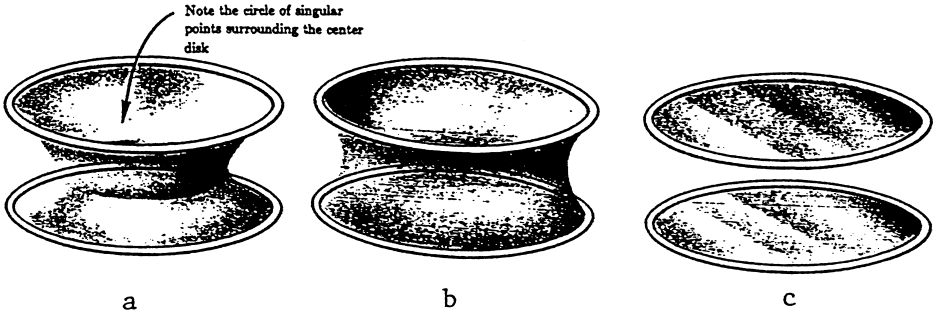


FIGURE 3 Soap films generated by a pair of rings.

Hausdorff measure of m-dimensional sets in higher dimensional space, varifolds, and integral currents. These provide the principle components of geometric measure theory formulated around 1960 by Almgren, Federer, and Fleming among others, and used to attack other extremal problems in higher dimensions that require intrinsic description. In particular, these methods have given new impetus to the study of minimal surface problems in $\mathbb{R}^d$ for $d > 3$ that was initiated in Douglas' work of 1931. Moreover, in $\mathbb{R}^3$ a few significant results have been obtained recently concerning surfaces with constant nonzero mean curvature that can represent closed soap bubbles.

Despite these efforts, many questions remain unanswered, especially in regard to multiply connected regions. In particular, in $\mathbb{R}^3$, it is still not known which surfaces of zero mean curvature (unfortunately referred to as *minimal surfaces*) have minimal area—at least among nearby competitors—thereby possessing the stability required of actual soap films. In certain cases, mathematical existence of unstable surfaces can be established, but just how many different surfaces are permitted by some boundary loop configurations and which of these will have singular sets is yet to be determined. However, the methods used to attack these questions have found other physical applications that range from characterizing types of crystalline structures in physical chemistry to establishing existence of black holes in relativistic analysis.

VIII. VARIATIONAL CALCULUS IN THE LARGE

Concern with unstable surfaces represents a modern trend in variational calculus, namely, consideration of *all* critical points of an integral function F, not just those that give local extremals. At each such point the Fréchet derivative F' vanishes, and just as in the case of ordinary functions of several variables, F can have directions of local maximal behavior as well as those of local minimal behavior, distinguished by signs of the second variations in these directions [recall Eq. (25)]. Thus, we are led to think of the graph of F as a "surface" lying over its domain $\mathcal{D}$ of functions. (Fig. 4).

Viewed in this light, we see that between a pair of isolated local minimum points there should be a region of $\mathcal{D}$ over which the graph of F resembles a range of mountains. Across this range, there should be at least one route connecting these points with lowest high point (a mountain pass), at which there are directions down as well as directions up the mountain. But this would correspond to a critical saddle point of F that is not a local extremal, and with care, this suggestive "mountain pass" terminology can be converted into a mathematically rigorous existence argument. It was first used successfully around 1940 by Courant to establish existence of unstable minimal surfaces, and it has since become an important technique in variational considerations.

However, the most remarkable consequence of regarding F over $\mathcal{D}$ as a whole, is the recognition that a deep relationship exists between the algebraic topological character of a space $\mathcal{D}$ and the variational functions F defined on it. Although first used by Birkhoff in 1917, this relationship was set forth around 1930 in works of Morse in America and, from a different viewpoint, Liusternik and Schnirelman in Russia. As a fundamental principle of nonlinear functional analysis, it now supplies an important source of results and insight in both variational calculus and nonlinear partial differential equations.

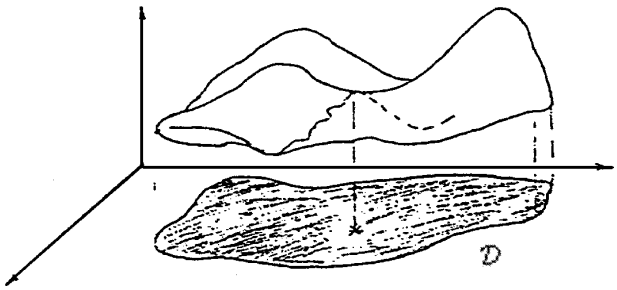


FIGURE 4 A mountain pass.

In its simplest form, Morse theory presupposes that F has only isolated non-degenerate critical points, and establishes a lower bound for their number based on numbers of homology classes of $\mathcal{D}$. This is achieved through recognition that the homology of the sublevel sets of F changes only at a critical point, and then only in a manner commensurate with the character of the critical point as reflected in that of the quadratic form of its second variation. In principle, Morse's theory even offers lower estimates of each type of critical point, but skill is required to assign a topology to the underlying function space $\mathcal{D}$ that has usable homology. Nevertheless, this theory was employed to establish existence of unstable minimal surfaces in 1939, and it has been extended to infinite dimensional Hilbert manifolds by Milnor, Palais, and Smale, and recently even to more general structures.

A simple version of Liusternik-Schnirelman theory provides lower estimates for the size of the critical set of a differentiable even function F (one for which $F(-u) = F(u)$) restricted to the surface $\mathcal{S}$ of the unit ball in a Banach space. In fact, the associated critical values of F at such points can be assessed through maximin estimates similar to those in Section III.B involving values of F on families of sets dependent on the topology of $\mathcal{S}$. This approach was used to prove existence of three non-self-intersecting closed geodesics (surface curves of stationary length) on any closed surface in $\mathbb{R}^3$ that smoothly resembles the surface of a sphere. Because this theory does not require assumptions about the critical set, it has had wide extensions and applications.

IX. CONCLUSION

Variational calculus supplies the analytic bridge linking ancient conjectures concerning an ideal universe with modern demands for optimal control of operating systems. It was instrumental in formulating variational principles of mechanics and physics and continues to supply insight into the relationship between these principles and their Euler-Lagrange systems of differential equations. Finally, it furnishes a wealth of nonlinear problems, many of physical origin, that appear simple but have consequences that are not. These problems have provided a motivating force in the creation of functional analysis, both linear and nonlinear as well as geometric measure theory.

SEE ALSO THE FOLLOWING ARTICLES

CALCULUS • STOCHASTIC PROCESSES

BIBLIOGRAPHY

Clarke, F. (1983). "Optimization and Nonsmooth Analysis," John Wiley and Sons, New York.

Courant, R. (1950). "Dirichlet's Principle, Conformal Mapping, and Minimal Surfaces," Interscience Publishers Inc., New York.

Federer, H. (1969). "Geometric Measure Theory," Springer-Verlag, New York.

Fleming, W., and Rishel, R. (1975). "Deterministic and Stochastic Optimal Control," Springer-Verlag, New York.

Goldstine, H. H. (1980). "A History of Calculus of Variations from the 17th through the 19th Century," Springer-Verlag, New York.

Hildebrandt, S., and Tromba, A. (1996). "The Parsimonius Universe: Shape and Form in the Natural World," Springer-Verlag Inc., New York.

Monna, A. F. (1975). "Dirichlet's Principle: A Mathematical Comedy of Errors and Its Influence on the Development of Analysis," Oosthoek, Scheltema and Holkema, Utrecht.

Osserman, R. (1986). "A Survey of Minimal Surfaces," Dover Publications, New York.

Struwe, M. (1988). "Plateau's Problem and the Calculus of Variations," Mathematical Notes, Princeton Univ. Press, New Jersey.

Tikhomirov, V. (1986). "Fundamental Principles of the Theory of Extremal Problems," John Wiley and Sons, New York.

Troutman, J. L. (1996). "Variational Calculus and Optimal Control: Optimization with Elementary Convexity," Second Edition, Springer-Verlag, New York.



Wavelets, Advanced

Su-Yun Huang
Academia Sinica, Taiwan

Zhidong Bai
National University of Singapore

I. Introduction: Why Wavelets?
II. Wavelets and Continuous Wavelet Transforms
III. Multiresolution Analysis
IV. Discrete Wavelet Transforms
V. Filters and Quadrature Filters
VI. Some Examples of Wavelets
VII. Wavelet Packets
VIII. Best Basis Selection in Wavelet Packet Analysis
IX. Wavelets in Statistical Estimation
X. Two-Dimensional Wavelet Algorithm and Image Processing

GLOSSARY

Cascade algorithm This algorithm gives a constructive and efficient way to compute approximate values for the scaling function with arbitrarily high precision.

Discrete wavelet transform The discrete wavelet transform calculates the coefficients of the wavelet series approximation for a discrete signal.

Multiresolution analysis A multiresolution analysis of $L_2(R)$ consists of a scaling function and a nested sequence of closed subspaces for approximating $L_2(R)$ functions. Subspaces in this nested sequence correspond to different resolution levels from coarser to finer scales.

Thresholding Many of the coefficients of a wavelet series are often close to or equal to zero. The thresholding method sets these small wavelet coefficients to zero so that the sequence of wavelet coefficients contains long strings of zeros. It can be used for data compression or for noise removal. There are some different thresholding rules in wavelet shrinkage motivated from statistical theory.

Wavelet packets Wavelet packet functions consist of a rich family of building block functions. Mixtures of orthogonal quadrature filters may be applied to scaling functions and wavelets to get new building block functions. Families of new orthonormal basis functions so produced are called wavelet packets. Wavelet packet functions are more flexible than wavelets in representing different types of signals.

Wavelets Wavelets are waveform-like functions with fast tail decay and are oscillatory and localized. They comprise the family of translations and dilations of a single function. They are versatile building blocks for representing functions or signals in various functional classes.

Wavelet shrinkage A methodology for denoising by shrinking the empirical wavelet coefficients toward zero.

Wavelet shrinkage—Bayesian approach A methodology for denoising by combining the wavelet shrinkage and Bayesian formulation.

I. INTRODUCTION: WHY WAVELETS?

Wavelets and multiresolution analysis are a very popular subject in mathematical science and engineering. Wavelets are waveform-like functions with fast tail decay, and they are oscillatory and localized. They comprise the family of translations and dilations of a single function. They are versatile building blocks for representing functions or signals in various functional classes.

The wavelet transform is the set of inner products of all dilated and translated wavelets with an analyzed function. The range of this linear map is characterized by a reproducing kernel Hilbert space, and this linear map is isometric. The wavelet transform is a time–frequency analysis tool that decomposes the analysed function into different frequency components. Each frequency component is matched with a resolution scale. The wavelet transform of a signal depends on two variables, frequency and time (or scale and space). It provides localized time–frequency information of the analysed function.

To study the spectral behavior of a function or a signal, the classical Fourier transform requires full knowledge of the signal in the time domain. A signal change in a small interval of time will affect the entire spectral behavior. In contrast, wavelets can provide good localization in both time and frequency domains. The short-time Fourier transform can also correct the deficiency of the Fourier transform by placing a window function around a certain time point of interest before taking the Fourier transform, and it has the effect of restricting the spectral information of the signal to the domain of influence of the window function. However, the short-time Fourier transform cuts up the signal (in the time domain) into slices of fixed length determined by the time window width. Slices of the same length in time domain are used to resolve all frequency components high or low, which results in redundant information for high-frequency components and inadequate information for low-frequency components. However, the wavelet transform provides ideal and adaptive time–frequency localization. It cuts up the signal with different time window widths to adapt to different frequency components. Lower frequency components are analyzed through a time interval of wider width and higher frequency components are analyzed through a time interval of shorter width.

An arbitrary function in $L_2(R)$ can be represented by wavelet series. The computational aspects of wavelet series are very attractive with fast algorithm and low complexity. Wavelet coefficients of lower scales can be obtained from coefficients of higher scales using the fast algorithm. In the reverse direction, coefficients of higher scales can be recovered from coefficients of lower scales, again using the fast algorithm.

In all, the wavelet analysis has good time–frequency localization and good adaptivity to local features of a signal. These advantages together with the fast algorithm make wavelets a very popular analysis tool for a wide range of applications, including signal and image processing, data compression, statistical estimation, and denoising.

II. WAVELETS AND CONTINUOUS WAVELET TRANSFORMS

Let $\{\psi_{a,b}(x)\}$, with $a, b \in R$ and $a \neq 0$, be a family of functions obtained from a single function $\psi(x) \in L_2(R)$ by dilation and translation:

$$\psi_{a,b}(x) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{x-b}{a}\right)$$

Further, ψ satisfies conditions $\int_{-\infty}^{\infty}\psi(x)\,dx = 0$ and $\int_{-\infty}^{\infty}\psi^2(x)\,dx = 1$ and also the admissibility condition

$$C_\psi = 2\pi\int_{-\infty}^{\infty}\frac{|\hat{\psi}(\omega)|^2}{|\omega|}\,d\omega < \infty$$

where

$$\hat{\psi}(\omega) = (\sqrt{2\pi})^{-1}\int_{-\infty}^{\infty}\psi(x)e^{-i\omega x}\,dx$$

is the Fourier transform of $\psi(x)$. The function $\psi(x)$ is called the mother wavelet function and $\psi_{a,b}(x)$ are called wavelets. The name "wavelet" comes from the fact that $\psi(x)$ is well localized and, as $\int_{-\infty}^{\infty}\psi(x)\,dx = 0$, its shape resembles small waves. The localization of wavelets can be made arbitrarily fine by appropriate rescaling.

For any function $f \in L_2(R)$, the continuous wavelet transform is defined as a bivariate function

$$\mathcal{T}f(a,b) = \int_{-\infty}^{\infty} f(x)\overline{\psi_{a,b}(x)}\,dx$$

The value of $\mathcal{T}f(a,b)$ is called the scale-space wavelet coefficient. Notice that $|\mathcal{T}f(a,b)| \leq \|f\|_2$.

A. Time–Frequency Analysis

The function $f(x)$ is analyzed by the continuous wavelet transform based on dilations and translations of one single wavelet function $\psi(x)$. The continuous wavelet transform can serve as a window function extracting local information around certain time-frequency span. We say a function

$w(x) \in L_2(R)$ is a window function if $xw(x)$ is in $L_2(R)$. The center c_w and the radius Δ_w of a window function $w(x)$ are defined by

$$c_w = \frac{1}{\|w\|_2^2} \int_{-\infty}^{\infty} x|w(x)|^2\,dx$$

and

$$\Delta_w = \frac{1}{\|w\|_2^2} \left\{ \int_{-\infty}^{\infty} (x - c_w)^2 |w(x)|^2\,dx \right\}^{1/2}$$

Suppose that $\psi(x)$ and $\hat{\psi}(\omega)$ are window functions. Denote their centers and radii by c_ψ, Δ_ψ, and $c_{\hat{\psi}}$, $\Delta_{\hat{\psi}}$, respectively.

The continuous wavelet transform by $\psi_{a,b}(x)$ localizes the signal $f(x)$ with a "time window" centered at $ac_\psi + b$ and with width $2a\Delta_\psi$. On the other hand, $\hat{\psi}_{a,b}(\omega)$ can serve as a window function, which localizes the spectral information of $\hat{f}(\omega)$ with a "frequency window" centered at $c_{\hat{\psi}}/a$ and with width $2\Delta_{\hat{\psi}}/a$. Notice that, by the Parseval identity, the continuous wavelet transform can be expressed as

$$\begin{aligned} \mathcal{T}_\psi f(a,b) &= \int_{-\infty}^{\infty} f(x)\overline{\psi_{a,b}(x)}\,dx \\ &= \int_{-\infty}^{\infty} \hat{f}(\omega)\overline{\hat{\psi}_{a,b}(\omega)}\,d\omega \end{aligned}$$

Notation $\mathcal{T}_\psi$ is used to indicate that the transform is wavelet dependent. For simplicity, we may suppress the subscript and use notation $\mathcal{T}$ instead, if no confusion. The continuous wavelet transform $\mathcal{T}f(a,b)$ localizes both the time and frequency information of the signal f. The window widths for time and frequency are $2a\Delta_\psi$ and $2\Delta_{\hat{\psi}}/a$, respectively. That is, the continuous wavelet transform $\mathcal{T}f(a,b)$ can serve as a time–frequency window with shorter time-window width for higher frequency and wider time-window width for lower frequency.

B. Regularity

A wavelet function $\psi(x)$ is called r-regular ($r \in N$) if it satisfies the condition

$$\left|\psi^{(\alpha)}(x)\right| \le C_m(1 + |x|)^{-m}$$

for each integer $m \in N$ and for every $\alpha = 0, 1, \ldots, r$.

C. Localization

In order to have good time-domain resolution, it is necessary that wavelets be localized in time. Such time localization may be quantified by the index γ:

$$|\psi(x)| \le c(1 + |x|^2)^{-\gamma/2}, \quad c > 0,\ \gamma > 1$$

The larger γ is, the better the localization is. In the same way the frequency localization can be quantified by the index $\tilde{\gamma}$:

$$|\hat{\psi}(\omega)| \le c(1 + |\omega|^2)^{-\tilde{\gamma}/2}, \quad c > 0,\ \tilde{\gamma} > 0$$

Again, the larger $\tilde{\gamma}$ is, the better the frequency localization is.

D. Vanishing Moments (or Oscillation Condition)

The wavelet function has at least the 0th degree moment vanished. This phenomenon can be extended to require that some more consecutive moments vanish

$$\int_{-\infty}^{\infty} \psi(x)x^k\,dx = 0, \quad k = 0, 1, \ldots, \ell \tag{1}$$

assuming that $\psi(x)$ is sufficiently localized. We say that the wavelet $\psi(x)$ has vanishing moments of order m (or of degree $m-1$) if and only if

$$\int_{-\infty}^{\infty} \psi(x)x^k\,dx = 0, \quad k = 0, \ldots, m-1,$$

and

$$\int_{-\infty}^{\infty} \psi(x)x^m\,dx \neq 0$$

Condition (1) is equivalent to

$$\hat{\psi}(\omega) = o(\omega^\ell), \quad (\omega \to 0)$$

whenever $\psi(x) \in L_1(R)$ and $x^\ell\psi(x) \in L_1(R)$. Notice that an r-regular wavelet function has at least up to rth degree vanishing moments. Usually the degree of vanishing moments is much higher than the degree of regularity.

E. Resolution of Identity

If $f, g \in L_2(R)$, then the following resolution of the identity formula holds:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{T}f(a,b)\overline{\mathcal{T}g(a,b)}\frac{1}{a^2}\,da\,db = C_\psi\langle f, g\rangle$$

where ψ satisfies the admissibility condition. A function can be reconstructed from its wavelet transform:

$$f(x) = \frac{1}{C_\psi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{T}f(a,b)\psi_{a,b}(x)\frac{1}{a^2}\,da\,db$$

by means of the resolution of the identity. The preceding equality is meant in the L_2-sense, not in the pointwise sense. Furthermore, by taking $f = g$ in the resolution of the identity formula, we obtain

$$C_\psi^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|\mathcal{T}f(a,b)|^2a^{-2}\,da\,db=\int_{-\infty}^{\infty}|f(x)|^2\,dx$$

That is, the continuous wavelet transform $\mathcal{T}_\psi$ maps $L_2(R,dx)$ isometrically into $L_2(R^2,a^{-2}C_\psi^{-1}\,da\,db)$.

F. The Associated Reproducing Kernel Hilbert Spaces

Denote the set of images of the wavelet transform $\{\mathcal{T}f: f\in L_2(R)\}$ by $\mathcal{H}$, which is a subspace of $L_2(R^2,a^{-2}C_\psi^{-1}\,da\,db)$. Then $\mathcal{H}$ is a reproducing kernel Hilbert space with the reproducing kernel

$$\mathcal{K}(u,v;a,b)=\int_{-\infty}^{\infty}\psi_{u,v}(x)\overline{\psi_{a,b}(x)}\,dx$$

and the following identity holds

$$C_\psi^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathcal{K}(u,v;a,b)\mathcal{T}f(u,v)u^{-2}\,du\,dv=\mathcal{T}f(a,b)$$

G. Shifting and Scaling Properties

Let $g(x)=f(x-x_0)$ and $h(x)=\sqrt{s}f(sx)$, then

$$\mathcal{T}g(a,b)=\mathcal{T}f(a,b-x_0)\text{ and }\mathcal{T}h(a,b)=\mathcal{T}f(sa,sb)$$

H. Function Regularity and the Order of Magnitude of Wavelet Coefficients

As we will see, a local regularity of the analyzed function implies a local decrease of the scale-space wavelet coefficients at fine resolutions.

We say that a funcation f is of local Hölder regularity α at x_0, if and only if f can be written as $f(x+x_0)=P_n(x)+\delta(x)$, where $n<\alpha\le n+1$, P_n is a polynomial of degree at most n and $\delta(x)=O(|x|^\alpha)$ as $x\to 0$. Suppose that the wavelet $\psi(x)$ has vanishing moments of order at least $n+1$ and that $x^\alpha\psi(x)\in L_1(R)$. If f is of local Hölder regularity of degree α at a point x_0, then

$$\mathcal{T}f(a,b+x_0)=O(|a|^{\alpha+1/2}+|a|^{1/2}|b|^\alpha),\quad a\neq 0 \tag{2}$$

as $a,b\to 0$. The asymptotic behavior (2) of $\mathcal{T}f(a,b+x_0)$ for a and b approaching zero implies that, for every fixed (a,b),

$$\mathcal{T}f(\lambda a,\lambda b+x_0)=O(|\lambda|^{\alpha+1/2})$$

as $\lambda\to 0$. In other words, the local Hölder regularity of degree α for f at a point x_0 is mirrored by a decrease of order $|\lambda|^{\alpha+1/2}$ along every straight line $\{(\lambda a,\lambda b+x_0):\lambda\in R\}$ in the plane.

III. MULTIRESOLUTION ANALYSIS

Wavelets can be used as an analysis tool to describe mathematically the increment in information needed to go from a coarser approximation to a higher resolution approximation. This insight was put into the framework of multiresolution analysis by Mallat (1989).

A multiresolution analysis of $L_2(R)$ consists of a nested sequence of closed subspaces $\cdots V_{j-1}\subset V_j\subset V_{j+1}\cdots$ for approximating $L_2(R)$ functions. The multiresolution analysis satisfies the following conditions:

1. $\bigcup_{j=-\infty}^{\infty}V_j$ is dense in $L_2(R)$.
2. $\bigcap_{j=-\infty}^{\infty}V_j=\{0\}$.
3. $f(x)\in V_j$ if and only if $f(2x)\in V_{j+1},\ \forall j\in Z$.
4. There exists a function $f(x)$ such that $\{f(x-k):k\in Z\}$ forms an unconditional basis (also known as Riesz basis) for V_0.

Notice that $\{f(2^jx-k):k\in Z\}$ is an unconditional basis for V_j.

Given this translation-invariant unconditional basis, we are seeking whether a translation-invariant orthonormal basis exists. The answer is affirmative. There exists a function $\phi(x)$ such that $\int_{-\infty}^{\infty}\phi(x)\,dx=1$ and that $\{\phi(x-k)\}_{k\in Z}$ forms an orthonormal basis for V_0. The function $\phi(x)$ is called a wavelet scaling function or the father wavelet.

The following is a brief description of construction of $\phi(x)$ from the unconditional basis $\{f(x-k)\}_{k\in Z}$ based on the Fourier method.

The family of integral translates of one single function $\phi^{(o.n.)}(x)\in L_1(R)\cap L_2(R)$ is an orthonormal set if and only if

$$\sum_{k\in Z}\left|\hat{\phi}^{(o.n.)}(\omega+2\pi k)\right|^2=\frac{1}{2\pi} \tag{3}$$

for all $\omega\in R$. Given the unconditional basis $\{f(x-k):k\in Z\}$, let

$$M(\omega)=\left(2\pi\sum_{k=-\infty}^{\infty}|\hat{f}(\omega+2k\pi)|^2\right)^{-1/2}$$

and let $\phi(x)$ be given by

$$\hat{\phi}(\omega)=M(\omega)\hat{f}(\omega)$$

It can be easily checked that $\phi(x)$ satisfies (3). Thus $\{\phi(x-k)\}_{k\in Z}$ forms an orthonormal basis for V_0. Notice that $\{\phi_{j,k}(x)=\sqrt{2^j}\phi(2^jx-k):k\in Z\}$ is then an orthonormal basis for V_j. Since $V_0\subset V_1$, the scaling function $\phi(x)$ can be represented as a linear combination of functions $\phi_{1,k}(x)$:

$$\phi(x) = \sum_{k\in Z} h_k \phi_{1,k}(x) \tag{4}$$

for some coefficients $h_k, k \in Z$. The preceding equation is called the two-scale equation or the scaling equation. The sequence of coefficients $\{h_k : k \in Z\}$ has the properties

$$\sum_{k\in Z} h_k = \sqrt{2} \qquad \text{and} \qquad \sum_{k\in Z} h_k^2 = 1$$

The two-scale equation (4) can be rewritten in terms of Fourier transform as

$$\hat{\varphi}(\omega) = \sum_{k\in Z} h_k \hat{\varphi}_{1,k}(\omega) = m_0(\omega/2)\hat{\varphi}(\omega/2) \tag{5}$$

where

$$m_0(\omega) = \frac{1}{\sqrt{2}} \sum_{k\in Z} h_k e^{-ik\omega}$$

Orthogonality of these $\phi(x-k)$ implies that

$$|m_0(\omega)|^2 + |m_0(\omega+\pi)|^2 = 1$$

By iterating (5), we have

$$\hat{\varphi}(\omega) = \hat{\phi}(0) \prod_{j=1}^{\infty} m_0\left(\frac{\omega}{2^j}\right) = \frac{1}{\sqrt{2\pi}} \prod_{j=1}^{\infty} m_0\left(\frac{\omega}{2^j}\right)$$

Define the orthogonal complementary subspaces $W_j = V_{j+1} - V_j$ so that $V_{j+1} = V_j \oplus W_j$, where $\oplus$ is an orthogonal direct sum. Then the space $L_2(R)$ can be decomposed as

$$L_2(R) = \bigoplus_{j\in Z} W_j$$

For an arbitrary multiresolution analysis of $L_2(R)$, there exists a function $\psi(x)$ such that $\{\psi_{j,k}(x) = \sqrt{2^j}\,\psi(2^j x - k) : k \in Z\}$ forms an orthonormal basis of the difference space W_j. Thus, $\{\psi_{j,k}(x) : j, k \in Z\}$ forms an orthonormal basis for $L_2(R)$. The function $\psi(x)$ is called a wavelet function or the mother wavelet. Since $\psi(x) \in W_0 \subset V_1$, it can be represented as

$$\psi(x) = \sum_{k\in Z} g_k\, \phi_{1,k}(x) \tag{6}$$

for some coefficients g_k, $k \in Z$. One possibility for the choice of these coefficients may be set to relate to h_k by

$$g_k = (-1)^k \bar{h}_{1-k}$$

That is, essentially only one two-scale sequence h governs the multiresolution analysis and its wavelet decomposition. The two-scale equation (6) can be rewritten in terms of Fourier transform as

$$\hat{\psi}(\omega) = m_1(\omega/2)\hat{\varphi}(\omega/2) \tag{7}$$

where

$$m_1(\omega) = \frac{1}{\sqrt{2}} \sum_{k\in Z} g_k e^{-ik\omega} = -e^{-i\omega}\,\overline{m_0(\omega+\pi)}$$

The sequence of coefficients $\{g_k : k \in Z\}$ has the properties

$$\sum_{k\in Z} g_k = 0 \qquad \text{and} \qquad \sum_{k\in Z} g_k^2 = 1$$

A multiresolution analysis $\{V_j : j \in Z\}$ of $L_2(R)$ is called r-regular, if the scaling function $\phi(x)$ (or the wavelet $\psi(x)$) can be so chosen to satisfy the condition

$$\left|\phi^{(\alpha)}(x)\right| \le C_m (1+|x|)^{-m}$$

for each integer $m \in N$ and for every $\alpha = 0, 1, \ldots, r$.

For the rest of this section, the multiresolution analysis $\{V_j : j \in Z\}$ is assumed r-regular, unless otherwise specified.

Let

$$\mathcal{V}(x,y) = \sum_{k\in Z} \phi(x-k)\overline{\phi(y-k)}$$

$$\mathcal{V}_j(x,y) = \sum_{k\in Z} 2^j \phi(2^j x - k)\overline{\phi(2^j y - k)}$$

$$\mathcal{W}(x,y) = \sum_{k\in Z} \psi(x-k)\overline{\psi(y-k)}$$

$$\mathcal{W}_j(x,y) = \sum_{k\in Z} 2^j \psi(2^j x - k)\overline{\psi(2^j y - k)}$$

By the r-regularity property, we immediately deduce that

$$\left|\partial_x^\alpha \partial_y^\beta\, \mathcal{V}(x,y)\right| \le C_m (1+|x-y|)^{-m}$$

for each integer $m \in N$ and for every $\alpha = 0, 1, \ldots, r$, and $\beta = 0, 1, \ldots, r$.

The wavelet subspaces V_j and W_j are reproducing kernel Hilbert spaces with the reproducing kernels $\mathcal{V}_j(x,y)$ and $\mathcal{W}_j(x,y)$, respectively. The orthogonal projection of $f \in L_2(R)$ onto the subspaces V_j and W_j can be written respectively as

$$\mathcal{P}_{V_j} f(x) = \int_{-\infty}^{\infty} \mathcal{V}_j(x,y) f(y)\, dy$$

and

$$\mathcal{P}_{W_j} f(x) = \int_{-\infty}^{\infty} \mathcal{W}_j(x,y) f(y)\, dy$$

For an r-regular multiresolution analysis, we have

$$\int_{-\infty}^{\infty} \mathcal{V}_j(x,y) y^k\, dy = x^k$$

and

$$\int_{-\infty}^{\infty} \mathcal{W}_j(x,y) y^k\, dy = 0$$

for every $j \in Z$ and $k = 0, 1, \ldots, r$.

A kernel $\mathcal{V}(x,y)$ is said to be of order m if and only if it satisfies the moment conditions

$$\int_{-\infty}^{\infty} \mathcal{V}(x,y) y^k\, dy = \begin{cases} 1, & k = 0 \\ x^k, & k = 1, \ldots, m-1 \\ \alpha(x) \ne x^m, & k = m \end{cases}$$

A projection kernel $\mathcal{V}(x, y)$ arising from a multiresolution analysis is of order m if and only if the associated wavelet $\psi(x)$ has vanishing moments of order m. Usually the order of vanishing moments is much higher than the degree of regularity.

For a multiresolution analysis, one may be interested in knowing the wellness of the projection approximation $\mathcal{P}_{V_j} f$ to f. The answer depends on the regularity of f and the regularity of functions in V_j. The higher these regularities are, the better the projection approximation is.

Theorem 1 *Suppose that the multiresolution analysis has regularity r. Then $f(x)$ is in the Sobolev space $W_2^s(R)$, $0 \le s \le r$, if and only if*

$$\limsup_{j\to\infty} 2^{js} \left\| f - \mathcal{P}_{V_j} f \right\|_2 < \infty$$

This theorem reveals the approximation order of the projection $\mathcal{P}_{V_j}$ in L_2-norm. It also characterizes the function space of $W_2^s(R)$.

To look further into the pointwise behavior of the projection $\mathcal{P}_{V_j}$, assume that $f(x)$ belongs to $C^{m,\alpha}(R) \cap L_2(R)$ for some $0 < \alpha \le 1$, where

$$C^{m,\alpha}(R) = \left\{ f \in C^m(R) : \left| f^{(m)}(x) - f^{(m)}(y) \right| \le A|x - y|^\alpha,\ A > 0 \right\}.$$

Suppose that $\mathcal{V}(x, y)$ is of order m. Define

$$b_m(x) = x^m - \int_{-\infty}^{\infty} \mathcal{V}(x, y) y^m \, dy$$

Notice that the function $b_m(x)$ is a continuous periodic function with period one. Assume that the kernel $\mathcal{V}(x, y)$, or $\phi(x)$ or $\psi(x)$, is sufficiently localized so that

$$\int_{-\infty}^{\infty} |\mathcal{V}(x, y)(y - x)^{m+\alpha}| \, dy < \infty$$

Then

$$f(x) - \mathcal{P}_{V_j} f(x) = \frac{1}{2^{jm}\, m!} f^{(m)}(x) b_m(2^j x) + O\left(2^{-j(m+\alpha)}\right) \tag{8}$$

A multiresolution analysis $\{V_j : j \in Z\}$ is said to be symmetric if the projection kernel satisfies the condition $\mathcal{V}(-x, y) = \mathcal{V}(x, -y)$. Notice that $\mathcal{V}(-x, y)$ is a time-reversed kernel, and $\int_{-\infty}^{\infty} \mathcal{V}(x, -y) f(y) \, dy$ corresponds to the transform of time-reversed signal f. If the multiresolution analysis is symmetric, then $b_m(-x) = (-1)^m b_m(x)$. When m is even, $b_m(x)$ is symmetric about zero. Since $b_m(x)$ is periodic with period one, $b_m(x)$ is also symmetric about all the points $x = k/2, k \in Z$. When m is odd $b_m(x)$ is antisymmetric about zero and hence antisymmetric about all the points $x = k/2, k \in Z$.

IV. DISCRETE WAVELET TRANSFORMS

In the continuous wavelet transform, we consider the family of wavelets

$$\psi_{a,b}(x) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{x - b}{a}\right)$$

where $a, b \in R$, $a \ne 0$ and $\psi(x)$ is admissible. In the multiresolution analysis, there exist such functions $\phi(x)$ and $\psi(x)$ that $\{\phi_{j,k}(x) : k \in Z\}$ constitutes an orthonormal basis for V_j and $\{\psi_{j,k}(x) : k \in Z\}$ constitutes an orthonormal basis for W_j. Thus, sampling information of $f(x)$ on the lattice points,

$$a = 2^j, \quad b \in 2^j Z \quad \text{for } j \in Z$$

is enough to gain full knowledge of f. In the discretization, we restrict a and b to the above lattice points. For every $f(x) \in L_2(R)$, $f(x)$ can be represented in terms of orthonormal wavelet series

$$f(x) = \sum_{k\in Z} c_{j,k} \phi_{j,k}(x) + \sum_{\ell \ge j} \sum_{k\in Z} d_{\ell,k} \psi_{\ell,k}(x)$$

or

$$f(x) = \sum_{\ell\in Z} \sum_{k\in Z} d_{\ell,k} \psi_{\ell,k}(x)$$

where

$$c_{j,k} = \langle f, \phi_{j,k} \rangle, \qquad d_{\ell,k} = \langle f, \psi_{\ell,k} \rangle$$

with $\langle \cdot, \cdot \rangle$ the inner product in L_2. By the two-scale equations (4) and (6), we have

$$\phi_{j-1,k}(x) = \sum_{\ell\in Z} h_{\ell-2k} \phi_{j,\ell}(x) \tag{9}$$

and

$$\psi_{j-1,k}(x) = \sum_{\ell\in Z} g_{\ell-2k} \phi_{j,\ell}(x) \tag{10}$$

By (9), (10), and the orthogonality property of wavelets, we obtain

$$c_{j-1,k} = \langle f, \phi_{j-1,k} \rangle = \left\langle f, \sum_{\ell\in Z} h_{\ell-2k} \phi_{j,\ell} \right\rangle = \sum_{\ell\in Z} h_{\ell-2k} c_{j,\ell}$$

Similarly,

$$d_{j-1,k} = \langle f, \psi_{j-1,k} \rangle = \left\langle f, \sum_{\ell\in Z} g_{\ell-2k}\ \phi_{j,\ell} \right\rangle = \sum_{\ell\in Z} g_{\ell-2k} c_{j,\ell}$$

In the reverse direction, coefficients in the finer resolution scale can be obtained from coefficients in the coarser resolution scale:

$$c_{j,k} = \langle f, \phi_{j,k}\rangle = \sum_{\ell} c_{j-1,\ell}\langle \phi_{j-1,\ell}, \phi_{j,k}\rangle + \sum_{\ell} d_{j-1,\ell}\langle \psi_{j-1,\ell}, \phi_{j,k}\rangle = \sum_{\ell} c_{j-1,\ell} h_{k-2\ell} + \sum_{\ell} d_{j-1,\ell} g_{k-2\ell}$$

In summary, we have the following fast wavelet algorithm.

Wavelet decomposition: $\begin{cases} c_{j-1,k} = \sum_{\ell\in Z} h_{\ell-2k} c_{j,\ell}, \\ d_{j-1,k} = \sum_{\ell\in Z} g_{\ell-2k} c_{j,\ell}, \end{cases}$

Wavelet reconstruction:

$$c_{j,k} = \sum_{\ell\in Z} h_{k-2\ell} c_{j-1,\ell} + \sum_{\ell\in Z} g_{k-2\ell} d_{j-1,\ell}$$

The preceding formulas give the so-called cascade algorithm. The algorithm gives a constructive and efficient way to compute approximate values of the scaling function $\phi(x)$ with arbitrarily high precision, provided that $\phi(x)$ has compact support. Use the notation $c_j = (\ldots, c_{j,-1}, c_{j,0}, c_{j,1}, \ldots)$ to denote the low-pass sequence and $d_j = (\ldots, d_{j,-1}, d_{j,0}, d_{j,1}, \ldots)$ to denote the high-pass sequence. Following is the cascade algorithm for approximate values of the scaling function.

1. We start with a low-pass sequence c_0, with $c_{0,0} = 1$ and $c_{0,k} = 0$ for $k \neq 0$, and a high-pass sequence $d_0 = 0$.
2. Set $j = j + 1$. Compute c_j based on c_{j-1} and $d_{j-1} = 0$ using the reconstruction formula. At each step of this cascade algorithm, twice as many values are computed. Values at "even points" $c_{j,2k}$ are refined from the previous iteration

$$c_{j,2k} = \sum_{\ell\in Z} h_{2(k-\ell)} c_{j-1,\ell}$$

and values at "odd points" $c_{j,2k+1}$ are computed by

$$c_{j,2k+1} = \sum_{\ell\in Z} h_{2(k-\ell)+1} c_{j-1,\ell}$$

3. Go to step 2, if $j < J$; otherwise, terminate the iteration and go to step 4 for approximate values of $\phi(x)$.
4. Set $\eta_J(2^{-J}k) = c_{J,k}$. Linearly interpolate the values $\eta_J(2^{-J}k)$ to obtain $\eta_J(x)$ for nondyadic x. This function $\eta_J(x)$ is used to approximate $\phi(x)$.

The error order of this approximation is given by

$$\|\phi - \eta_J\|_\infty = O(2^{-J\alpha})$$

where $\phi(x)$ is assumed to be Hölder continuous with exponent α. Thus, by letting the number of iterations $J \to \infty$, the approximation to $\phi(x)$ can be made with arbitrarily high accuracy. An attractive feature of the cascade algorithm is that it allows one to zoom in on particular features of $\phi(x)$.

For a discrete signal of finite extent $f = (f_1, \ldots, f_n)$, the discrete wavelet transform calculates the coefficients of its wavelet transform approximation. This transformation maps the vector f to a vector of n wavelet coefficients. The discrete wavelet transform can be obtained as $w = Wf$, where W is an orthogonal matrix corresponding to the discrete wavelet transform. To get the wavelet coefficients w, we do not actually perform the matrix multiplication. Instead we use the fast algorithm with complexity $O(n)$.

Functions in the space $L_2(R)$ can be represented by orthonormal wavelet series. We shall notice that wavelet series are just as effective in the analysis of other spaces, such as $L_p(R)$ for $1 < p < \infty$, Sobolev spaces, Hölder spaces, Hardy spaces, and Besov spaces.

Suppose that the mother wavelet $\psi(x)$ arising from a multiresolution analysis of $L_2(R)$ has regularity of degree $r \geq 1$. Then the set $\{\psi_{j,k}(x): j, k \in Z\}$ constitutes an unconditional basis for $L_p(R)$, $1 < p < \infty$. It clearly does not apply to the spaces $L_1(R)$ and $L_\infty(R)$, which have no unconditional bases. Moreover, a wavelet series $f(x) = \sum_{j,k\in Z} \langle f, \psi_{j,k}\rangle \psi_{j,k}(x)$ belongs to $L_p(R)$ if and only if

$$\left(\sum_{j,k\in Z} \langle f, \psi_{j,k}\rangle^2 \psi_{j,k}^2(x)\right)^{1/2} \in L_p(R)$$

if and only if

$$\left(\sum_{j,k\in Z} \langle f, \psi_{j,k}\rangle^2 I(2^{-j}k \leq x \leq 2^{-j}(k+1))\right)^{1/2} \in L_p(R)$$

where I is the indicator function.

Similarly, wavelets provide unconditional bases and characterizations for Sobolev spaces $W_2^s(R)$, $s \geq 0$. The Sobolev spaces $W_2^s(R)$ are defined by

$$W_2^s(R) = \left\{ f \in L_p(R): \int_{-\infty}^{\infty} (1 + |\omega|^2)^s |\hat{f}(\omega)|^2 \, d\omega < \infty \right\}$$

The set $\{\psi_{j,k}(x): j, k \in Z\}$ constitutes an unconditional basis for $W_2^s(R)$ for $0 \leq s < r$. A wavelet series $f(x) = \sum_{j,k\in Z} \langle f, \psi_{j,k}\rangle \psi_{j,k}(x)$ belongs to $W_2^s(R)$ if and only if

$$\sum_{j,k\in Z} \langle f, \psi_{j,k}\rangle^2 (1 + 2^{2js}) < \infty$$

V. FILTERS AND QUADRATURE FILTERS

The discrete wavelet analysis and the associated fast algorithm can be conveniently put into the framework of filtering. Consider a discrete signal $f = \{f_n\}$, which could be finite or infinite. If not otherwise specified, we assume

that the sequence is doubly infinite, i.e., the index n goes from $-\infty$ to $+\infty$.

We use the term filter to denote a convolution operator. Suppose that h and f are both in $\ell_1(Z)$. The filter operator H acting on f is given by

$$(Hf)_n = \sum_{k\in Z} h_{n-k} f_k$$

If the filter sequence h_k is finitely supported, we have a finite impulse response (FIR) filter; otherwise, we have an infinite impulse response (IIR) filter.

We use the term quadrature filter (QF) to denote a linear operator that convolves and decimates. Define a convolution–decimation QF operator H_Q and its adjoint H_Q^* by the following formulas:

$$(H_Q f)_k = \sum_{\ell\in Z} h_{\ell-2k} f_\ell, \quad k \in Z$$

and

$$\left(H_Q^* f\right)_k = \sum_{\ell\in Z} \bar{h}_{k-2\ell} f_\ell, \quad k \in Z$$

An individual quadrature filter is not invertible, as it loses information in the decimation step. However, it is possible to find a pair of QFs so that each preserves the information lost by the other. Such a pair of QFs can be combined into an invertible operator. For convenience, we may suppress the subscript Q for notation of quadrature filters, if there is no confusion.

A pair of QFs H and G satisfying

$$g_k = (-1)^k \bar{h}_{1-k}$$

is called a set of quadrature mirror filters (QMFs). A pair of QFs H and G is called a set of orthogonal QFs, if the following conditions hold:

$$HH^* = GG^* = I$$

$$GH^* = HG^* = 0$$

$$H^*H + G^*G = I$$

$$H\mathbf{1} = \sqrt{2}\,\mathbf{1},\ \mathbf{1} = (\ldots, 1, 1, 1, \ldots)^T$$

There are no symmetric orthogonal exact reconstruction FIR QMFs. The orthogonal QF conditions just given imply that

$$H^*\mathbf{1} = \frac{1}{\sqrt{2}}\mathbf{1}, \qquad G\mathbf{1} = 0, \qquad |G^*\mathbf{1}| = \frac{1}{\sqrt{2}}\mathbf{1}$$

where the absolute value in $|G^*\mathbf{1}|$ is meant entrywise.

Associated with the filter H, define

$$H(\omega) = \sum_{k\in Z} h_k e^{-ik\omega}$$

The filter H is said to be low-pass if the support of $H(\omega)$ contains 0; otherwise, it is said to be high-pass.

The fast wavelet algorithm can be represented through QFs:

Wavelet decomposition: $c_{j-1} = Hc_j,\ d_{j-1} = Gc_j$

Wavelet reconstruction: $c_j = H^*c_{j-1} + G^*d_{j-1}$

VI. SOME EXAMPLES OF WAVELETS

A. Haar Wavelets

The scaling function is

$$\phi(x) = I(0 \le x < 1), \text{ where } I \text{ is the indicator function}$$

The wavelet is

$$\psi(x) = I(0 \le x < 1/2) - I(1/2 \le x < 1)$$

The associated orthogonal QF coefficients are

$$h_0 = h_1 = \frac{1}{\sqrt{2}}, \qquad g_0 = \frac{1}{\sqrt{2}}, \qquad g_1 = -\frac{1}{\sqrt{2}}$$

The Haar wavelet is extremely localized in time, but it has poor regularity (not even continuous). Its Fourier transform satisfies the condition

$$|\hat{\psi}(\omega)| = O(|\omega|^{-1}), \quad |\omega| \to \infty$$

That is, the Haar wavelet has poor localization in the frequency domain.

B. Littlewood–Paley Wavelets or Shannon's Wavelets

The Haar wavelets have very good time localization but poor frequency localization. The other extreme is the Shannon wavelets, also known as the Littlewood–Paley analysis. They have very good frequency localization but poor time localization. They are time-scale mirror images of Haar wavelets.

A function $f \in L_2(R)$ is called bandlimited if its Fourier transform $\hat{f}(\omega)$ has compact support. For simplicity, we assume the support is on $[-\pi, \pi]$. Then $\hat{f}(\omega)$ can be represented by its Fourier series

$$\hat{f}(\omega) = \sum_{k\in Z} f(n) \frac{e^{-ik\omega}}{\sqrt{2\pi}}, \quad \omega \in [-\pi, \pi]$$

It follows that

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \hat{f}(\omega) e^{i\omega x}\, d\omega = \sum_{k\in Z} f(k)\, \frac{\sin \pi(x-k)}{\pi(x-k)}$$

Let $V_j \subset L_2(R)$ be the space of bandlimited functions $f(x)$ whose Fourier transform is supported on $[-2^j\pi, 2^j\pi]$. It can be shown that $\{V_j : j \in Z\}$ constitutes

a multiresolution analysis of $L_2(R)$. The scaling function and its Fourier transform are given by

$$\phi(x) = \frac{\sin(\pi x)}{\pi x}$$

and

$$\hat{\phi}(\omega) = \frac{1}{\sqrt{2\pi}}\, I(-\pi \le \omega < \pi)$$

The set $\{\phi(x-k): k \in Z\}$ forms an orthonormal basis for V_0. It is easy to see that

$$m_0(\omega) = \sum_{k \in Z} I(-\pi/2 + 2k\pi \le \omega < \pi/2 + 2k\pi)$$

Notice that the filter coefficients h_k are given by $H(\omega) = \sqrt{2}\, m_0(\omega)$ for $\omega \in [-\pi, \pi]$. Thus,

$$h_k = \frac{\sqrt{2}}{2\pi} \int_{-\pi}^{\pi} m_0(\omega) e^{i\omega k}\, d\omega = \frac{1}{\sqrt{2}} \frac{\sin(\pi k/2)}{(\pi k)/2}$$

for $k \in Z$. The Shannon filter sequence h_k does not have finite support. It has not quite fast enough decay, $h_k = O(|k|^{-1})$ as $|k| \to \infty$, for absolute summability. Though the filter has poor time localization, it has very good frequency localization property.

By some straightforward calculation, we get the Shannon wavelet function and its Fourier transform given by

$$\psi(x) = \frac{\sin \pi(x - 1/2) - \sin 2\pi(x - 1/2)}{\pi(x - 1/2)}$$

and

$$\hat{\psi}(\omega) = \frac{-e^{-i\omega/2}}{\sqrt{2\pi}}\, I(-2\pi \le \omega < -\pi, \quad \pi \le \omega < 2\pi)$$

C. Daubechies' Wavelets

Daubechies was first to construct compactly supported orthogonal wavelets with a preassigned order, denoted by N, of vanishing moments. It is also of interest to make $\phi(x)$ and $\psi(x)$ reasonably regular.

Start with $m_0(\omega)$ of the form

$$m_0(\omega) = \left(\frac{1 + e^{-i\omega}}{2}\right)^N \mathcal{L}(\omega), \quad N \ge 1 \qquad (11)$$

where $\mathcal{L}(\omega)$ is a trigonometric polynomial. A trigonometric polynomial m_0 of the form (11) satisfies the condition

$$|m_0(\omega)|^2 + |m_0(\omega + \pi)|^2 = 1 \qquad (12)$$

if and only if the function $L(\omega) = |\mathcal{L}(\omega)|^2$ can be written as

$$L(\omega) = P\left(\sin^2 \frac{\omega}{2}\right) \qquad (13)$$

where

$$P(y) = \sum_{k=0}^{N-1} \binom{N + k - 1}{k} y^k + y^N R(1/2 - y) \qquad (14)$$

with $R(y)$ an odd polynomial so chosen that $P(y) \ge 0$ for $y \in [0, 1]$.

To get the function $m_0(\omega)$, we need to extract the square root from $L(\omega)$ using the spectral factorization. The solution is not unique. We can get all possible constructions of $m_0(\omega)$ of the form (11) that satisfy the condition $|m_0(\omega)|^2 + |m_0(\omega + \pi)|^2 = 1$. However, it is not yet sufficient to ensure that such m_0 lead to an orthonormal wavelet basis.

The following theorem gives a necessary and sufficient condition for orthonormality.

Theorem 2 *Suppose* $m_0(\omega) = \frac{1}{\sqrt{2}} \sum_{k=0}^{n} h_k e^{-ik\omega}$ *is a trigonometric polynomial such that* (12) *holds and* $m_0(0) = 1$. *Define* ϕ *and* ψ *by*

$$\hat{\phi}(\omega) = \frac{1}{\sqrt{2\pi}} \prod_{j=1}^{\infty} m_0(2^{-j}\omega)$$

$$\hat{\psi}(\omega) = -e^{-i\omega/2} \hat{\phi}(\omega/2) \overline{m_0(\omega/2 + \pi)}$$

Then ϕ *and* ψ *are compactly supported functions in* $L_2(R)$, *satisfying*

$$\phi(x) = \sqrt{2} \sum_{k=0}^{n} h_k \phi(2x - k)$$

$$\psi(x) = \sqrt{2} \sum_{k=0}^{n} (-1)^k h_{1-k} \phi(2x - k)$$

Moreover, the set of functions $\{\psi_{j,k}(x): j, k \in Z\}$ *constitutes an orthonormal basis for* $L_2(R)$ *if and only if the eigenvalue* 1 *of the* $(2n - 1) \times (2n - 1)$ *matrix* A *given by*

$$A(i, j) = \sum_{k=0}^{n} h_k \overline{h_{j-2i+k}}$$

is nondegenerate.

A family of wavelets corresponding to $m_0(\omega)$ satisfying (11) to (14) with $R \equiv 0$ was first constructed by Daubechies. For each N, the ${}_N m_0$ has $2N$ nonvanishing coefficients; we can choose the phrase of ${}_N m_0$ so that ${}_N m_0(\omega) = \frac{1}{\sqrt{2}} \sum_{k=0}^{2N-1} {}_N h_k e^{-ik\omega}$. Lists of low-pass filter coefficients ${}_N h_k$ for $N = 2, 3, \ldots, 10$ of Daubechies wavelets can be found in Daubechies (1992). However, there are no explicit expressions for the scaling function ${}_N\phi(x)$ and the wavelet ${}_N\psi(x)$. They are computed via the cascade algorithm. These functions ${}_N\phi(x)$ and ${}_N\psi(x)$, referred to as the extremal phase family, are not symmetric.

D. Symmlets

There are no symmetric orthonormal wavelets with compact support. However, it is possible to construct wavelets with compact support that are "more symmetric." The symmlets were also constructed by Daubechies. They were constructed to be as nearly symmetric as possible. Symmlets are also called the least asymmetric Daubechies wavelets. Low-pass filter coefficients for symmlets with vanishing moments of order N, for $N = 4, 5, \ldots, 10$, can be found in Daubechies (1992).

E. Spline Wavelets

The spline wavelets are associated with multiresolution analysis ladders consisting of spline function spaces. A cardinal spline of order m is a function in C^{m-2} such that the restriction of the function to any interval $[k, k+1)$, $k \in Z$, is an order m (or degree less than or equal to $m-1$) polynomial. We take V_0 the $L_2(R)$ subspace of cardinal splines of order m and $V_j = \{f(2^j x)\colon f(x) \in V_0\}$.

If we choose $\phi(x)$ to be the piecewise constant spline,

$$\phi(x) = I(0 \le x < 1)$$

then we end up with the Haar wavelet.

The next example is the piecewise linear spline. The linear cardinal B-spline with support on $[-1, 1]$ is given by

$$B(x) = \begin{cases} 1 - x & x \in [0, 1] \\ 1 + x & x \in [-1, 0] \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

$$B(x) = (1 - |x|) I(0 \le |x| \le 1)$$

The set of B-splines $\{B(x-k)\colon k \in Z\}$ forms an unconditional basis for the $L_2(R)$ subspace V_0 consisting of cardinal splines of order 2. Define

$$H(x) = \sum_{k=-\infty}^{\infty} B(x-k) c_1 r_1^{|k|}$$

where $r_1 = -2 + \sqrt{3}$ and $c_1 = \sqrt{3}$. Notice that for a fixed x the above sum is actually a finite sum of two terms. Let

$$\phi(x) = \frac{3+\sqrt{3}}{6} H(x) + \frac{3-\sqrt{3}}{6} H(x-1) \tag{16}$$

Then $\{\phi(x-k)\}_{k \in Z}$ constitutes an orthonormal basis for V_0. The scaling function $\phi(x)$ has exponential decay. Its Fourier transform satisfies the condition

$$\hat{\phi}(\omega) = O(|\omega|^{-2}), \quad |\omega| \to \infty$$

The associated low-pass and high-pass filter coefficients can be computed using the explicit expression (16): $h_k = \langle \phi, \phi_{1,k} \rangle$ and $g_k = (-1)^k h_{1-k}$.

For spline wavelets, they can be chosen in C^k. The decay rate is exponential order in the time domain and polynomial order $k+2$ in the frequency domain.

F. Coiflets

Coiflets are wavelets that both the wavelet $\psi(x)$ and the scaling function $\phi(x)$ have vanishing moments of order N, i.e.,

$$\int_{-\infty}^{\infty} x^k \phi(x)\, dx = 0, \quad k = 1, \ldots, N-1$$

$$\int_{-\infty}^{\infty} x^k \psi(x)\, dx = 0, \quad k = 0, \ldots, N-1$$

If $f(x)$ is smooth enough, the following formulas are accurate up to error of order N:

$$\int_{-\infty}^{\infty} f(x) \frac{1}{h} \phi\left(\frac{x - x_0}{h}\right) dx = f(x_0) + O(h^N), \quad h \to 0$$

and

$$\int_{-\infty}^{\infty} f(x) \frac{1}{h} \psi\left(\frac{x - x_0}{h}\right) dx = O(h^N), \quad h \to 0$$

Thus, the wavelet series expansion for $f(x)$ is

$$f(x) = \sum_k f(2^{-j}k) \phi(2^j x - k) + O(2^{-jN})$$

provided that $f \in C^N(R) \cap L_2(R)$ and $\|f^{(N)}\|_\infty < \infty$. The function

$$\sum_k f(2^{-j}k) \phi(2^j x - k)$$

can be intepreted as a "blurred version" of f.

Coiflets are almost symmetric. The price to pay for this additional symmetry is a larger support. The length of the support for coiflets is $3N-1$ compared to $2N-1$ for the standard Daubechies family. The filter coefficients for the coiflets with $N = 2, 4, 6, 8, 10$ can be found in Daubechies (1992).

VII. WAVELET PACKETS

Wavelet packet functions consist of a rich family of building block functions. Mixtures of H and G may be applied to functions $\phi(x)$ and $\psi(x)$ to get new building block functions. Families of new orthonormal basis functions so produced are called wavelet packets. Wavelet packet functions are more flexible than wavelets in representing different types of signals. In particular, they are better at representing signals that exhibit oscillatory or periodic behavior.

A. Unit-Scale Wavelet Packets

Let us use the notation

$$\begin{cases}\mu_0(x) = \phi(x)\\ \mu_1(x) = \psi(x)\end{cases}$$

Define the following sequence of functions:

$$\begin{cases}\mu_{2m}(x) &= \sum_{\ell\in Z} h_\ell \sqrt{2}\,\mu_m(2x-\ell)\\ \mu_{2m+1}(x) &= \sum_{\ell\in Z} g_\ell \sqrt{2}\,\mu_m(2x-\ell)\end{cases} \tag{17}$$

Functions $\{\mu_n(x-k): n\in Z_+, k\in Z\}$ are called wavelet packets associated to H and G (or to ϕ and ψ). Notice that the scaling equations (17) can be rewritten as

$$\begin{cases}\mu_{2m}(x-k) &= \sum_{\ell\in Z} h_{\ell-2k} \sqrt{2}\,\mu_m(2x-\ell)\\ \mu_{2m+1}(x-k) &= \sum_{\ell\in Z} g_{\ell-2k} \sqrt{2}\,\mu_m(2x-\ell)\end{cases} \tag{18}$$

Theorem 3 *If H and G are orthogonal QFs, then $\{\mu_n(x-k): n\in Z_+, k\in Z\}$ forms an orthonormal basis for $L_2(R)$.*

Let Λ_n be the closed linear subspace of $L_2(R)$ spanned by $\{\mu_n(x-k): k\in Z\}$. If H and G are orthogonal QFs, then $\{\Lambda_n: n\geq 0\}$ is an orthogonal decomposition of $L_2(R)$ into subspaces of unit-scale functions of different frequencies.

B. Multiscale Wavelet Packets

A wavelet packet of scale index s, frequency index n, and translation index k is given by

$$\mu_{snk}(x) = 2^{s/2}\mu_n(2^s x - k)$$

Denote by Λ_{sn} the closed linear subspace of $L_2(R)$ spanned by $\{\mu_{snk}(x): k\in Z\}$. Define a dyadic interval I_{sn} by

$$I_{sn} = [2^s n, 2^s(n+1))$$

Notice that for any two dyadic intervals, either they are disjoint or one contains the other. There is a natural one-to-one correspondence between the dyadic interval I_{sn} and the subspace Λ_{sn}. Thus, the dyadic intervals can be used to keep track of multiscale wavelet packets.

Lemma 1 *Suppose H and G are orthogonal QFs. Functions $\mu_{snk}(x)$ and $\mu_{s'n'k'}(x)$ are orthogonal to each other if I_{sn} and $I_{s'n'}$ are disjoint, or if I_{sn} and $I_{s'n'}$ are identical but $k\neq k'$.*

Denote an arbitrary collection of disjoint dyadic intervals by $\mathcal{I}$ and its corresponding subspace by $\Lambda_{\mathcal{I}}$, where

$$\Lambda_{\mathcal{I}} = \overline{\cup_{I_{sn}\in\mathcal{I}} \Lambda_{sn}}\,.$$

If two collections of dyadic intervals $\mathcal{I}_1$ and $\mathcal{I}_2$ are disjoint, then $\Lambda_{\mathcal{I}_1}$ and $\Lambda_{\mathcal{I}_2}$ are linearly independent. If $\mathcal{I}$ covers R^+, then $\Lambda_{\mathcal{I}} = L_2(R)$ and the space $L_2(R)$ can be decomposed via dyadic intervals in $\mathcal{I}$:

$$L_2(R) = \sum_{I_{sn}\in\mathcal{I}} \Lambda_{sn}$$

Theorem 4 *If H and G are orthogonal QFs and $\mathcal{I}$ is a disjoint dyadic covering of R^+, then $\{\mu_{snk}(x): k\in Z, I_{sn}\in\mathcal{I}\}$ forms a complete orthonormal basis for $L_2(R)$.*

The above theorem may be called graph theorem because disjoint dyadic coverings can be viewed as graphs from partitions of R^+ to bases for $L_2(R)$. The orthonormal basis associated with each disjoint dyadic covering is called orthonormal graph basis. Graph bases provide a large library of bases. Some examples of wavelet packet graph bases are

Walsh-type basis:	$\Lambda_0 \oplus \Lambda_1 \oplus \cdots \Lambda_n \oplus \cdots$	$n\in N$
subband basis:	$\Lambda_{s0} \oplus \Lambda_{s1} \oplus \cdots \Lambda_{sn} \oplus \cdots$	$n\in N$
wavelet basis:	$\cdots \oplus \Lambda_{-1,1} \oplus \Lambda_{0,1} \oplus \cdots \oplus \Lambda_{s1}\cdots$	$s\in Z$

For any disjoint dyadic covering $\mathcal{I}$ of R^+, every function $f\in L_2(R)$ can be represented by wavelet packets as

$$f(x) = \sum_{I_{sn}\in\mathcal{I}, k\in Z} \langle f, \mu_{snk}\rangle\, \mu_{snk}(x)$$

Example 1 (Haar–Walsh wavelet packets) *Consider the scaling function $\phi(x) = I(0\leq x<1)$ and the wavelet $\psi(x) = I(0\leq x<1/2) - I(1/2\leq x<1)$ from the Haar basis. Define a system of functions in a recursive manner as*

$$\begin{cases}\mu_0(x) = \phi(x)\\ \mu_{2m+1}(x) = \mu_m(2x) - \mu_m(2x-1), & m\geq 0\\ \mu_{2m}(x) = \mu_m(2x) + \mu_m(2x-1), & m\geq 1\end{cases}$$

The set of functions

$$\{\mu_n(x-k): n\in Z_+, k\in Z\}$$

forms an orthonormal basis for $L_2(R)$, called the Haar–Walsh basis, and the set of functions

$$\{\mu_{2^m}(x-k), \mu_{2^m+1}(x-k), \ldots, \mu_{2^{m+1}-1}(x-k): k\in Z\}$$

forms an orthonormal basis for the wavelet subspace W_m. The scaled functions $\mu_{snk}(x) = 2^{s/2}\mu_n(2^s x - k)$ are called Haar–Walsh wavelet packets.

Like the discrete wavelet transform, wavelet packet coefficients can be calculated in an efficient way. The fast algorithm for discrete wavelet transform extends in a straightforward manner to wavelet packet analysis. For a given function $f\in L_2(R)$, define

$$\lambda_{snk} = \langle f, \mu_{snk}\rangle = \int_{-\infty}^{\infty} f(x)\overline{\mu_{snk}(x)}\,dx$$

Let $\lambda_{sn} = \{\lambda_{snk} : k \in Z\}$ be a sequence of numbers. Then we have

$$\begin{aligned}(H\lambda_{sn})_k &= \sum_{\ell\in Z} h_{\ell-2k}\lambda_{sn\ell} \\ &= \sum_{\ell\in Z} h_{\ell-2k}\int_{-\infty}^{\infty} f(x)\overline{2^{s/2}\mu_n(2^s x-\ell)}\,dx \\ &= \int_{-\infty}^{\infty} f(x)\left(\sum_{\ell\in Z} h_{\ell-2k}\overline{2^{s/2}\mu_n(2^s x-\ell)}\right)dx \\ &= \int_{-\infty}^{\infty} f(x)\overline{2^{(s-1)/2}\mu_{2n}(2^{s-1}x-k)}\,dx \\ &= \lambda_{s-1,2n,k}\end{aligned}$$

Similarly, we have

$$(G\lambda_{sn})_k = \sum_{\ell\in Z} g_{\ell-2k}\lambda_{sn\ell} = \lambda_{s-1,2n+1,k}$$

To obtain the wavelet packet analysis of a function, we first find its coefficient sequence in the root subspace; then H or G is applied down the branches of wavelet packet coefficient tree:

$$H\lambda_{sn} = \lambda_{s-1,2n} \qquad \text{and} \qquad G\lambda_{sn} = \lambda_{s-1,2n+1}$$

The wavelet packet analysis gives a decomposition of the function into the whole collection of its wavelet packet components, which are more than a basis set. A basis subset can be chosen by checking that its associated dyadic intervals form a disjoint covering of R^+.

In the reverse direction, we can synthesize the root coefficients via the adjoints H^* and G^* in the following way:

$$\lambda_{sn} = H^*\lambda_{s-1,2n} + G^*\lambda_{s-1,2n+1}$$

VIII. BEST BASIS SELECTION IN WAVELET PACKET ANALYSIS

The wavelet packet components of a funciton contain redundant information. To select among wavelet packet bases, we first have to define cost criteria for judging the bases.

Let $\mathcal{L}$ be a library of bases (the collection of wavelet packets is an example). Let $\mathcal{B} \subset \mathcal{L}$ be a basis. Given the data $x = (x_1, \ldots, x_n)$, let the cost of the basis $\mathcal{B}$ for representation of x be denoted by $C(\mathcal{B}x)$. There are some papular cost measures:

- The entropy cost: $C(y) = -\sum_{i=1}^n p_i \log p_i$, where $p_i = |y_i|^2/\|y\|_2^2$.
- The threshold cost: $C(y) = \sum_{i=1}^n I(|y_i| \geq \delta)$, $\delta > 0$.
- The ℓ_p-norm cost: $C(y) = (\sum_{i=1}^n |y_i|^p)^{1/p}$ for $0 < p < 2$. (Since orthogonal transform preserves the ℓ_2-norm, the costs for $p = 2$ are the same for all bases.)
- The SURE (Stein's unbiased risk estimate) cost:

$$C(y) = \sigma^2\left(n - 2\sum_{i=1}^n I(|y_i| \leq \sigma\delta) + \sum_{i=1}^n \min\{(y_i/\sigma)^2, \delta^2\}\right)$$

where $\delta = \sqrt{2\log_e(n\log_2 n)}$ and σ is the noise level of data.

Suppose that the data is of length 2^J and the height of the packet tree is J. Recall that λ_{sn} has two children $\lambda_{s-1\ 2n}$ and $\lambda_{s-1\ 2n+1}$. The algorithm leading to the selection of best basis is briefly described as follows.

0. Let $\lambda_{00} = (\lambda_{000}, \ldots, \lambda_{00\ 2^J-1})$ come from data. Order the wavelet packet components $\lambda_{sn} = (\lambda_{sn0}, \ldots, \lambda_{sn\ 2^{j_0+s}-1})$ into a tree, where the resolution scale ranges over $s = 0, -1, \ldots, -J$, and $n = 0, \ldots, 2^{|s|} - 1$. The top of the tree is λ_{00}, the second row is $\lambda_{-1,0}$ and $\lambda_{-1\ 1}$, and the bottom of the tree is $\lambda_{-J\ 0}, \lambda_{-J\ 1}, \ldots, \lambda_{-J+1\ 2^J-1}$. Put all these components into the library $\mathcal{L}$. Start with $\mathcal{B}$, an empty set, as a candidate set of best basis.

1. Consider the current wavelet packet component $\lambda_{sn} \in \mathcal{L}$ in the wavelet packet tree as a candidate component for inclusion into $\mathcal{B}$.

2. If $s = -J$ (the coarsest resolution level) or if

$$C(\lambda_{sn}) \leq C(\lambda_{s-1\ 2n}) + C(\lambda_{s-1\ 2n+1})$$

then go to the next step. Else discard λ_{sn} from $\mathcal{L}$ and return to step 1.

3. Move λ_{sn} to $\mathcal{B}$. Remove λ_{sn} and all its descendants from $\mathcal{L}$. Stop if $\mathcal{L}$ is empty, otherwise return to step 1.

IX. WAVELETS IN STATISTICAL ESTIMATION

With the basic introduction to wavelets, we now turn to their application to nonparametric statistical estimation.

A. Wavelet Density Estimation

Let $X_1, \ldots, X_n$ be independently and identically distributed random variables from a distribution with probability density function $f(x)$. Consider the problem of estimating $f(x)$ assuming that $f(x)$ lies a priori in a functional class with certain regularity conditions.

Represent $f(x)$ in terms of wavelet expansion

$$f(x) = \sum_{k\in Z} c_{j,k}\phi_{j,k}(x) + \sum_{\ell\geq j}\sum_{k\in Z} d_{\ell,k}\psi_{\ell,k}(x)$$

The coefficients $c_{j,k}$ and $d_{\ell,k}$ can be estimated by

$$\hat{c}_{j,k} = \frac{1}{n}\sum_{i=1}^{n}\phi_{j,k}(X_i) \qquad \text{and} \qquad \hat{d}_{\ell,k} = \frac{1}{n}\sum_{i=1}^{n}\psi_{\ell,k}(X_i)$$

These estimates $\hat{c}_{j,k}$ and $\hat{d}_{\ell,k}$ are called empirical wavelet coefficients. Notice that the empirical wavelet coefficients are unbiased; that is,

$$E\hat{c}_{j,k} = E\phi_{j,k}(X) = c_{j,k}$$

and

$$E\hat{d}_{\ell,k} = E\psi_{\ell,k}(X) = d_{\ell,k}$$

Also notice that

$$Var(\hat{c}_{j,k}) = \frac{1}{n}\int_{-\infty}^{\infty}\phi_{j,k}^2(x)f(x)\,dx = O\left(\frac{2^j}{n}\right)$$

and

$$Var(\hat{d}_{\ell,k}) = \frac{1}{n}\int_{-\infty}^{\infty}\psi_{\ell,k}^2(x)f(x)\,dx = O\left(\frac{2^\ell}{n}\right)$$

To estimate $f(x)$, we have to truncate at a certain resolution level, say, J, and get

$$\hat{f}(x) = \sum_{k\in Z}\hat{c}_{j,k}\phi_{j,k}(x) + \sum_{\ell=j}^{J-1}\sum_{k\in Z}\hat{d}_{\ell,k}\psi_{\ell,k}(x) \qquad (19)$$

which is equivalent to

$$\hat{f}(x) = \sum_{k\in Z}\hat{c}_{J,k}\phi_{J,k}(x) \qquad (20)$$

For compactly supported $\phi(x)$ and $\psi(x)$, the above summations $\sum_{k\in Z}$ in (19) and (20) are finite sums. For other infinitely supported wavelets with rapid decay

$$|\psi(x)| \leq c(1+|x|)^{\gamma}, \quad c>0,\ \gamma>1$$

the summations $\sum_{k\in Z}$ just given can be truncated to finite sums with arbitrary prescribed precision.

The linear estimator (19), or equivalently (20), can be written as a projection type kernel estimator. As seen previously, the projection of a function $f(x)\in L_2(R)$ onto the subspace V_J is given by $\mathcal{P}_{V_J}f(x) = \int_{-\infty}^{\infty}\mathcal{V}_J(x,y)f(y)\,dy$. Notice that

$$\mathcal{V}_J(x,y) = 2^J\mathcal{V}(2^Jx, 2^Jy)$$

Instead of dyadic bandwidths, we may also consider kernels with continuously varying bandwidths

$$\mathcal{V}_h(x,y) = \frac{1}{h}\mathcal{V}\left(\frac{x}{h},\frac{y}{h}\right)$$

A density estimator based on the projection kernel $\mathcal{V}_h(x,y)$ is given by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_h(x,X_i) = \frac{1}{nh}\sum_{i=1}^{n}\mathcal{V}\left(\frac{x}{h},\frac{X_i}{h}\right) \qquad (21)$$

The asymptotic bias and variance of the linear wavelet density estimator (21) are discussed here after. Suppose that $\mathcal{V}(x,y)$ is of order m. Recall that

$$b_m(x) = x^m - \int_{-\infty}^{\infty}\mathcal{V}(x,y)y^m\,dy \qquad (22)$$

For $h>0$, we have

$$x^m - \int_{-\infty}^{\infty}\mathcal{V}_h(x,y)y^m\,dy = h^m b_m\left(\frac{x}{h}\right)$$

Theorem 5 *Assume that* $f(x)\in C^{m,\alpha}(R)\cap L_2(R)$, *for some* $0<\alpha\leq 1$, *that* $\int_{-\infty}^{\infty}|\mathcal{V}(x,y)(y-x)^{m+\alpha}|\,dy<\infty$, *and that* $h\to 0$, $nh\to\infty$, *as* $n\to\infty$. *Then for a fixed* x, *we have the following pointwise asymptotic bias*:

$$E\hat{f}(x) - f(x) = \frac{-1}{m!}f^{(m)}(x)b_m\left(\frac{x}{h}\right)h^m + O(h^{m+\alpha})$$

Moreover, if $f^{(m)}$ *is in* $L_2(R)$, *then the integrated squared bias is*

$$\|E\hat{f} - f\|_2^2 = \frac{b_{2m}}{(2m)!}\|f^{(m)}\|_2^2 h^{2m} + O\left(h^{2(m+\alpha)}\right)$$

where

$$b_{2m} = (2m)!(m!)^{-2}\int_0^1 b_m^2(x)\,dx$$

Example 2 (Spline wavelets) *Consider a density estimator using spline wavelets of order* m. *Then we have*

$$b_m(x) = B_m(x) \text{ for } x\in(0,1) \text{ and } b_{2m} = |B_{2m}|$$

where $B_m(x)$ *is the* m*th Bernoulli polynomial and* B_{2m} *is the* $2m$*th Bernoulli number*.

Example 3 (Daubechies' wavelets) *Consider a density estimator using* ${}_N\phi(x)$. *An expression for the bias* $b_N(x)$ *is*

$$b_N(x) = x^N - \sum_{\ell=0}^{N}C(N,\ell)\,a_N^{\ell}\,\Phi_N^{N-\ell}(x)$$

where $C(N,\ell) = N!/\ell!(N-\ell)!$ *and*

$$a_N^{\ell} = \int_0^{2N-1}{}_N\phi(x)x^{\ell}\,dx,$$

$$\Phi_N^{N-\ell}(x) = \sum_{k=-2N+2}^{0}k^{N-\ell}\,{}_N\phi(x-k)$$

Theorem 6 *Suppose that* $f(x)\in C^1(R)$ *and that* $f(x)$ *and* $f'(x)$ *are uniformly bounded. For a fixed* x, *we have*

$$\text{Var }\hat{f}(x) = \frac{1}{n}f(x)v_h(x) + O\left(\frac{1}{n}\right)$$

where

$$v_h(x) = \int_{-\infty}^{\infty} \mathcal{V}_h^2(x, y)\, dy = \frac{1}{h}\mathcal{V}\left(\frac{x}{h}, \frac{x}{h}\right) \tag{23}$$

Moreover, the integrated variance is given by

$$\int_{-\infty}^{\infty} \mathrm{Var}\ \hat{f}(x)\, dx = \frac{1}{nh} + O\left(\frac{1}{n}\right)$$

In a wavelet method, the bandwidth selection problem is often considered in a discrete manner, i.e., $h = 2^{-j}$. However, from the kernel point of view, bandwidths can be chosen in a continuous manner. Any automatic bandwidth selector for a convolution type kernel estimator can be tried out for estimator (21). Theoretical optimal bandwidth can be obtained by minimizing the asymptotic IMSE

$$\mathrm{IMSE} = E\|\hat{f} - f\|_2 \simeq \frac{b_{2m}}{(2m)!}\left\|f^{(m)}\right\|_2^2 h^{2m} + \frac{v}{nh} \tag{24}$$

The optimal bandwidth is

$$h_{\mathrm{opt}} = \left(\left\|f^{(m)}\right\|_2^2\right)^{-1/(2m+1)} \left(\frac{(2m-1)!v}{b_{2m}}\right)^{1/(2m+1)} \times n^{-1/(2m+1)} \tag{25}$$

This optimal bandwidth suggests that the resolution level J should be chosen to satisfy the order

$$2^J = O\left(n^{1/(2m+1)}\right)$$

The preceding formulas (24) and (25) are valid for a convolution type kernel $\mathcal{V}(x)$ with $v = \int_{-\infty}^{\infty} \mathcal{V}^2(x)\, dx$ and $b_{2m} = (2m)!(m!)^{-2}(\int_{-\infty}^{\infty} \mathcal{V}(x)x^m\, dx)^2$. Plugging h_{opt} into (24), we have

$$\mathrm{IMSE}_{\mathrm{opt}} \simeq \frac{2m+1}{2m}\left(\frac{b_{2m}}{(2m-1)!}\right)^{1/2m+1} \times \left\|f^{(m)}\right\|_2^{2/2m+1}\left(\frac{v}{n}\right)^{2m/2m+1} \tag{26}$$

Let $C_m(\mathcal{V}) = b_{2m}^{1/2m+1} v^{2m/2m+1}$. The relative efficiency of two kernels $\mathcal{V}^\star$ to $\mathcal{V}$ is defined as

$$\text{rel eff} = \{C_m(\mathcal{V})/C_m(\mathcal{V}^\star)\}^{(2m+1)/2m} \tag{27}$$

Wavelet estimators have good asymptotic behavior not only in the L_2 norm, but also in general L_p norms, and not only for Hölder continuous functions, but also for functions in the Besov class. In the following theorem, wavelet estimators can attain the minimax convergence rate judged by $E\|\hat{f} - f\|_p$ on the Besov class.

Theorem 7 *Suppose that the underlying density f is in $F_p^{sq}(M) = \{f \in B_p^{sq}: \|f\|_{B_p^{sq}} \le M\}$, $s > 0$, $2 \le p < \infty$, $1 \le q < \infty$, and that the multiresolution analysis has regularity $r > s$. Then, taking the resolution level J so that $2^J = O(n^{1/(2s+1)})$, the estimator $\hat{f}$ in (20) has the upper bound of convergence rate*

$$\sup_{f \in F_p^{sq}(M)} E\|\hat{f} - f\|_p^p \le c_1\, n^{-sp/(2s+1)},$$

for some constant $c_1 > 0$

This upper bound is still true for $1 < p < 2$ if one requires in addition that $f(x) < w(x)$, $x \in R$, for some symmetric unimodal function $w(x) \in L_{p/2}(R)$.

The rate $O(n^{-sp/(2s+1)})$ is also known to be the minimax rate on Besov class in the sense that

$$\inf_{\hat{f} \in F^*} \sup_{f \in F_p^{sq}(M)} E\|\hat{f} - f\|_p^p \ge c_2\, n^{-sp/(2s+1)},$$

for some constant $c_2 > 0$

where F^* is the set of estimators taking their values in a class containing $F_p^{sq}(M)$. Thus, linear wavelet estimators can attain the minimax rate on Besov class.

Instead of linear estimators, we may consider nonlinear estimators to allow wavelets to perform innate adaptive fit by shrinking the wavelet coefficients toward zero.

B. Wavelet Shrinkage

In the density estimation setting, we first compute the empirical wavelet coefficients $\hat{c}_{j,k}$ and $\hat{d}_{\ell,k}$ for $\ell = j, j+1, \ldots, J-1$, and $k \in Z$. Let $\Delta(\cdot, \delta)$ be a shrinkage/thresholding rule. Examples of $\Delta(\cdot, \delta)$ include

Hard threshold: $\Delta^H(d, \delta) = dI(|d| \ge \delta)$

Soft threshold: $\Delta^S(d, \delta) = sign(d)(|d| - \delta)I(|d| \ge \delta)$

Diagonal shrinkage: $\Delta^D(d, \delta) = \left(1 - \frac{\delta^2}{d^2}\right) dI(d^2 \ge \delta^2)$

A nonlinear wavelet density estimator of $f(x)$ can be obtained via a certain shrinkage/thresholding rule Δ:

$$\hat{f}(x) = \sum_k \hat{c}_{j,k}\phi_{j,k}(x) + \sum_{\ell=j}^{J-1}\sum_k \Delta(\hat{d}_{\ell,k}, \delta)\psi_{\ell,k}(x)$$

A suggested choice for the finest resolution level is $J = [\log_2 n]$, where $[t]$ is the greatest integer less than or equal to t. For simplicity, we may assume that the data size is $n = 2^J$.

The wavelet shrinkage estimation is particularly useful for regression function estimation and denoising. Consider a nonparametric regression problem with observations y_i arising from the following regression model:

$$y_i = f(x_i) + \sigma\epsilon_i, \quad i = 1, \ldots, n, x_i \in [0, 1] \tag{28}$$

where $x_1, \ldots, x_n$ are equally spaced design points, $\sigma > 0$ is a noise level, and ϵ_i are independently and identically distributed $N(0, 1)$ random errors. In vector form, the model (28) can be written as

$$y = f + \sigma\epsilon$$

where $y=(y_1,\ldots,y_n)$, $f=(f(x_1),\ldots,f(x_n))$, and $\epsilon=(\epsilon_1,\ldots,\epsilon_n)$. The goal is to estimate the regression function $f(x)$ based on the observations y.

Represent f in terms of wavelet expansion

$$f(x)=\sum_k c_{j,k}\phi_{j,k}(x)+\sum_{\ell\geq j}\sum_k d_{\ell,k}\psi_{\ell,k}(x)$$

The empirical wavelet coefficients are given by

$$\hat{c}_{j,k}=\frac{1}{n}\sum_{i=1}^{n}\phi_{j,k}(x_i)y_i \quad \text{and} \quad \hat{d}_{\ell,k}=\frac{1}{n}\sum_{i=1}^{n}\psi_{\ell,k}(x_i)y_i$$

Order these empirical wavelet coefficients into a vector form

$$w=\sqrt{n}(\hat{c}_{j,1},\ldots,\hat{c}_{j,2^j},\hat{d}_{j,1},\ldots,\hat{d}_{J-1,2^{J-1}}) \qquad (29)$$

and w is indexed as

$$w_{j,k}=\sqrt{n}\,c_{j,k}, \qquad w_{\ell+1,k}=\sqrt{n}\,d_{\ell,k}$$

The vector w can be obtained by an orthogonal transform of data $w=Wy$, where W is the orthogonal matrix corresponding to the discrete wavelet transform. Represent w as

$$w=\theta+\sigma e, \quad \text{where } \theta=Wf \text{ and } e\sim N(0,I) \qquad (30)$$

Now θ is the object of interest and we wish to estimate it with ℓ_2 loss. Define the risk measure

$$R(\hat{\theta},\theta)=E\|\hat{\theta}-\theta\|_{\ell_2}^2$$

Suppose that we had available an oracle that would supply for us the ideal coefficients $I(|\theta_i|\geq\sigma)$ optimal for use in the keep-or-kill scheme. The ideal scheme consists in estimating only those θ_i larger than the noise level. The ideal estimate of θ is then given by

$$\hat{\theta}_{\ell,k}=w_{\ell,k}I(|\theta_{\ell,k}|\geq\sigma), \quad k=1,\ldots,2^{\ell},\ \ell=j,\ldots,2^J$$

The ideal risk is

$$R_{\text{ideal}}(\hat{\theta},\theta)=\sum_{\ell,k}\min\{\theta_{\ell,k}^2,\sigma^2\}$$

In general, the ideal risk can not be attained by any estimator, linear or nonlinear. However, simple estimates do come remarkably close. The shrinkage rules Δ^H, Δ^S, and Δ^D had been known in statistics before the discovery of wavelets. They can apply to the scaling and wavelet coefficients simultaneously, or apply only to wavelet coefficients and retain the scaling coefficients unchanged.

The preceding shrinkage rules retain only data that exceed a certain amount of the noise level. There are several methods for setting up the threshold parameter δ.

The universal threshold. The threshold parameter is set to

$$\delta=\sigma\sqrt{2\log n}$$

The minimax threshold. It minimizes a theoretical upper bound on the asymptotic risk. They are always smaller than the universal threshold for a given sample size, and thus results in less smoothing. Values of the minimax threshold for various sample sizes are computed and listed in Table 2 of Donoho and Johnstone (1994).

SURE. It is based on the principle of minimizing the Stein unbiased risk estimate at each resolution level. The level-dependent threshold value at resolution level ℓ is given by

$$\delta_\ell^{\text{SURE}}=\sigma\left(\arg\min_{0\leq\delta\leq\sqrt{2\log n_\ell}}\text{SURE}(\sigma^{-1}w_\ell,\delta)\right)$$

where

$$\text{SURE}(\sigma^{-1}w_\ell,\delta)=n_\ell-2\sum_{k=1}^{n_\ell}I(\sigma^{-1}|w_{\ell,k}|\leq\delta)+\sum_{k=1}^{n_\ell}\min\{\sigma^{-2}w_{\ell,k}^2,\delta^2\}$$

with n_ℓ is the number of coefficients in the resolution level. In this article, $n_\ell=2^\ell$.

The SURE threshold should only be used with the soft shrinkage. The SURE threshold can perform poorly if the coefficients are very sparse. Define

$$s_\ell^2=\frac{1}{n_\ell}\sum_{k=1}^{n_\ell}(w_{\ell,k}^2-1) \qquad \text{and} \qquad \gamma_\ell=\frac{(\log_2 n_\ell)^{3/2}}{\sqrt{n_\ell}}$$

Wavelet coefficients at a level ℓ are considered sparse if $s_\ell^2\leq\gamma_\ell$. To correct the drawback in situations of sparse wavelet coefficients, the following hybrid rule combining SURE and universal threshold is suggested. When the coefficients are not sparse, use the SURE threshold; otherwise, use the universal threshold. This hybrid method is proposed by Donoho and Johnstone (1995). It works as follows. At a resolution ℓ,

$$\hat{\theta}_{\ell,k}=\begin{cases}\Delta^S(w_{\ell,k},\delta_\ell^{\text{SURE}}), & s_\ell^2>\gamma_\ell\\ \Delta^S(w_{\ell,k},\sigma\sqrt{2\log n}), & s_\ell^2\leq\gamma_\ell\end{cases}$$

C. Bayesian Wavelet Shrinkage

Consider still the nonparametric regression setting (28) on an finite interval; without loss of generality, $x\in[0,1]$. The main objective of this section is to study the wavelet estimation from a Bayesian viewpoint. The Bayesian formulation here is inspired by models of smoothing splines.

1. The Prior Model

The underlying function of interest, $f(x)$, has the wavelet series representation

$$f(x) = \sum_{k=1}^{2^j} c_{j,k}\phi_{j,k}(x) + \sum_{\ell \geq j}\sum_{k=1}^{2^\ell} d_{\ell,k}\psi_{\ell,k}(x), \quad x \in [0,1] \tag{31}$$

where $\phi_{j,k}(x)$ and $\psi_{\ell,k}(x)$ are scaling functions and wavelets on the interval $[0, 1]$. These scaling functions and wavelets are assumed to retain orthonormality. It is also assumed that this system is with regularity $r > 0$ and $\psi(x)$ has vanishing moments of order $N > r$. The properties of regularity and vanishing moments ensure that these ϕs and ψs form an unconditional basis for Besov spaces $B^s_{p,q}$ for $0 < s < r$ and $1 \leq p, q \leq \infty$. The construction of wavelets on an interval can be found in Cohen *et al.* (1993a,b).

The scaling coefficients $c_{j,k}$ in (31) are assume fixed but unknown and the wavelet coefficients $d_{\ell,k}$ are assumed uncorrelated random variables with zero mean and a common variance σ^2. Such model is a nonparametric mixed-effects model.

We briefly summarize the definition of Besov space. For $0 < b \leq 1$ and $1 \leq p, q \leq \infty$, $J^b_{p,q}$ is defined by

$$J^b_{p,q}(f) = \begin{cases} \left(\int_{h=0}^{1/(1+[b])} h^{-qb-1} \left\|\Delta_h^{1+[b]} f\right\|^q_{L_p[0,1-(1+[b])h]} dh\right)^{1/q}, & q < \infty \\ \sup_{h\in[0,1/(1+[b])]} h^{-b} \left\|\Delta_h^{1+[b]} f\right\|_{L_p[0,1-(1+[b])h]}, & q = \infty \end{cases}$$

where Δ^k_h is the kth-order difference and $[b]$ is the greatest integer less than or equal to b. When $s > 0$ with $s = a + b$ for a positive integer a and $0 < b \leq 1$, the Besov space $B^s_{p,q}$ is the collection of functions f such that $f, f^{(1)}, \ldots, f^{(a)} \in L_p[0, 1]$ and $J^b_{p,q}(f^{(a)}) < \infty$. This space $B^s_{p,q}$ is equipped with the norm $\|f\|_{B^s_{p,q}} = \|f\|_{L_p} + \sum_{k=1}^a J^b_{p,q}(f^{(k)})$.

Expand $f(x)$ in terms of the wavelet basis:

$$f(x) = \sum_k c_{j,k}\phi_{j,k}(x) + \sum_{\ell\geq j}\sum_k d_{\ell,k}\psi_{\ell,k}(x)$$

Let $\|c_j\|_{\ell_p} = (\sum_k c^p_{j,k})^{1/p}$ and $\|d_\ell\|_{\ell_p} = (\sum_k d^p_{\ell,k})^{1/p}$ with the usual modification for $p = \infty$. The above norm $\|f\|_{B^s_{p,q}}$ is equivalent to the sequence norm

$$\|c_j\|_{\ell_p} + \left\{\sum_{\ell\geq j}\left(2^{\ell(s+1/2-1/p)}\|d_\ell\|_{\ell_p}\right)^q\right\}^{1/q}$$

with the usual modification for $q = \infty$. Note that $B^s_{2,2} = W^s_2$ for $s > 0$.

The relation between the prior parameters and the Besov spaces can be characterized as follows. For $1 \leq p \leq \infty$, the prior sample path of $f(t)$ is in $B^{s-1/2}_{p,\infty}$ $a.s.$ if and only if $\lambda_\ell = O(2^{-2\ell s})$, where $s > 1/2$.

Consider the functional class consisting of sample paths of prior model (31) with a common upper bound $\limsup_{\ell\to\infty} 2^{2\ell s}\lambda_\ell \leq C$ imposed upon prior parameters. Let π denote the induced prior probability measure. We have that $\|f\|_{B^{s-1/2}_{p,\infty}} \leq C$ almost surely. That is, $\pi\{\|f\|_{B^{s-1/2}_{p,\infty}} \leq C\} = 1$. It indicates that we are working over a compact prior functional class.

2. The Bayesian and Empirical Bayesian Approach

The posterior mean of $f(x)$ given the observations y can be calculated in a straightforward manner:

$$E(f(x)|y) = \mu(x) + \sigma^{-2}w^T(x)\, M^{-1}(y - \mu) \tag{32}$$

where $\mu(x) = \sum_{k=1}^{2^j} c_{j,k}\phi_{j,k}(x), \mu = [\mu(x_1), \ldots, \mu(x_n)]^T$, $w(x) = [\mathcal{W}(x, x_1), \ldots, \mathcal{W}(x, x_n)]^T$ with $\mathcal{W}(x, s) = \sum_{\ell=j}^{J-1}\sum_k \lambda_\ell\psi_{\ell,k}(x)\psi_{\ell,k}(s)$, $M = R + \sigma^{-2}W$, and W is an $n \times n$ matrix with the (i, j)th entry given by $\mathcal{W}(x_i, x_j)$. The foregoing posterior mean (32) is the Bayes rule under the squared error loss.

If the coefficients c are not known, one needs to estimate them from the data y. This is known as the empirical Bayes approach. A generalized least squares estimate is proposed to estimate c here, which is

$$\hat{c} = (X^T M^{-1} X)^{-1} X^T M^{-1} y \tag{33}$$

where X is the design matrix for fixed effects. The empirical Bayes estimator becomes

$$\hat{f}(x) = \hat{\mu}(x) + \sigma^{-2}w^T(x)M^{-1}(y - \hat{\mu}) \tag{34}$$

where $\hat{\mu}(x) = \sum_{k=1}^{2^j} \hat{c}_{j,k}\,\phi_{j,k}(x)$ and $\hat{\mu} = [\hat{\mu}(x_1), \ldots, \hat{\mu}(x_n)]^T$.

The preceding empirical Bayes estimator has an asymptotic expression as

$$\hat{f}(x) \approx \hat{f}_L(x) + \hat{f}_{res,DS}(x), \quad x \in [0,1] \tag{35}$$

where

$$\hat{f}_L(x) = \sum_{k=1}^{2^j} \hat{c}_{j,k}\,\phi_{j,k}(x),$$

$$\hat{f}_{res,DS}(x) = \sum_{\ell=j}^{J-1}\sum_{k=1}^{2^\ell} \frac{\lambda_\ell}{\lambda_\ell + \sigma^2/n}\,\hat{d}_{\ell,k}\psi_{\ell,k}(x)$$

and $\hat{c}_{j,k}$ and $\hat{d}_{\ell,k}$ are empirical wavelet coefficients, which can be obtained by the fast algorithm of discrete wavelet transform.

When the parameter values for σ and λ_ℓ are not available, adaptive estimates are necessary. Since

$$E\left(\hat{d}^2_{\ell,k}\right) \approx \lambda_\ell + \frac{\sigma^2}{n}$$

an alternative estimate using diagonal shrinkage rule is given by

$$\begin{aligned}\hat{f}(x) &= \sum_{k=1}^{2^j} \hat{c}_{j,k}\,\phi_{j,k}(x) \\ &\quad + \sum_{\ell=j}^{J-1}\sum_{k=1}^{2^\ell}\left(1-\frac{\sigma^2/n}{\hat{d}^2_{\ell,k}}\right)_+ \hat{d}_{\ell,k}\psi_{\ell,k}(x) \\ &= \sum_{k=1}^{2^j} \hat{c}_{j,k}\,\phi_{j,k}(x) + \sum_{\ell=j}^{J-1}\sum_{k=1}^{2^\ell}\Delta^D\left(d_{\ell,k},\sigma^2/n\right)\psi_{\ell,k}(x)\end{aligned} \tag{36}$$

Often σ^2 is not known, the threshold value can be chosen by generalized cross validation:

$$\hat{f}(x) = \sum_{k=1}^{2^j} \hat{c}_{j,k}\phi_{j,k}(x) + \sum_{\ell=j}^{J-1}\sum_{k=1}^{2^\ell}\Delta^D(d_{\ell,k},\delta_{\mathrm{GCV}})\psi_{\ell,k}(x) \tag{37}$$

3. The BLUP

The estimator (34) turns out to be the unique BLUP in the following sense.

Definition *A predictor $\hat{f}(x)$ is the best linear unbiased prediction (BLUP) for $f(x)$ if and only if (i) $\hat{f}(x)$ is linear in y, (ii) $\hat{f}(x)$ is unbiased in the sense that $E\hat{f}(x) = Ef(x) = \sum_{k=1}^{2^j} c_k\phi_{j,k}(x)$ for all $x \in [0,1]$ and all $c \in R^{2^j}$, (iii) $\hat{f}(x)$ has the minimum mean squared error, among all linear unbiased estimators $\tilde{f}(x)$, i.e., $E(\hat{f}(x)-f(x))^2 \le E(\tilde{f}(x)-f(x))^2$ for all $x \in [0,1]$ and all $c \in R^m$.*

The empirical Bayes estimator $\hat{f}(x)$ is a shrinkage estimator toward $\hat{\mu}(x)$, where $\hat{\mu}(x)$ is the generalized least squares fit of data to the low dimensional subspace H_0 spanned by $\{\phi_{j,k}(x): k = 1,\ldots,2^j\}$. Let $H_{\mathcal{W}}$ be the closure (not the L_2 closure, but rather the closure with respect to the norm given below) of linear space spanned by $\{\psi_{\ell,k}(x): k = 1,\ldots,2^\ell;\ \ell \ge j\}$. The norm in $H_{\mathcal{W}}$ is defined by

$$\left\|\sum_{\ell\ge j}\sum_{k=1}^{2^\ell} d_{\ell,k}\psi_{\ell,k}\right\|^2_{H_{\mathcal{W}}} = \sum_{\ell\ge j}\sum_{k\in\mathcal{J}_\ell} d^2_{\ell,k}/\lambda_\ell$$

where $0/0$ is defined to be zero. The BLUP (34) can be obtained as a penalized least squares estimate.

Theorem 8 *The empirical Bayes estimator $\hat{f}(x)$ in (34) can be also obtained as the solution to the following minimization problem*:

$$\min_{f\in H_0\oplus H_{\mathcal{W}}} n^{-1}(y-f)^T R^{-1}(y-f) + \lambda\|f\|^2_{H_{\mathcal{W}}} \text{ with } \lambda = \frac{\sigma^2}{n} \tag{38}$$

and $f = (f(x_1),\ldots,f(x_n))^T$.

The penalized least squares estimate in (38) is a method of regularization based on the Sobolev norm $\|\cdot\|_{H_{\mathcal{W}}}$. By choosing the order for λ_ℓ, we can control the degree of global smoothness of the estimator $\hat{f}$. When λ_ℓ is of order $O(2^{-2\ell s})$ as $\ell \to \infty$, the posterior space $H_0 \oplus H_{\mathcal{W}}$ is simply $W_2^s[0,1]$, i.e., $\hat{f}$ is in $W_2^s[0,1]$.

The method of regularization in (38) penalizes on random effects. The penalty on random effects is magnified more and more as the resolution level goes finer and finer. Such penalty highly discourages high-frequency wavelet coefficients and hence diminishes high-frequency fluctuation to get smoother posterior curves than the prior curves.

The empirical Bayes estimator $\hat{f}(x)$ (34), or equivalently (38), is the minimax linear estimator of $f(x)$ under mean squared error for f in the functional class

$$W_2^s(\delta) = \left\{f \in W_2^s[0,1], s > 0, \text{ and } \|f\|^2_{H_{\mathcal{W}}} \le 1\right\}$$

D. Cycle Spinning

In view of expressions (22) and (23), the pointwise asymptotic bias and variance depend on their positions relative to grid points, or dyadic points in multiresolution analysis. The oscillatory effects appearing in bias and variance can be removed or lessened by shifting the grid points and then averaging over the shifts.

Consider an averaged shifted kernel

$$\mathcal{V}_h^{(q)}(x,y) = \frac{1}{q}\sum_{\ell=0}^{q-1}\mathcal{V}_h\left(x+\frac{\ell h}{q}, y+\frac{\ell h}{q}\right) \tag{39}$$

The kernel $\mathcal{V}_h^{(q)}(x,y)$ has the same order as the original kernel $\mathcal{V}_h(x,y)$. An estimator based on the averaged shifted kernel is given by

$$\hat{f}_{ASKE,q}(x) = \frac{1}{n}\sum_{i=1}^n \mathcal{V}_h^{(q)}(x,X_i)$$

for density estimation

$$\hat{f}_{ASKE,q}(x) = \frac{1}{n}\sum_{i=1}^n \mathcal{V}_h^{(q)}(x,x_i)y_i$$

for regression estimation

Assume that $f(x) \in C^{m,\alpha}(R) \cap L_2(R)$, for some $0 < \alpha \le 1$, and that $\int_{-\infty}^{\infty}|\mathcal{V}(x,y)(y-x)^{m+\alpha}|\,dy < \infty$. We have, as $h \to 0$ and $n \to \infty$,

$$E\hat{f}_{ASKE,q}(x) - f(x) = \frac{-f^{(m)}(x)}{m!}\frac{1}{q}\left(\sum_{\ell=0}^{q-1} b_m\left(\frac{x}{h}+\frac{\ell}{q}\right)\right)h^m + O(h^{m+\alpha})$$

where $O(h^{m+\alpha})$ is uniform in q. Notice that the number of cycles q can be chosen independently of n and h. By letting $q \to \infty$, we have

$$\frac{1}{q}\left(\sum_{\ell=0}^{q-1} b_m\left(\frac{x}{h}+\frac{\ell}{q}\right)\right) \to \int_0^1 b_m\left(\frac{x}{h}+t\right)dt = 0$$

That is, the asymptotic bias tends to order $O(h^{m+\alpha})$ as the number of cycles q goes to infinity.

As for the variance part, we have

$$\mathcal{V}^{(q)}\left(\frac{x}{h},\frac{x}{h}\right) \to \int_0^1 \mathcal{V}(x,x)\,dx = 1$$

The variance is getting more and more stable in the sense that

$$\sup_{x\in R}\left|\mathcal{V}^{(q)}(x,x) - 1\right| \to 0, \quad \text{as } q \to \infty$$

Therefore, we get

$$\text{Var}\ \hat{f}_{ASKE,q}(x) \le \frac{1}{nh}f(x)K_0^{(q)}\left(\frac{x}{h},\frac{x}{h}\right) + O\left(\frac{1}{n}\right)$$

where $O(n^{-1})$ is uniform in q.

The action of shift-and-average is a smoothing operation on the grid points rather than on the data points. It is well known that when a smoothing operation is applied to the data points, there is a trade-off between bias and variance. However, this is not the case for a smoothing operation applied to grid points. This operation results in bias reduction and is variance stable.

X. TWO-DIMENSIONAL WAVELET ALGORITHM AND IMAGE PROCESSING

Suppose $\{V_j : j \in Z\}$ is a multiresolution analysis of $L_2(R)$ and $\phi(x)$ and $\psi(x)$ are the associated orthonormal scaling function and wavelet. Then the tensor product $\{V_j \otimes V_j : j \in Z\}$ forms a multiresolution analysis of $L_2(R^2)$. The set of functions

$$\{\phi(x-k)\phi(y-k') : k, k' \in Z\}$$

is an orthonormal basis for $V_0 \otimes V_0$. Moreover, the spaces $V_j \otimes V_j$, $V_j \otimes W_j$, $W_j \otimes V_j$ and $W_j \otimes W_j$ have as bases the sets of functions

$$\{\phi_{j,k}(x)\phi_{j,k'}(y) : k, k' \in Z\}$$
$$\{\phi_{j,k}(x)\psi_{j,k'}(y) : k, k' \in Z\}$$
$$\{\psi_{j,k}(x)\phi_{j,k'}(y) : k, k' \in Z\}$$
$$\{\psi_{j,k}(x)\psi_{j,k'}(y) : k, k' \in Z\}$$

respectively. The two-dimensional fast algorithm can be easily formulated from its one-dimensional counterpart. Subscripts H, V, D stand for horizontal, vertical, and diagonal, as the subspaces $V_j \otimes W_j$, $W_j \otimes V_j$, and $W_j \otimes W_j$ tend to emphasize horizontal, vertical, and diagonal features of a two-dimensional image.

The decomposition formula is

$$\begin{aligned} c^{j-1} &= H_c H_r c^j \\ d_H^{j-1} &= G_c H_r c^j \\ d_V^{j-1} &= H_c G_r c^j \\ d_D^{j-1} &= G_c G_r c^j \end{aligned}$$

where $H_r = H$ and $G_r = G$ are quadrature filters acting on each row of the underlying array; similarly, $H_c = H$ and $G_c = G$ are quadrature filters acting on each column of the underlying array. That is,

$$\begin{aligned} \left(H_c H_r c^{j-1}\right)_{m,n} &= \sum_{k\in Z} h_{k-2m}\left(H_r c^j\right)_{k,n} \\ &= \sum_{k\in Z}\sum_{\ell\in Z} h_{k-2m}h_{\ell-2n}c^j_{k,\ell} \\ \left(G_c H_r c^{j-1}\right)_{m,n} &= \sum_{k\in Z} g_{k-2m}\left(H_r c^j\right)_{k,n} \\ &= \sum_{k\in Z}\sum_{\ell\in Z} g_{k-2m}h_{\ell-2n}c^j_{k,\ell} \\ \left(H_c G_r c^{j-1}\right)_{m,n} &= \sum_{k\in Z} h_{k-2m}\left(G_r c^j\right)_{k,n} \\ &= \sum_{k\in Z}\sum_{\ell\in Z} h_{k-2m}g_{\ell-2n}c^j_{k,\ell} \\ \left(G_c G_r c^{j-1}\right)_{m,n} &= \sum_{k\in Z} g_{k-2m}\left(G_r c^j\right)_{k,n} \\ &= \sum_{k\in Z}\sum_{\ell\in Z} g_{k-2m}g_{\ell-2n}c^j_{k,\ell} \end{aligned}$$

The reconstruction formula is

$$c^j = H_c^* H_r^* c^{j-1} + G_c^* H_r^* d_H^{j-1} + H_c^* G_r^* d_V^{j-1} + G_c^* G_r^* d_D^{j-1}$$

where H^* and G^* are adjoint operators and the subscripts c and r indicate respectively the operators acting on each column and each row.

Let $c^0 = (c^0_{m,n})_{r\times c}$ be an $r \times c$ matrix representing a discretized image intensity. The entries $c^0_{m,n}$ are called the pixel values or gray-scale levels. In practice these pixel values are in a finite range. Typical values are $0, 1, \ldots, 255$ (eight bits per pixel).

To compress an image, the image can first be decomposed into wavelet coefficients, then coefficients smaller

than a certain threshold value can be thrown away, and finally the image can be reconstructed with new coefficients.

SEE ALSO THE FOLLOWING ARTICLES

IMAGE PROCESSING • STATISTICS, NON-PARAMETRIC • WAVELETS, INTRODUCTION

BIBLIOGRAPHY

Chui, C. K. (1992). "An Introduction to Wavelets," Academic Press, New York.

Cohen, A., Daubechies, I., Jawerth, B., and Vial, P. (1993a). "Multiresolution analysis, wavelets and fast algorithms on an interval," *C. R. Acad. Sci. Paris series I* **316,** 417–421.

Cohen, A., Daubechies, I., and Vial, P. (1993b). "Wavelets on the interval and fast wavelet transforms," *Appl. Comp. Harmonic Anal*. **1,** 54–81.

Daubechies, I. (1992). "Ten Lectures on Wavelets," CBMS Lecture Notes, No. 61, SIAM, Philadelphia.

Donoho, D. L., and Johnstone, I. M. (1994). "Ideal spatial adaptation by wavelet shrinkage," Biometrika **81,** 425–455.

Donoho, D. L., and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc*. **90,** 1200–1224.

Donoho, D. L., and Johnstone, I. M. (1998). "Minimax estimation via wavelet shrinkage," *Ann. Statist*. **26,** 879–921.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). "Wavelet shrinkage: asymptopia?" *J. Roy. Statist. Soc. Ser. B* **57,** 301–369.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). "Wavelets, Approximation, and Statistical Applications," Lecture Notes in Statistics 129, Springer-Verlag, Berlin.

Huang, S. Y. (1999). "Density estimation by wavelet-based reproducing kernels," *Statistica Sinica* **9,** 137–151.

Huang, S. Y., and Lu, H. H. S. (2000). "Bayesian wavelet shrinkage for nonparametric mixed-effects models," *Statistica Sinica* **10,** 1021–1040.

Kerkyacharian, G., and Picard, D. (1992). "Density estimation in Besov spaces," *Statist. Probab. Lett.* **13,** 15–24.

Kerkyacharian, G., and Picard, D. (1993). "Density estimation by kernel and wavelets methods: optimality of Besov spaces," *Statist. Probab. Lett.* **18,** 327–336.

Mallat, S. G. (1989). "Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$," *Trans. Am. Math. Soc.* **315,** 69–87.

Meyer, Y. (1992). "Wavelets and Operators," Cambridge University Press, Cambridge, U.K.

Meyer, Y. (1993). "Wavelets: Algorithm and Applications," SIAM, Philadelphia.

Vidakovic, B. (1999). "Statistical Modeling by Wavelets," John Wiley & Sons, New York.

Wickerhauser, M. V. (1994). "Adapted Wavelet Analysis from Theory to Software," A. K. Peters, Ltd., Wellesley, MA.



Wavelets, Introduction

Edward Aboufadel
Steven Schlicker
Grand Valley State University

I. Introduction
II. Basic Techniques
III. Two Applications
IV. Signals as Functions
V. The Haar Family of Wavelets
VI. Dilation Equations and Operators
VII. Localization
VIII. Multiresolution Analysis
IX. Daubechies Wavelets
X. Conclusion

GLOSSARY

Compression Manipulating data so it can be stored in a smaller amount of space.

Daubechies wavelets Smooth, orthogonal wavelets, with compact support, that can be used to decompose and recompose polynomial functions with no error.

Detail coefficients Numbers generated from applying a high-pass filter.

Differencing Computing certain linear combinations of components in a signal to create detail coefficients.

Dilation equation An equation relating the father wavelet (scaling function) or mother wavelet to translates and dilations of the father wavelet. Dilation equations generate the coefficients of filters.

Filter Operators on signals that act via averaging and differencing to produce new signals.

Multiresolution analysis A nested sequence of inner product spaces satisfying certain properties. Members of a wavelet family form bases for the spaces in a multiresolution analysis.

Refinement coefficients Coefficients found in a dilation equation.

Thresholding A process in which certain wavelet coefficients are modified in order to achieve an application (e.g., compression or denoising) while preserving the essential structure of a signal.

Wavelets Special functions, consisting of the father wavelet and other functions generated by scalings and translations of a basic function called the mother wavelet.

THE FUNDAMENTAL problem in wavelet research and applications is to find or create a set of basis functions

that will capture, in some efficient manner, the essential information contained in a signal or function. Beginning with a *scaling function* ϕ, we define a *mother wavelet*, ψ, and then *wavelets* $\psi_{m,k}(t) = \psi(2^m t - k)$ generated by scalings and translations of the mother wavelet. The scaling function may be chosen based on many requirements, e.g., regularity, depending on the application. These functions form a basis that provides information about a signal or a function on both a local and global scale. Key ideas in this subject—multiresolution, localization, and adaptability to specific requirements—have made wavelets an important tool in many areas of science.

I. INTRODUCTION

Information surrounds us. Technology has made it easy to collect tremendous amounts of data. The sheer size of the available data makes it difficult to analyze, store, retrieve, and disseminate it. To cope with these challenges, more and more people are using *wavelets*.

As an example, the FBI has more than 25 million cards containing fingerprints. On each card is stored 10 rolled fingerprint impressions, producing about 10 megabytes of data per card. It requires an enormous amount of space just to store all of this information. Without some sort of image compression, a sortable and searchable electronic fingerprint database would be next to impossible. For this purpose, the FBI has adopted a wavelet compression standard for fingerprint digitization. Using this standard, they are able to obtain a compression ratio of about 20:1, without sacrificing the necessary detail required by law.

The history of wavelets began with the development of Fourier analysis. In the early 1800s, Joseph Fourier showed that any periodic function (one that repeats itself like a sine wave) can be represented as an infinite sum of sines and cosines. Within this process, it is possible to approximate such a periodic function as closely as one likes with a finite sum of sines and cosines. To do this, it is necessary to determine the coefficients (amplitudes) of the sines and cosines, along with appropriate translations, in order to "fit" the function being approximated. The ideas of Fourier analysis are used today in many applications.

Although Fourier analysis is a useful and powerful tool, it does have its limitations. The major drawback to Fourier methods is that the basis functions (the sines and cosines) are periodic on the entire real line. As a result, Fourier methods supply global information (i.e., the "big picture") about functions, but not local information (or details). The desire for analytic techniques that provide simultaneous information on both the large and small scale led to the development of wavelets.

The roots of this subject are many. Researchers in different areas such as optics, quantum physics, geology, speech, computer science, and electrical engineering developed wavelet tools to attack problems as diverse as modeling human vision and predicting earthquakes. The main idea was to find basis functions that would allow one to analyze information at both coarse and fine resolutions, leading ultimately to the notion of a *multiresolution analysis*. Work in this area can be traced back to the 1930s. Because of the lack of communication between scientists working in these disparate fields, the tools that were developed in the early years were not seen as being related.

The formal field of wavelets is said to have been introduced in the early 1980s when the geophysicist Jean Morlet developed wavelets as a tool used in oil prospecting. Morlet sought assistance with his work from the physicist Alexander Grossmann, and together the two broadly defined wavelets based in a physical context. In the mid-1980s, Stéphane Mallat and Yves Meyer further developed the field of wavelets by piecing together a broader framework for the theory and constructing the first nontrivial wavelets. In 1987, Ingrid Daubechies used Mallat's work to construct a family of wavelets that satisfied certain conditions (smooth, orthogonal, with compact support) that have become the foundation for current applications of wavelets.

Examples of applications of wavelets today can be found in medical imaging, astronomy, sound and image compression and recognition, and in the studies of turbulence in systems, understanding human vision, and eliminating noise from data.

II. BASIC TECHNIQUES

In this section we will introduce wavelets and discuss how they can be used to process data. The Haar wavelets will be used to illustrate the process.

Any ordered collection of data will be called a *signal*. To obtain an example of a signal, we will collect output from the damped oscillation $\cos(3t)e^{-t}$ on the interval [0,5] at 32 evenly spaced points. This results in the signal

$$\begin{aligned}\mathbf{s} = [&0.763, 0.433, 0.103, -0.160, -0.320, -0.371,\\ &-0.332, -0.235, -0.116, -0.005, 0.077, 0.121,\\ &0.129, 0.108, 0.070, 0.029, -0.008, -0.033,\\ &-0.045, -0.044, -0.034, -0.020, -0.006, 0.006,\\ &0.013, 0.016, 0.015, 0.011, 0.006, 0.001, -0.003,\\ &-0.005].\end{aligned}$$

Note: To save space, all data will be rounded to three decimal places.

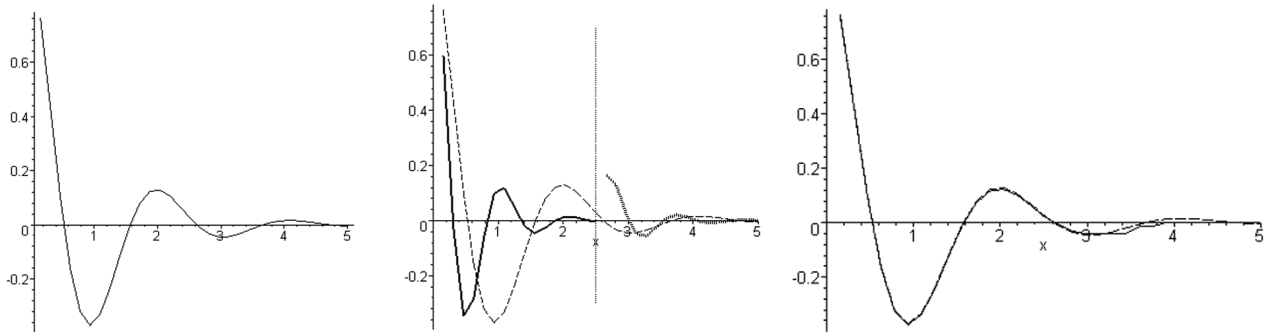


FIGURE 1 (*Left*) Plot of original data. (*Center*) Result of low- and high-pass filters. The plot of the original signal is dashed, the result of the low-pass filter is solid, and the result of the high-pass filter is dotted. (*Right*) Reconstruction of signal after thresholding. The plot of the original signal is dashed; the reconstructed signal is solid.

To graph this signal, we plot t values along the horizonal axis and the corresponding sample values along the vertical axis. We then connect the resulting points with line segments. A plot of this data is shown in Fig. 1.

To process a signal with wavelets, we use *filters* that are determined by the wavelets. There are two components to each filter—a low-pass filter and a high-pass filter. These filters come from what are called the *father* and *mother wavelets* that define a family of wavelets. Each family of wavelets gives us different filters. We will use the Haar wavelets to illustrate.

The *Haar father wavelet*, ϕ, and *Haar mother wavelet*, ψ, are defined by

$$\phi(t) = \begin{cases} 1, & \text{if } 0 \le t < 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{and}$$
$$\psi(t) = \begin{cases} 1, & \text{if } 0 \le t < \frac{1}{2} \\ -1, & \text{if } \frac{1}{2} \le t < 1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The father wavelet is also called the *scaling function*. Note that the father and mother wavelets are related by

$$\begin{aligned} \phi(t) &= \phi(2t) + \phi(2t-1) \quad \text{and} \\ \psi(t) &= \phi(2t) - \phi(2t-1). \end{aligned} \tag{2}$$

The low-pass filter is determined by the father wavelet. This filter uses the coefficients from (2) to perform an averaging of values of a signal. For the Haar wavelets, the two coefficients (1 and 1) are divided by 2 and then used to compute linear combinations of pairs of elements in the signal (later we will see from where the factor of 2 comes). This results in computing an average of pairs of elements. For example, the average of the first pair of data points in our sample is $(0.763 + 0.433)/2 = 0.598$, the average of the second pair of sample points is $(0.103 - 0.161)/2 = -0.029$, and so on. The result is a new signal half the length of the old one. In particular, the result of applying the low-pass filter to our original signal is

$$\begin{aligned} \mathbf{s}_l = [&0.598, -0.029, -0.346, -0.284, -0.061, \\ &0.099, 0.119, 0.050, -0.021, -0.045, -0.027, \\ &0, 0.015, 0.013, 0.004, -0.004] \end{aligned}$$

The high-pass filter is determined by the mother wavelet. This time, instead of adding the data in pairs and dividing by 2, we subtract and divide by 2. This is called *differencing*. For example, using the first two components of $\mathbf{s}$ we obtain $0.763/2 - 0.433/2 = 0.165$, the second pair yields $0.103/2 - 0.161/2 = 0.132$, and so on. When the high-pass filter is applied to the original signal we obtain

$$\begin{aligned} \mathbf{s}_h = [&0.165, 0.132, 0.026, -0.049, -0.056, -0.022, \\ &0.011, 0.021, 0.013, -0.001, -0.007, -0.006, \\ &-0.002, 0.002, 0.003, 0.001] \end{aligned}$$

Note that both $\mathbf{s}_l$ and $\mathbf{s}_h$ are half the length of the original signal. We can think of these shorter signals as signals of length 32 by appending zeros at either the beginning or the end of the signals. In particular, we can identify $\mathbf{s}_l$ with a signal $\mathbf{s}'_l$ of length 32 by attaching 16 zeros to the end of the signal. Similarly, we will view $\mathbf{s}_h$ as a signal $\mathbf{s}'_h$ of length 32 by adding 16 zeros to the beginning of the signal. We attach the zeros in this way so that we can view the result of applying the low- and high-pass filters to $\mathbf{s}$ as $\mathbf{s}' = \mathbf{s}'_l + \mathbf{s}'_h$. We can then plot the signal $\mathbf{s}'$ against the signal $\mathbf{s}$ and compare. A graph is shown in Fig. 1. The result of the low-pass filter is seen on the interval [0,2.5]; the output from the high-pass filter appears on [2.5,5]. We see that the low-pass filter, or the averaging, makes a copy of the original signal but on half scale.

The high-pass filter, or the differencing, keeps track of the detail lost by the low-pass filter. The entries in the signal $\mathbf{s}_h$ obtained from the high-pass filter are called *detail coefficients*.

To see how these coefficients keep track of the details, it is only necessary to see how they are used to recover the original signal. Note that if we have an average $A = \frac{a}{2} + \frac{b}{2}$ of data points a and b and we know the corresponding detail coefficient $C = \frac{a}{2} - \frac{b}{2}$, then we can recover both a and b via $a = A + C$ and $b = A - C$. In this way, the process of applying the low- and high-pass filters is completely reversible.

By applying the low-pass filter, a copy of the original signal $\mathbf{s}_l$ at half scale is created. We now apply the low- and high-pass filters to this new copy to obtain another reduced copy of the original signal at one-quarter scale, plus more detail coefficients. This gives us another new copy of the original signal at one-quarter scale:

$$\mathbf{s}_{ll} = [0.285, -0.315, 0.019, 0.085, -0.033, -0.014, 0.014, 0]$$

plus corresponding detail coefficients

$$\mathbf{s}_{lh} = [0.314, -0.031, -0.080, 0.035, 0.012, -0.014, 0.001, 0.004].$$

Again, we can attach 0's at appropriate ends of these signals to produce a signal

$$\mathbf{s}'_{ll} + \mathbf{s}'_{lh} + \mathbf{s}'_h$$

of length 32.

We can continue with this process until we obtain one number that contains a copy of the original signal at the smallest possible scale. Through this process we have constructed a new signal consisting of the average of all the elements of the original signal, plus all of the detail coefficients. In our example, this new signal is

$$\mathbf{s}_{new} = [0.005, 0.014, -0.034, -0.016, 0.300, -0.033, -0.010, 0.007, 0.314, -0.031, -0.080, 0.035, 0.012, 0.014, 0.001, 0.004, 0.165, 0.132, 0.026, -0.049, -0.056, -0.022, 0.011, 0.021, 0.013, -0.001, -0.007, -0.006, -0.002, 0.002, 0.003, 0.001].$$

This new signal has the same length as the original and contains all of the information the original signal does. In fact, as we discussed above, each step in this processing can be reversed to recover the original signal from this new one. Since no information is lost as a result of applying these filters, this process is called *lossless*. The entries in $\mathbf{s}_{new}$ obtained in this way are called *wavelet coefficients*.

III. TWO APPLICATIONS

A. Thresholding and Compression

Although the preceding process shows a new way to store a signal, it is important to see what benefits this provides. Recall that the detail coefficients are obtained by averaging the differences of pairs of data points. Notice that the data points collected from our damped oscillation are close together in value near the end of the graph. When we process with our high-pass filter, the differencing generates values that are close to 0 in our signal, $\mathbf{s}_{new}$. When the detail coefficients are near 0, there is not much detail in certain parts of the original signal, that is, the signal has fairly constant values there. If we replace all of the detail coefficients that are close to 0 with 0, then reverse the processing with the low- and high-pass filters, we should obtain a signal that is reasonably close to the original. In other words, when the data is similar in small groups like this, we will lose little information in the signal if we replace all of these values that are "close together" with the same value.

Returning to our example, let us assume that "close" to 0 means within 0.01. In other words, replace all of the entries in the processed signal that are closer to 0 than 0.01 with a value of 0. This is called *hard thresholding*. After hard thresholding we obtain the processed signal

$$[0, 0.014, -0.034, -0.016, 0.300, -0.033, 0, 0, 0.314, -0.031, -0.080, 0.035, 0.012, -0.014, 0, 0, 0.165, 0.132, 0.026, -0.049, -0.056, -0.022, 0.011, 0.021, 0.013, 0, 0, 0, 0, 0, 0, 0]$$

To reverse the processing, the first step is to use the final average, 0, along with the final detail coefficient, 0.014, to recover the signal on $\frac{1}{32}$ scale. Note that we introduce some error in this process by rounding. However, we will tolerate this in the interest of saving space in the discussion. This gives us $[0 + 0.014, 0 - 0.014] = [0.014, -0.014]$. The result of our first step in deprocessing the signal is

$$[0.014, -0.014, -0.034, -0.016, 0.300, -0.033, 0, 0, 0.314, -0.031, -0.080, 0.035, 0.012, -0.014, 0, 0, 0.165, 0.132, 0.026, -0.049, -0.056, -0.022, 0.011, 0.021, 0.013, 0, 0, 0, 0, 0, 0, 0]$$

Continuing in this manner, we reconstruct an imperfect copy of the original signal on full scale:

$$[0.759, 0.429, 0.098, -0.166, -0.325, -0.377, -0.338, -0.240, -0.121, -0.009, 0.073, 0.117, 0.127, 0.105, 0.067, 0.025, -0.005, -0.031, -0.042, -0.042, -0.044, -0.044, -0.016, -0.016, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002]$$

If we plot the original data and this reconstructed data after thresholding on the same set of axes, we get the picture in Fig. 1. Notice that it is difficult to tell the two plots apart.

What is the point of thresholding? One important use of wavelets is in the compression of data. If we simply compute wavelet coefficients, we lose no information from our signal, but we usually gain little in the way of storage space. However, if through thresholding we are able to introduce long strings of zeros in our signal, this allows us to store the information in a smaller amount of space. Instead of storing each 0 individually, we set a flag to tell us that we have a string of zeros, then follow that flag with the number of zeros in the string. So instead of storing, say, a string of 10 separate zeros, we only need to use two storage spaces, one for the flag, and one for the number of zeros in the string. In this way, we can reduce the amount of storage space from 10 bytes to 2 bytes. If we are willing to sacrifice some of the original signal, impressive compression results can be obtained through processing and thresholding. Since we lose some information through thresholding, we call this a *lossy* process. It should be noted that hard thresholding is only one of many different thresholding techniques.

B. Noise Removal

Noise, which is unwanted and irrelevant data, often creeps into data sets during collection. A few of the most common types of noise are random noise (which occurs through a signal), localized random noise (which appears in small intervals in the signal), and other noise (for example, pops that occur in old phonograph recordings).

As we have seen, the detail coefficients indicate where important details are located in a data set. If some of these coefficients are very small in relation to others, eliminating them may not substantially alter the original data set. In this manner, wavelets can be used to filter out unwanted noise from a data set. This process is called *denoising*.

In this process we are making the assumption that data we collect is of the form $\mathbf{q} = \mathbf{s} + noise$, where $\mathbf{q}$ is the received signal that is contaminated with noise and $\mathbf{s}$ is the actual signal. We also assume that *noise* is random and that $|noise| < c$ for some constant c. In other words, the noise is uniformly bounded on both sides of the actual signal. Our goal is to recover the signal $\mathbf{s}$.

As an example, we have artificially introduced noise onto the piecewise constant function f, with values of 2 on (0,1], 4 on (1,2], and 2 on (2,3], and then sampled at 1024 points to obtain a signal $\mathbf{s}$. (Note: The "noise" introduced here is defined as $0.01t\sin(150t)$ on (0,1], $0.01\sin(150t)$ on (1,2], and $0.01(t-2.5)^2\sin(150t)$ on (2,3].) The data is shown in Fig. 2.

We process this signal to obtain the signal $\mathbf{s}_{new}$, which consists of wavelet coefficients. A plot of $\mathbf{s}_{new}$ is shown in Fig. 2. The vertical axis in this figure is scaled so that the noise can be clearly seen. Other detail coefficients do not completely fit in this window.

Since the actual signal $\mathbf{s}$ is mostly constant, we should expect the majority of detail coefficients to be 0. However, the added noise makes many of these entries small, but nonzero. If we apply thresholding to eliminate what appears to be the added noise, we should be able to recover the signal $\mathbf{s}$. After applying hard thresholding with a tolerance $\lambda = 0.0075$ and reversing the processing, we obtain the "denoised" data shown in Fig. 2.

Similar ideas may be used to restore damaged video, photographs, or recordings, or to detect corrosion in metallic equipment. This approach has also been used in the correction of seismic data in Japan. However, it should be noted here that the Haar wavelets were effective in denoising the signal in this section because of the piecewise constant nature of the signal. Later we will see examples of other types of wavelets that are better choices for the applications mentioned here.

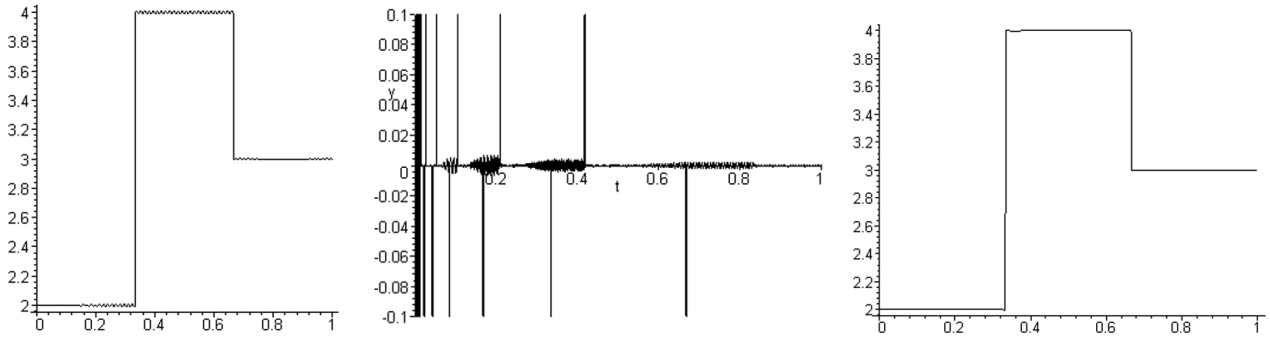


FIGURE 2 (*Left*) Piecewise constant signal with added noise. (*Center*) Wavelet coefficients of noisy data. (*Right*) Reconstructed piecewise constant signal after thresholding.

IV. SIGNALS AS FUNCTIONS

When working with wavelets it is convenient to describe our data in terms of functions. There is a simple way to do this, which we describe in this section.

In the previous section we began with an initial signal, **s**, of length 32 or 2^5. Note that this is only one of infinitely many signals of length 32. For convenience, we will denote the collection of all signals of length 2^n by $\mathbb{R}^{2^n}$ and write $\mathbf{s} \in \mathbb{R}^{32}$.

There is a natural connection between the space $\mathbb{R}^{2^n}$ of signals and the space of real-valued functions. We can think of each signal in $\mathbb{R}^{2^n}$ as representing a piecewise defined function on the interval [0,1] as follows. Break the interval into 2^n subintervals of equal length. The function identified with the signal has constant values on each subinterval, with the value of the function on the ith subinterval given by the ith entry of the signal. For example, the signal [1,2,3,4] in $\mathbb{R}^4$ corresponds to the function f that has values of 1 on $[0, \frac{1}{4})$, 2 on $[\frac{1}{4}, \frac{1}{2})$, 3 $[\frac{1}{2}, \frac{3}{4})$, 4 on $[\frac{3}{4}, 1)$, and 0 elsewhere.

We will denote the collection of all piecewise constant functions on such intervals of length $\frac{1}{2^n}$ by V_n. Our signal from the first section is then viewed as a function in V_5. Similarly, the father and mother Haar wavelets (1) can be viewed as functions in V_2 and identified with the vectors [1,1,1,1] and [1,1,−1,−1] in $\mathbb{R}^4$.

It is important to make the connection between the standard basis for $\mathbb{R}^{2^n}$ and the corresponding basis for V_n. Since V_0 consists of the constant functions on the interval [0,1], the father wavelet, $\phi(t)$, will span V_0. In V_1, we consider functions that are piecewise constant on the intervals $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$. The standard basis for V_1 is then $\{\phi(2t), \phi(2t-1)\}$. In general, we can see that the set $S_n = \{\phi(2^n t - k) : 0 \le k < 2^n - 1\}$ will be a basis for V_n corresponding to the standard basis for $\mathbb{R}^{2^n}$.

V. THE HAAR FAMILY OF WAVELETS

In the previous sections we defined the father and mother Haar wavelets and saw how these wavelets were used to define two filters that we used to process data. In this section we will introduce the Haar family of wavelets and explain how the functional view of data discussed in the previous section actually determines the filters.

The function setting is the natural one in which to study wavelets. To see why, we must view each of the sets V_n as defined in the previous section as an *inner product space*.

Definition Let V be a vector space. An *inner product* on V is a function that assigns to each pair of vectors $\mathbf{u}$, $\mathbf{v}$ in V a real number, denoted $\langle \mathbf{u}, \mathbf{v} \rangle$, satisfying the following:

1. $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$ for all $\mathbf{u}, \mathbf{v} \in V$.
2. $\langle k\mathbf{u}, \mathbf{v} \rangle = k\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, k\mathbf{v} \rangle$ for all $\mathbf{u}, \mathbf{v} \in V$ and $k \in \mathbb{R}$.
3. $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$.
4. $\langle \mathbf{v}, \mathbf{v} \rangle \ge 0$ for all $\mathbf{v} \in V$ with equality if and only if $\mathbf{v} = 0$.

If V is a vector space on which an inner product, $\langle , \rangle$, is defined, we call the pair $(V, \langle , \rangle)$, an *inner product space*. A familiar example of an inner product space is $\mathbb{R}^n$ with the dot product as inner product.

Each V_n can be considered an inner product space using the inner product $\langle f, g \rangle = \int_{-\infty}^{\infty} f(t)g(t)\,dt$.

It is important to notice that signals and Haar wavelets have values of 0 outside the interval [0,1]. (Functions that are 0 outside of a closed and bounded interval are said to have *compact support*.) Each finite signal **s** corresponds to a function $f_{\mathbf{s}}$ satisfying $\int_{-\infty}^{\infty} f_{\mathbf{s}}(t)^2\,dt < \infty$. So, each V_n is a subspace of the larger inner product space $L^2(\mathbb{R})$, which consists of those functions $f : \mathbb{R} \to \mathbb{R}$ whose *norm*

$$\|f\| = \langle f, f \rangle^{1/2} = \left(\int_{-\infty}^{\infty} f(t)^2\,dt \right)^{1/2} \tag{3}$$

is finite.

Inner product spaces are useful in that we can measure lengths of vectors and angles between vectors in these spaces. We measure lengths of functions in V_n using the norm given in (3). Although we are not usually interested in angles between functions, we are interested in *orthogonality*. Recall that in the inner product space $\mathbb{R}^n$, using the dot product as inner product, two vectors **u** and **v** are perpendicular if $\mathbf{u} \cdot \mathbf{v} = 0$. Orthogonality is a generalization of this idea. In an inner product space V, two vectors **u** and **v** are *orthogonal* if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$. As we will see, orthogonality is a very useful tool when computing projections.

Now we return to wavelets. Recall that we defined the father and mother wavelets by the equations in (1). As the terminology suggests, the father and mother generate "children." These children are determined by *scalings* and *translations* of the parents. In general, the nth-*generation children* are defined by

$$\psi_{n,k}(t) = \psi(2^n t - k), \quad 0 \le k \le 2^n - 1 \tag{4}$$

Note that there are 2^n children in the nth generation. The graphs of each of these children look like compressed and translated copies of the mother wavelet.

In the previous section we saw that there is a correspondence between the vector spaces V_n and $\mathbb{R}^{2^n}$. Consequently, the dimension of V_n is 2^n. In fact, we saw earlier that S_n is the "standard" basis for V_n. Another useful basis for V_n is

$$B_n = \left\{ \phi, \psi, \psi_{m,k} : 0 \le k \le 2^m - 1, 1 \le m \le n - 1 \right\}$$

which consists of father, mother, and children wavelets. The basis B_n is important in the world of wavelets precisely because the elements in B_n *are* the wavelets. The basis B_n also has the property that $\langle f, g\rangle = \int_{-\infty}^{\infty} f(t)g(t)\,dt = 0$ for any $f, g \in B_n$, with $f \neq g$. In other words, B_n is an *orthogonal basis* for V_n.

Any time we have an orthogonal basis for a subspace of an inner product space, we can *project* any vector in the space onto the subspace. What is more, there is an elegant way to do this given by the *orthogonal decomposition theorem*.

The orthogonal decomposition theorem If $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\}$ is an orthogonal basis for a finite-dimensional subspace W of an inner product space V, then any $\mathbf{v} \in V$ can be written uniquely as $\mathbf{v} = \mathbf{w} + \mathbf{w}_\perp$, with $\mathbf{w} \in W$. Moreover,

$$\mathbf{w} = \frac{\langle \mathbf{v}, \mathbf{w}_1\rangle}{\langle \mathbf{w}_1, \mathbf{w}_1\rangle}\mathbf{w}_1 + \cdots + \frac{\langle \mathbf{v}, \mathbf{w}_k\rangle}{\langle \mathbf{w}_k, \mathbf{w}_k\rangle}\mathbf{w}_k = \sum_{i=1}^{k} \frac{\langle \mathbf{v}, \mathbf{w}_i\rangle}{\langle \mathbf{w}_i, \mathbf{w}_i\rangle}\mathbf{w}_i$$

and $\mathbf{w}_\perp = \mathbf{v} - \mathbf{w}$. The vector $\mathbf{w}$ in the orthogonal decomposition theorem is called the *projection of* $\mathbf{v}$ *onto* W.

This theorem is important in many respects. One way to use this theorem is for approximations. The vector $\mathbf{w}$ determined by the theorem is the "best" approximation to the vector $\mathbf{v}$ by a vector in W, in the sense that $\mathbf{w}_\perp$ is orthogonal to W. In fact, $\mathbf{w}_\perp$ is orthogonal to every vector in W. The collection of all vectors in V that are orthogonal to every vector in W is a subspace of V. This subspace is called the *orthogonal complement of* W *in* V and is denoted $W^\perp$. We then represent the result of the orthogonal decomposition theorem as $V = W \oplus W^\perp$.

To see how this applies to wavelets, let us return to our earlier example:

$$\begin{aligned}\mathbf{s} = [&0.763, 0.433, 0.103, -0.160, -0.320, -0.371,\\ &-0.332, -0.235, -0.116, -0.005, 0.077, 0.121,\\ &0.129, 0.108, 0.070, 0.029, -0.008, -0.033,\\ &-0.045, -0.044, -0.034, -0.020, -0.006, 0.006,\\ &0.013, 0.016, 0.015, 0.011, 0.006, 0.001, -0.003,\\ &-0.005]\end{aligned}$$

Recall that $\mathbf{s}$ is an element of V_5. Consider the subspace of V_5 spanned by the set $C_4 = \{\psi_{4,k} : 0 \le k \le 2^4 - 1\}$. Now C_4 is an orthogonal basis for the space it spans. Let W_4 be the span of C_4. Using this subspace W_4, let us compute the projection of $\mathbf{s}$ onto the space W_4 as described in the orthogonal decomposition theorem. Call this vector $\mathbf{w}$. First, note that $\langle \psi_{4,k}, \psi_{4,k}\rangle = \frac{1}{2^4}$ for any value of k. Let $\mathbf{s}_i$ denote the ith component of $\mathbf{s}$. Now notice that $\langle \mathbf{s}, \psi_{4,k}\rangle = (\mathbf{s}_{2k+2} - \mathbf{s}_{2k+1})/2^5$ for each k. Then the coefficient of $\psi_{4,k}$ in the orthogonal decomposition theorem is

$$\frac{\langle \mathbf{s}, \psi_{4,k}\rangle}{\langle \psi_{4,k}, \psi_{4,k}\rangle} = \frac{\frac{\mathbf{s}_{2k+1} - \mathbf{s}_{2k+2}}{2^5}}{\frac{1}{2^4}} = \frac{\mathbf{s}_{2k+1} - \mathbf{s}_{2k+2}}{2}$$

which is a detail coefficient as described earlier. The factor of 2 occurs naturally as a result of the squares of the norms of our basis vectors. Let $a_{4,k}$ represent the coefficient of $\psi_{4,k}$ in $\mathbf{w}$, so that $a_{4,k} = (\mathbf{s}_{2k+2} - \mathbf{s}_{2+1})/2$. We then have

$$\mathbf{w} = \sum_{k=0}^{15} a_{4,k}\psi_{4,k} \tag{5}$$

We must treat this projection with caution. Notice that we have only 16 coefficients in (5). As a result, we can treat $\mathbf{w}$ as a signal in $\mathbb{R}^{16}$. In this sense, the coefficients we calculated when projecting $\mathbf{s}$ onto W_4 are exactly those we obtained from applying the high-pass filter in the first section. In other words, we can say $\mathbf{w} = \mathbf{s}_h$. When using the orthogonal decomposition theorem, however, we must view this projection as an element of the larger space, or V_5 in this case. To do this, recall that the nonzero components of the vectors in $\mathbb{R}^{32}$ that correspond to $\psi_{4,k} \in V_5$ are 1 and -1. When viewed in this way, the vector in $\mathbb{R}^{32}$ corresponding to $\mathbf{w}$ (from (5)) will have components

$$[a_{4,0}, -a_{4,0}, a_{4,1}, -a_{4,1}, \ldots, a_{4,15}, -a_{4,15}]$$

This will be important when we compute $\mathbf{w}_\perp$. (This is different from how we earlier extended $\mathbf{s}_h \in \mathbb{R}^{16}$ to a signal in $\mathbb{R}^{32}$. In that situation, we were interested in interpreting our results graphically. In this case, we are guided by the orthogonal decomposition theorem.)

Now let us compute $\mathbf{w}_\perp = \mathbf{s} - \mathbf{w}$. Computing the first two components of $\mathbf{w}_\perp$ yields

$$\mathbf{s}_1 - \mathbf{w}_1 = \mathbf{s}_1 - a_{4,0}\psi_{4,0}(t) = \mathbf{s}_1 - \left(\frac{\mathbf{s}_1 - \mathbf{s}_2}{2}\right) = \frac{\mathbf{s}_1 + \mathbf{s}_2}{2}$$

and

$$\mathbf{s}_2 - \mathbf{w}_2 = \mathbf{s}_2 + a_{4,0}\psi_{4,0}(t) = \mathbf{s}_2 + \left(\frac{\mathbf{s}_1 - \mathbf{s}_2}{2}\right) = \frac{\mathbf{s}_1 + \mathbf{s}_2}{2}$$

The remaining components can be computed in a similar manner. Looking at the result componentwise, we identify $\mathbf{w}_\perp$ with the signal

$$\left[\frac{\mathbf{s}_1 + \mathbf{s}_2}{2}, \frac{\mathbf{s}_1 + \mathbf{s}_2}{2}, \frac{\mathbf{s}_3 + \mathbf{s}_4}{2}, \frac{\mathbf{s}_3 + \mathbf{s}_4}{2}, \ldots, \frac{\mathbf{s}_{31} + \mathbf{s}_{32}}{2}, \frac{\mathbf{s}_{31} + \mathbf{s}_{32}}{2}\right]$$

In other words, if we ignore the duplication, we can view $\mathbf{w}_\perp$ as the result of applying the low-pass filter to the signal.

Recall that the averaging process (the low-pass filter) produces a copy of the original signal on half scale. So we can consider the copy, $\mathbf{s}_l$, of our signal as a function in V_4. It can be shown that each basis element in B_4 is orthogonal

to the functions in V_4. Consequently, every function in W_4 is orthogonal to every function in V_4. So V_4 and W_4 are orthogonal complements in V_5. In other words, $W_4 = V_4^{\perp}$ and $V_5 = V_4 \oplus V_4^{\perp}$.

The next step in processing is to decompose the new signal $\mathbf{s}_l$ in V_4. When we apply the orthogonal decomposition theorem to $\mathbf{s}_l$, we decompose it into two pieces, one in V_3 and one in $V_3^{\perp}$ as subspaces of V_4. This gives us another decomposition of $\mathbf{s}$ in $V_3 \oplus V_3^{\perp} \oplus V_4^{\perp}$. Again, this decomposition uses the idea that we are identifying V_3 as subspace of V_4, which is a subspace of V_5.

This process continues until we have obtained a vector in

$$V_0 \oplus V_0^{\perp} \oplus V_1^{\perp} \oplus V_2^{\perp} \oplus V_3^{\perp} \oplus V_4^{\perp}$$

In our earlier work, this was the signal $\mathbf{s}_{new}$. Since the coefficients that are determined in this process are multipliers on wavelets, we can see why they are called wavelet coefficients. We should note here that the process described in the previous sections of constructing new signals through averaging and differencing has involved successively producing versions of the original signal on half scales. Since we continually reduce the scale by half, the process can only be applied to signals of length 2^n some integer n.

VI. DILATION EQUATIONS AND OPERATORS

In this section we will see how wavelet processing can be viewed as *operators* that arise from *dilation equations*. Again, we use the Haar wavelets to motivate the discussion.

The processing we have done has depended only on the equations (2). These equations are called *dilation, equations*, and they completely determine the process through which the wavelet coefficients are found. In many situations, the properties that we desire our wavelets to have determine the dilation equations for the wavelets.

Once we have dilation equations for a family of wavelets, we can use them to describe low- and high-pass filters for that family. To make computations more efficient, however, we *normalize* the functions in equations (2). Recall that a vector is a *normal vector* if its norm or length is 1. In the case of our wavelets, we use the $L^2(\mathbb{R})$ integral norm defined in (3). A simple integration by substitution shows us that $\|\phi(2t)\| = \sqrt{2} = \|\phi(2t-1)\|$.

Multiplying $\phi(2t)$ and $\phi(2t-1)$ in the original dilation equations by $\sqrt{2}$ to normalize them produces the dilation equations

$$\begin{aligned} \phi(t) &= h_0\sqrt{2}\phi(2t) + h_1\sqrt{2}\phi(2t-1) \quad \text{and} \\ \psi(t) &= g_0\sqrt{2}\phi(2t) - g_1\sqrt{2}\phi(2t-1) \end{aligned} \tag{6}$$

where $h_0 = h_1 = \frac{1}{\sqrt{2}}$, and $g_0 = \frac{1}{\sqrt{2}}$ and $g_1 = -\frac{1}{\sqrt{2}}$ for the Haar wavelets.

Suppose we have a signal $\mathbf{s} = [\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_{2^n-1}]$ of length 2^n. The low- and high-pass filters can then be described in general by two *operators*, H (the low-pass operator) and G (the high-pass operator), defined by

$$(H\mathbf{s})_k = \sum_{j\in\mathbb{Z}} h_{j-2k}\mathbf{s}_j \quad \text{and} \quad (G\mathbf{s})_k = \sum_{j\in\mathbb{Z}} g_{j-2k}\mathbf{s}_j \tag{7}$$

The coefficients h_i and g_i in these sums are called *filter coefficients*. There are two items to be aware of here:

- Note the translation in the subscripts of h and g by $2k$. If we do a standard convolution (translate by k), the entries we want occur in every other component. We saw this happen earlier when we used the orthogonal decomposition theorem. Hence, we *downsample* at $2k$.
- It might seem natural that the high-pass operator should be denoted by H, but the convention is to use G instead.

Notice that for the Haar wavelets these operators give us exactly the results (up to a factor of $\sqrt{2}$) that we obtained in the previous section. The importance of this operator approach is that to process signals with wavelets we only need to know the coefficients in the dilation equations. As we will see later, other wavelet families have dilation equations whose coefficients are different from those of the Haar wavelets. However, the method of processing signals with these wavelet families is the same.

To undo the processing with H and G, we define the *dual operators* denoted H^* and G^*. The dual operators are defined by

$$(H^*\mathbf{s}^*)_k = \sum_{j\in\mathbb{Z}} h_{k-2j}\mathbf{s}_j^* \quad \text{and} \quad (G^*\mathbf{s}^*)_k = \sum_{j\in\mathbb{Z}} g_{k-2j}\mathbf{s}_j^* \tag{8}$$

Note the change in indices from the definitions of H and G.

It is not difficult to see that $H^*(H\mathbf{s}) + G^*(G\mathbf{s}) = \mathbf{s}$ for any signal $\mathbf{s}$ of length 2^n using the Haar wavelets. However, the orthogonality plays a critical role in this process (recall how orthogonality arose in the projections onto the subspaces V_n and $V_n^{\perp}$), so these operators will not perform in the same way if the wavelets we use are not orthogonal. Also, the operators defined by (7) are designed to process signals of infinite length, so there may be difficulties in recomposing a finite signal at either end if the number of nonzero coefficients in the dilation equations is different from 2. This is called a *boundary problem*, and there are several methods to cope with such difficulties. We will discuss three such methods in a later section.

VII. LOCALIZATION

Another important property of wavelets is their ability to identify important information on small scale. This is called *localization*.

In (4) we defined the nth-generation children wavelets as contractions and translations of the mother wavelet. By selecting a large enough value of n, we are able to analyze our data on very small intervals. There is, however, no reason to restrict ourselves to small intervals. In fact, the beauty and power of wavelets lies in their ability to perform analysis on both large and small scales simultaneously. To see how this works, we extend the definitions of the previously defined sets V_n to include negative values of n as well as to cover functions defined (and nonzero) outside of [0,1].

Recall that for nonnegative integer values of n, V_n consists of the functions that are piecewise constant on intervals of length $1/2^n$ within the interval [0,1]. If we allow for translations outside of the interval [0,1], then we can extend the definition of each V_n. We will need to be careful, however, to insist that each of our functions remains in $L^2(\mathbb{R})$. With this in mind, V_0 will be the space of all functions in $L^2(\mathbb{R})$ that are piecewise constant on intervals of length 1 with integer end points. The space V_0 will be generated by translations of the father wavelet of the form $\phi(t-k)$ for $k \in \mathbb{Z}$. Next, V_1 will include all functions in $L^2(\mathbb{R})$ that are piecewise constant on intervals of length $\frac{1}{2}$ with possible breaks at integer points or at points of the form $(2n+1)/2$ from some integer n. We can generate functions in V_1 with the collection $\{\phi(2t-k): k \in \mathbb{Z}\}$. Similarly, V_2 will be spanned by $\{\phi(2^2t-k): k \in \mathbb{Z}\}$ and will contain all functions with possible breaks at rational points with denominators of $4 = 2^2$, and so on.

This perspective allows us to define V_n for negative values of n as well. In these cases, instead of contracting the intervals on which our functions are constant, we expand them. For example, V_{-1} will consist of all functions in $L^2(\mathbb{R})$ that are piecewise constant on intervals of length 2, with integer end points. More generally, V_n will contain piecewise constant functions in $L^2(\mathbb{R})$ with possible jumps at points of the form $m \times 2^{-n}$ for any integer n. The set $\{\phi(2^n t-k): k \in \mathbb{Z}\}$ will generate V_n. An advantage to this approach is that we can now work with data that is defined on intervals other than [0,1].

To illustrate this advantage, consider the methods of Fourier analysis, which were discussed in the Introduction. As was previously stated, one disadvantage of Fourier methods is that they produce *global* information about signals, but not *local* information. For instance, if a low-pass filter, based on Fourier analysis, is applied to a signal, then high frequencies are eliminated throughout the whole signal, which may not be desired. The global approach of Fourier methods can also be seen in the formula for the discrete Fourier transform, where *every* data point contributes to *every* Fourier coefficient.

A wavelet analysis of a signal will bring out local information. As an example, consider the signal represented on the left in Fig. 3. The key behavior of this signal can be found on the intervals $[-4, -2]$ and $[1, 1.5]$. If we analyze this signal with wavelets, then specific wavelets will pick up the information on specific intervals. For instance, wavelet coefficients corresponding to the Haar wavelet $\psi(2^{-1}t+2) \in V_0$, along with scalings of this wavelet that are in $V_1, V_2, \ldots$, will be nonzero, because of the impulse on the interval $[-4, -2]$. Similarly, $\psi(2t-1) \in V_2$ and its scalings in $V_3, V_4, \ldots$ will notice the impulse on the other interval. All other wavelet coefficients will be zero.

Consequently, translations of ψ can be used to find specific time intervals of interest, and the scalings of these translations will reveal finer and finer details. This idea is represented on the right in Fig. 3. There, the shaded regions correspond to the wavelets we can use to analyze this function.

VIII. MULTIRESOLUTION ANALYSIS

In this section we will introduce the idea of a *multiresolution analysis* within the context of the Haar wavelets.

Let us return to our example processing a signal. There we applied the low- and high-pass filters to our original

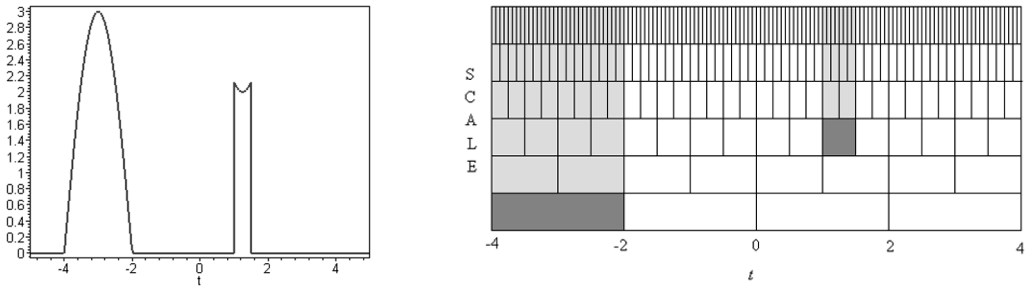


FIGURE 3 (*Left*) A signal with two short impulses. (*Right*) Selecting wavelets of different scales to analyze a signal.

signal $\mathbf{s}$ to obtain two new signals, $\mathbf{s}_l$ and $\mathbf{s}_h$, each in $\mathbb{R}^{16}$. They (or, consequently, the corresponding functions in V_4) can be thought of in at least two different ways as belonging to $\mathbb{R}^{32}$ (or V_5). With such inclusions, we can say that $\mathbb{R}^{16}$ is a subset of $\mathbb{R}^{32}$ ($\mathbb{R}^{16} \subset \mathbb{R}^{32}$) or, in the function setting, V_4 is a subset of V_5 ($V_4 \subset V_5$). After the first round of processing with the Haar wavelets, the original signal $\mathbf{s}$ in $\mathbb{R}^{32}$ is decomposed into two signals in $\mathbb{R}^{32}$.

The second round of processing consists of applying the low- and high-pass filters to $\mathbf{s}_l$, which yields two new signals $\mathbf{s}_{ll}$ and $\mathbf{s}_{lh}$ in $\mathbb{R}^8$. Again, each of these signals can be identified with signals in $\mathbb{R}^{16}$ and $\mathbb{R}^{32}$ by an appropriate extension. At each stage of the processing, new signals on half scale are obtained. The processing ends when we are reduced to a single number, a signal in $\mathbb{R}$.

In the function setting, when we process the signal with the low-pass filter, a copy of the corresponding function from V_5 is produced on half scale in V_4. Continued processing constructs additional copies on smaller scales in V_3, then V_2, and so on until V_0 is reached. In other words, we are reproducing our function in a nested sequence of spaces,

$$V_0 \subset V_1 \subset V_2 \subset V_3 \subset V_4 \subset V_5$$

Of course, there is no reason to restrict ourselves to signals of a given length. We can expand this idea to build a nested collection of sets

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset V_3 \subset V_4 \cdots$$

that continues for as long as we need in either direction. It can also be shown that the set $\{\phi(t-k)\}_{k\in\mathbb{Z}}$ forms an orthogonal basis for V_0. These are the basic ideas behind *multiresolution analysis*.

We have so far only encountered the Haar wavelets, but there are many other families of wavelets. Wavelets that are typically used in applications are constructed to satisfy certain criteria. The standard approach is to first build a multiresolution analysis (MRA) and then construct the wavelet family with the desired criteria from the MRA. All of the features of the Haar wavelets can be seen in the following definition.

Definition A *multiresolution analysis* (MRA) is a nested sequence

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots$$

of subspaces of $L^2(\mathbb{R})$ with a scaling function ϕ such that

1. $\bigcup_{n\in\mathbb{Z}} V_n$ is dense in $L^2(\mathbb{R})$
2. $\bigcap_{n\in\mathbb{Z}} V_n = \{0\}$
3. $f(t) \in V_n$ if and only if $f(2^{-n}) \in V_0$
4. $\{\phi(t-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis for V_0 (that is, $\{\phi(t-k)\}_{k\in\mathbb{Z}}$ is an orthogonal basis for V_0 in which $\|\phi(t-k)\| = 1$ for each $k \in \mathbb{Z}$)

In every multiresolution analysis there is a *dilation equation* of the form

$$\phi(t) = \sum_k c_k \phi(2t-k) \tag{9}$$

The constants c_k in a dilation equation are called *refinement coefficients*. Equations (2) are examples of dilation equations. As we will see later, dilation equations allow us to construct a wide variety of different types of wavelets.

IX. DAUBECHIES WAVELETS

A. Dilation Equation for Daubechies Wavelets

The idea of a scaling function having compact support is featured in a type of wavelet family that is named after Ingrid Daubechies, who pioneered their development. The scaling functions for *Daubechies wavelets* have compact support and, more importantly, can be used to accurately decompose and recompose polynomial functions. This property of Daubechies wavelets is called the *regularity property*, and it is what makes these wavelets special.

There is more than one family of Daubechies wavelets, and one of the simplest families is generated by the scaling function called D_4. This function has compact support $[0,3]$, and because of this compact support, only four of the refinement coefficients are nonzero (hence the subscript 4). In other words, we can write

$$D_4(t) = c_0 D_4(2t) + c_1 D_4(2t-1) + c_2 D_4(2t-2) + c_3 D_4(2t-3) \tag{10}$$

The values of the four refinement coefficients are determined by using the fourth property of MRAs, an averaging condition for the scaling function, and a regularity condition. The orthogonality of the translates of D_4 yields the following two equations:

$$c_0^2 + c_1^2 + c_2^2 + c_3^2 = 2 \quad \text{and} \quad c_0 c_2 + c_1 c_3 = 0 \tag{11}$$

The *averaging condition* for the scaling function is simply that its average value over the real line is 1. This leads to the equation

$$c_0 + c_1 + c_2 + c_3 = 2 \tag{12}$$

The regularity condition for D_4 will lead to two more equations. The condition states that constant and linear functions can be reproduced by D_4 and its translates. More specifically, for any real numbers α and β, there exists a sequence of coefficients $\{a_k\}$ such that $\alpha t + \beta = \sum_k a_k D_4(t-k)$. This ability to reproduce polynomials

is referred to as *regularity* or *smoothness*, and it distinguishes the different families of Daubechies wavelets. For instance, the regularity condition for D_6, Daubechies scaling function with six refinement coefficients, is that constant, linear, and quadratic functions can be perfectly reproduced.

In some places in the literature, the regularity condition is presented in terms of the mother wavelet ψ that corresponds with D_4, and the following *vanishing moment conditions* are stated instead: $\int_{-\infty}^{\infty} \psi(t)\,dt = 0$ and $\int_{-\infty}^{\infty} t\psi(t)\,dt = 0$. It should be clear that for the family that is generated by D_6, there is a third moment condition to be included: $\int_{-\infty}^{\infty} t^2\psi(t)\,dt = 0$.

Regularity is important for several reasons. First, in image processing, smooth images, where colors and shades change gradually rather than suddenly, are better analyzed with smoother wavelets, as edge effects and other errors are avoided. Also, smoother wavelets lead to a cleaner separation of signals into low-pass and high-pass pieces. Finally, numerical methods such as the fast wavelet transform have been found to work more efficiently with wavelets that are more regular.

The regularity condition for D_4 leads to these two equations:

$$-c_0 + c_1 - c_2 + c_3 = 0 \quad \text{and} \quad -c_1 + 2c_2 - 3c_3 = 0 \tag{13}$$

The system of equations (11), (12), and (13) has two solutions. By convention, we use the solution that gives this dilation equation:

$$D_4(t) = \frac{1+\sqrt{3}}{4} D_4(2t) + \frac{3+\sqrt{3}}{4} D_4(2t-1) + \frac{3-\sqrt{3}}{4} D_4(2t-2) + \frac{1-\sqrt{3}}{4} D_4(2t-3) \tag{14}$$

See Fig. 4 for a graph of D_4. The other solution of the system of equations leads to a function whose graph is a mirror image of this graph, which is why we can ignore it.

The scaling function D_4 has a distinctive jagged shape. Although it may be difficult to label D_4 as *smooth*, it is continuous, and given the regularity condition, it should not come as a surprise that as n gets larger, D_n itself looks smoother and smoother. The method used to create this graph is described in the next section.

B. The Cascade Algorithm

The Haar wavelets are the simplest, and least regular, type of Daubechies wavelets, reproducing constant functions only. Also, the Haar scaling function is the only Daubechies scaling function that can represented with a simple formula. For other Daubechies wavelets families, the scaling functions are generated from the dilation equation using the *cascade algorithm*.

The cascade algorithm is based on the dyadic scaling and dilation equations (9) that are at the heart of an MRA. The equation (14) is another dilation equation. These equations can be used to determine function values at smaller scales, if function values at larger scales are known. For instance, if ϕ is known at all multiples of $\frac{1}{2}$, then (9) can be used to determine the value of ϕ at all multiples of $\frac{1}{4}$.

The cascade algorithm is a fixed-point method. Assuming the refinement coefficients are known, we can define a mapping from functions to functions by

$$F(u(t)) = \sum_k c_k u(2t-k). \tag{15}$$

Fixed points of F will satisfy (9), and it is possible to find these fixed points through an iterative process.

We will demonstrate how this works on (14), and consequently generate D_4. To begin, create an "initial guess" for a fixed point of (15), called u_0, defined only on the integers. Let u_0 be this guess: The function is zero on all of the integers except that $u_0(0) = 1$. Then, to get a good picture, connect these points with line segments, as is done is Fig. 5. (This is a reasonable first guess for D_4, as we know the scaling function has compact support and reaches some sort of maximum on the interval $[0,3]$. Connecting the

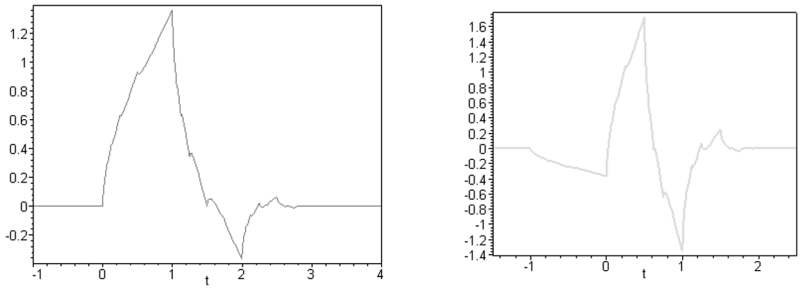


FIGURE 4 (*Left*) Scaling function D_4 of Daubechies. (*Right*) Daubechies' mother wavelet.

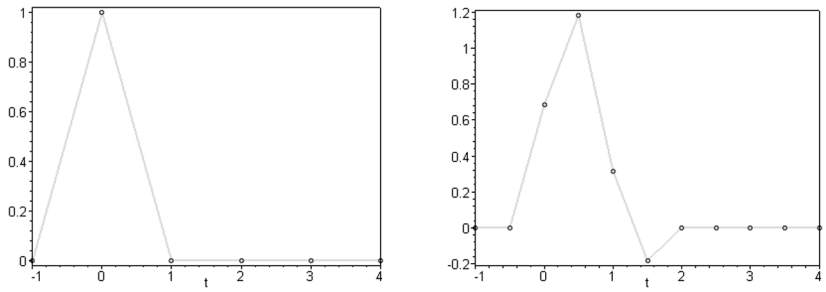


FIGURE 5 (*Left*) Initial guess for the cascade algorithm. (*Right*) First iteration of the cascade algorithm.

points is not necessary, except to get good images of the iterates. Also, for normalization reasons, it is important that the values of u_0 sum to 1.)

Next, use (14) to compute $u_1 = F(u_0)$ on all multiples of $\frac{1}{2}$, including the integers. For example,

$$u_1\left(\frac{3}{2}\right) = \frac{1+\sqrt{3}}{4}u_0(3) + \frac{3+\sqrt{3}}{4}u_0(2) + \frac{3-\sqrt{3}}{4}u_0(1) + \frac{1-\sqrt{3}}{4}u_0(0) = \frac{1-\sqrt{3}}{4}$$

Once the values on the multiples of $\frac{1}{2}$ are determined, connect the points to create Fig. 5.

A second iteration will yield u_2, defined on multiples of $\frac{1}{4}$. Graphs of the function u_2 and other iterates, demonstrating convergence to D_4, are found in Fig. 6. Of note is that, at convergence, the scaling function equals zero on all of the integers, except that $D_4(1) = (1+\sqrt{3})/2$ and $D_4(2) = (1-\sqrt{3})/2$.

As can be seen from this example, two important actions are occuring simultaneously in the cascade algorithm. First, each iterate is being defined on a finer scale that the previous one, so the algorithm can be run until a desired level of refinement is reached. Second, the iterates are converging to the scaling function. For the types of dilation equations described in this article, the cascade algorithm will converge, and proofs of this convergence can be found in the literature.

As in the case of the Haar wavelets, we can create a mother wavelet that corresponds with D_4. The key is the well-known relationship between the filter coefficients of the low-pass operator H and the high-pass operator G, and the relationship between the filter coefficients of H and the refinement coefficients in (9):

$$g_k = (-1)^k h_{1-k} \quad \text{and} \quad h_k = \frac{c_k}{\sqrt{2}} \tag{16}$$

These identities were demonstrated for the Haar wavelets in (6).

Since we know the values of c_k for D_4, we can determine h_k, and then g_k, and finally an equation as in (2) that relates D_4 to its mother wavelet:

$$\psi(t) = \frac{1-\sqrt{3}}{4}D_4(2t+2) - \frac{3-\sqrt{3}}{4}D_4(2t+1) + \frac{3+\sqrt{3}}{4}D_4(2t) - \frac{1+\sqrt{3}}{4}D_4(2t-1) \tag{17}$$

See Fig. 4 for a graph of ψ. This function is often referred to as simply the "Daubechies wavelet."

Refinement coefficients for other Daubechies scaling functions (e.g, D_6, D_8) can also be found in the literature. The cascade algorithm can be used to generate these functions, and equations such as (17) can be used to create the corresponding mother wavelets.

C. Analysis of Polynomials by Daubechies Wavelets

Constant and linear functions can be reproduced exactly by the Daubechies scaling function D_4 and its translates.

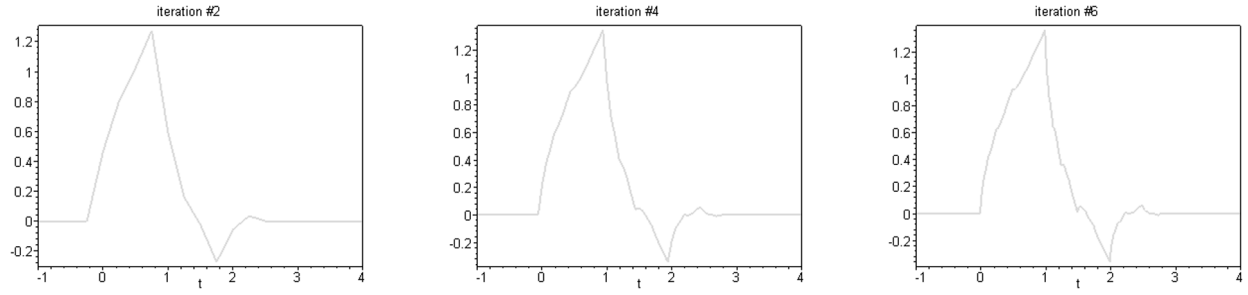


FIGURE 6 Selected iterations of the cascade algorithm.

This is surprising, as D_4 has a rather jagged shape. However, we will demonstrate how a linear function g can be written as a linear combination of D_4 and its translates. That is, we will show how to determine the values of a_k that satisfy

$$g(t) = \sum_k a_k D_4(t-k) \tag{18}$$

To find the coefficients in the linear combination, the orthogonal decomposition theorem can be used. For example, according to the theorem, a_{-2}, the coefficient that goes with $D_4(t+2)$, is defined by

$$a_{-2} = \frac{\langle g(t),\, D_4(t+2)\rangle}{\langle D_4(t+2),\, D_4(t+2)\rangle}$$

The denominator is 1 because the basis is orthonormal, so the coefficient that goes with $D_4(t+2)$ is

$$a_{-2} = \int_{-\infty}^{\infty} g(t) D_4(t+2)\, dt$$

Since the compact support of D_4 is $[0,3]$, this integral can be written as

$$a_{-2} = \int_{-2}^{1} g(t) D_4(t+2)\, dt \tag{19}$$

As discussed earlier, there is no simple formula for D_4, as it is simply defined through its dilation equation, so integrals such as (19) must be computed numerically to as much accuracy as desired.

A second approach to determining a_k begins by applying a simple change of variables to (18) to get

$$g(t) = \sum_j a_{t-j} D_4(j)$$

Using the values of D_4 on the integers yields

$$\begin{aligned} g(t) &= a_{t-1} D_4(1) + a_{t-2} D_4(2) \\ &= a_{t-1}\frac{1+\sqrt{3}}{2} + a_{t-2}\frac{1-\sqrt{3}}{2} \end{aligned} \tag{20}$$

If we think of a_k as a function in k, then it can be proved that a_k is a polynomial with the same degree as g.

We will use this fact to investigate what happens in the cases where g is the constant function 1, and where g is the linear function t. If g is constant, then a_k is the same constant for all k, which we will label a. Then (20) becomes

$$1 = a\frac{1+\sqrt{3}}{2} + a\frac{1-\sqrt{3}}{2} = a$$

So, $a_k = 1$ for all values of k, and

$$1 = \sum_k D_4(t-k) \tag{21}$$

If g is the linear function t, then $a_k = \gamma k + \delta$ for some constants γ and δ, and (20) is

$$t = (\gamma(t-1)+\delta)\frac{1+\sqrt{3}}{2} + (\gamma(t-2)+\delta)\frac{1-\sqrt{3}}{2}$$

which simplifies to

$$t = \gamma t + \left(\delta - \gamma\left(\frac{3-\sqrt{3}}{2}\right)\right)$$

Therefore, $\gamma = 1$, $\delta = (3-\sqrt{3})/2$, and we have the identity

$$t = \sum_k \left(k + \frac{3-\sqrt{3}}{2}\right) D_4(t-k) \tag{22}$$

We now apply this analysis to the specific case of $g = 3t+4$. Combining the identities (21) and (22) leads to

$$3t+4 = \sum_k \left(3k + \frac{17-3\sqrt{3}}{2}\right) D_4(t-k) \tag{23}$$

Although this sum is infinite, we can illustrate this identity on a finite interval. (It is tempting to conclude that $g = 3t+4$ is an element of V_0 for this wavelet family, as it is a linear combination of the translates of D_4. However, $V_0 \subset L^2(\mathbb{R})$, but g is not in $L^2(\mathbb{R})$. So, while any function in V_0 can be written uniquely as a linear combination of D_4 and its translates, not every linear combination is in V_0. The counterexamples are *infinite* sums where $\{a_k\} \notin l^2(\mathbb{R})$.) Here is the sum, using k from -1 to 1, with the coefficients evaluated to three decimal places:

$$2.902 D_4(t+1) + 5.902 D_4(t) + 8.902 D_4(t-1)$$

The graph of this linear combination is in Fig. 7. Note that it reproduces $3t+4$ exactly on the interval $[0,2]$, but not outside this interval.

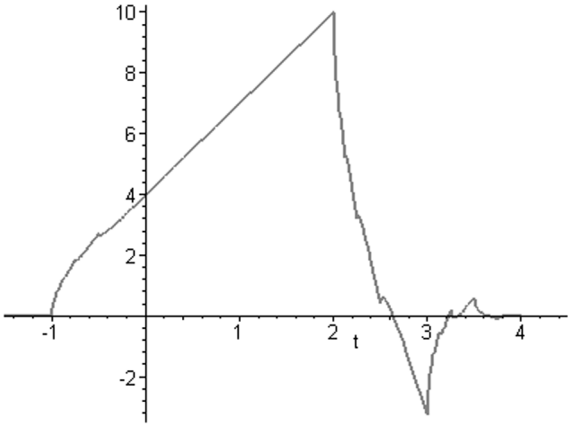


FIGURE 7 A truncated linear combination of the translates of D_4 reproduces a linear function on the interval [0,2].

D. Filters Based on Daubechies Wavelets

We now return to a discrete situation, where we wish to process a signal $\mathbf{s}$, rather than analyze a linear function g, with Daubechies wavelets. Two cases will be discussed: an infinite signal and a finite signal.

The low-pass and high-pass operators, H and G, defined by (7), will be used to process signals. For wavelets based on D_4, the values of the coefficients in the filter come from dividing the coefficients in (14) and (17) by $\sqrt{2}$, giving us

$$h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}} \approx 0.48296 \quad h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}} \approx 0.83652$$

$$h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}} \approx 0.22414 \quad h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}} \approx -0.12941$$

$$g_{-2} = \frac{1-\sqrt{3}}{4\sqrt{2}} \approx -0.12941$$

$$g_{-1} = -\frac{3-\sqrt{3}}{4\sqrt{2}} \approx -0.22414$$

$$g_0 = \frac{3+\sqrt{3}}{4\sqrt{2}} \approx 0.83652$$

$$g_1 = -\frac{1+\sqrt{3}}{4\sqrt{2}} \approx -0.48296$$

Since these are the only filter coefficients that are non-zero, the equations in (7) become

$$(H\mathbf{s})_k = h_0\mathbf{s}_{2k} + h_1\mathbf{s}_{2k+1} + h_2\mathbf{s}_{2k+2} + h_3\mathbf{s}_{2k+3} \qquad (24)$$

and

$$(G\mathbf{s})_k = g_{-2}\mathbf{s}_{2k-2} + g_{-1}\mathbf{s}_{2k-1} + g_0\mathbf{s}_{2k} + g_1\mathbf{s}_{2k+1} \qquad (25)$$

For example, consider the infinite signal that is analogous to the example in the previous section, $g = 3t + 4$. For all j, let $\mathbf{s}_j = 3j + 4$. Then, (24) and (25) become

$$(H\mathbf{s})_k = h_0(6k+4) + h_1(6k+7) + h_2(6k+10) + h_3(6k+13) = \sqrt{2}\left(6k + \frac{17-3\sqrt{3}}{2}\right)$$

and

$$(G\mathbf{s})_k = g_{-2}(6k-2) + g_{-1}(6k+1) + g_0(6k+4) + g_1(6k+7) = 0$$

There are many observations to make about these results. The output of the low-pass operator is, as in the case of the Haar wavelets, an average, but this time it is a weighted average, using the four weights h_0, h_1, h_2, and h_3. Those specific weights are important because then the high-pass operator G will filter the linear signal completely, leaving an output of zero. In other words, the details that are critical here are the ones that deviate from a linear trend, and, in this case, there are no details. *This is why the Daubechies wavelets are important.*

Also of note is that the output of the low-pass operator is similar to the coefficients in (23). Integrals such as (19) also compute weighted averages, but this time in the continuous domain, so it is reasonable that the results would be quite similar. The different coefficient on k is due to normalization and downsampling.

E. Analysis of Finite Signals

In reality, signals that require processing are finite, which presents a problem at the ends of the signal. If a signal $\mathbf{s} = [\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_7]$ has length 8, then we can compute $(H\mathbf{s})_k$ only for $k = 0$, 1, and 2. Similarly, we can compute $(G\mathbf{s})_k$ only for $k = 1$, 2, and 3. If we are ever going to have any hope of reconstructing the original signal from average and detail coefficients, then we will need at least eight of them. There are many ways of addressing this issue, all of which involve concatenating infinite strings on both ends of $\mathbf{s}$, giving us an infinite signal that we can process as we did earlier. Then, we can generate as many average and detail coefficients as we need.

One approach is called *zero padding*, and we simply add infinite strings of zeros to both ends of $\mathbf{s}$, giving us

$$\ldots, 0, 0, 0, \mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4, \mathbf{s}_5, \mathbf{s}_6, \mathbf{s}_7, 0, 0, 0, \ldots$$

A second approach is called the *periodic method*, where we repeat $\mathbf{s}$ as if it were periodic, yielding

$$\ldots, \mathbf{s}_5, \mathbf{s}_6, \mathbf{s}_7, \mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4, \mathbf{s}_5, \mathbf{s}_6, \mathbf{s}_7, \mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \ldots$$

A third way is called the *symmetric extension method*, where we create symmetry in the string by concatenating the the reverse of $\mathbf{s}$ and repeating, leading to

$$\ldots, \mathbf{s}_2, \mathbf{s}_1, \mathbf{s}_0, \mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4, \mathbf{s}_5, \mathbf{s}_6, \mathbf{s}_7, \mathbf{s}_7, \mathbf{s}_6, \mathbf{s}_5, \ldots$$

Each method has its advantages. Zero padding is easy to implement. The periodic method leads to periodic output from the filters. Symmetric extension is popular because it does not introduce a sudden change in the signal as the other two do. It is important to observe that each method described here will lead to different results when applying H and G to the extended signals.

For the filters based on D_4, there is an overlapping dependence of filter values on data points. For this reason, we must use one of these boundary methods. This is not a problem with the Haar wavelets, however, so signals being processed with the Haar filters do not need to be extended. Finally, the shortest signal that can be processsed

by a filter is a concern. The Haar filters can be applied to signals of length 2, but the shortest possible signal for the D_4 filters has length 4.

F. A Final Example

At the beginning of this article, we investigated the analysis of a finite signal using the Haar wavelets. In this section, the filters based on D_4 will be used to analyze a signal that has a linear trend, along with some random noise. Through this example, the various ideas discussed in this article will come into play.

The following string is a truncated version of the infinite string from earlier, $3j+4$, with some added random noise (see Fig. 8):

$$\mathbf{s}_{init} = [4.03, 7.82, 9.66, 12.45, 16.03, 15.97, 22, 27.11, 28.58, 31.98, 34, 36.99, 40, 44.1, 44.87, 49.13]$$

We will apply the periodic boundary method to create a new signal that we can filter. This signal would be

$$\tilde{\mathbf{s}} = [\ldots, 34, 36.99, 40, 44.1, 44.87, 49.13, \mathbf{4.03}, \mathbf{7.82}, \mathbf{9.66}, \mathbf{12.45}, \mathbf{16.03}, \mathbf{15.97}, \mathbf{22}, \mathbf{27.11}, \mathbf{28.58}, \mathbf{31.98}, \mathbf{34}, \mathbf{36.99}, \mathbf{40}, \mathbf{44.1}, \mathbf{44.87}, \mathbf{49.13}, 4.03, 7.82, 9.66, 12.45, 16.03, 15.97, \ldots]$$

where the periodic "core" of this signal is in bold. Applying (24) and (25) to this signal yields

$$\tilde{\mathbf{s}}_l = [9.042, 16.606, 22.524, 35.571, 43.389, 50.622, 59.908, 62.660]$$

and

$$\tilde{\mathbf{s}}_h = [-17.224, -0.206, 1.656, -0.344, -0.461, -0.290, -0.529, -1.254]$$

We observe that the low-pass filter H is computing weighted averages of the values in the signal, while the high-pass filter G captures the deviation from a linear sequence, which, in this case, is the noise. (The Haar filters capture the deviation from a constant sequence.) Unlike the earlier example, the detail coefficients are not zero because noise is detected. Also, because the linear trend in the data is broken at the periodic extension (from $\tilde{\mathbf{s}}_{-1}$ to $\tilde{\mathbf{s}}_0$, for instance) there is a detail coefficient, -17.224, that is not very close to zero.

As we did at the beginning of this article, we can apply the filters to $\tilde{\mathbf{s}}_l$, extended periodically, to perform a second level of processing. Applying (24) and (25) produces

$$\tilde{\mathbf{s}}_{ll} = [18.704, 43.808, 68.620, 81.227]$$

and

$$\tilde{\mathbf{s}}_{lh} = [-22.254, -3.230, .960, 2.891]$$

We can apply the filters one more time to $\tilde{\mathbf{s}}_{ll}$, which has period 4 (the minimum length for the D_4 filters). This will give us:

$$\tilde{\mathbf{s}}_{lll} = [50.548, 99.612] \quad \text{and} \quad \tilde{\mathbf{s}}_{llh} = [-32.598, 5.933]$$

Combining $\tilde{\mathbf{s}}_{lll}$ and $\tilde{\mathbf{s}}_{llh}$ with the detail coefficients given previously, we create the new signal

$$\tilde{\mathbf{s}}_{new} = [50.548, 99.612, -32.598, 5.933, -22.254, -3.230, .960, 2.891, -17.224, -.206, 1.656, -.344, -.461, -.290, -.529, -1.254]$$

This signal, $\tilde{\mathbf{s}}_{new}$, of length 16 contains all of the information needed to reproduce $\mathbf{s}_{init}$.

The next step in processing this signal is to apply thresholding as a way to remove the noise. Applying a tolerance of 1.8 to $\mathbf{s}_{new}$ to eliminate some of the detail coefficients produces

$$\hat{\mathbf{s}} = [50.548, 99.612, -32.598, 5.933, -22.254, -3.230, 0, 2.891, -17.224, 0, 0, 0, 0, 0, 0, 0]$$

Finally, we apply the dual operators (8) to $\hat{\mathbf{s}}$ and compare our results with $\mathbf{s}_{init}$. Reconstructing the original signal requires applying the dual operators three times. The first time, we calculate $H^*(\tilde{\mathbf{s}}_{lll}) + G^*(\tilde{\mathbf{s}}_{llh})$, creating

$$[18.704, 43.808, 68.620, 81.227]$$

Continuing this process ultimately yields

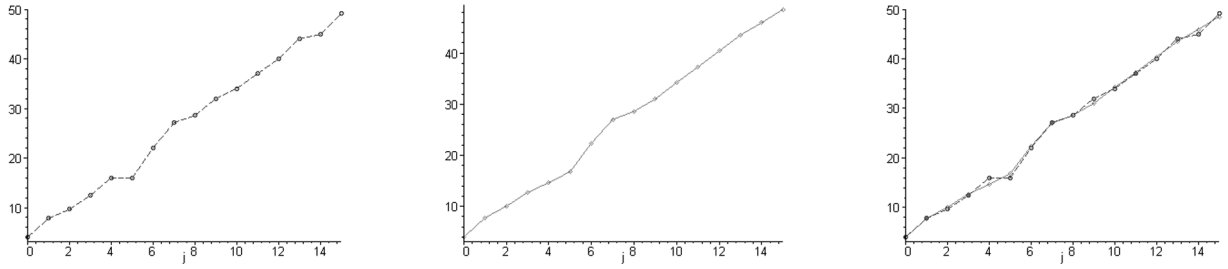


FIGURE 8 (*Left*) The original signal—a linear trend, but with noise. (*Center*) The smoother processed signal. (*Right*) Both signals together.

[4.003, 7.773, 10.046, 12.721, 14.660, 16.796,
22.359, 27.004, 28.588, 30.993, 34.217, 37.223,
40.383, 43.503, 45.919, 48.524]

The original signal and the new signal are compared in Fig. 8. Notice that the effect of processing with thresholding is to smooth the original signal and make it more linear. If we had wished to denoise only the second half of the original signal, then we would have altered only the detail coefficients that are based on that part of the signal.

This example demonstrates many of the ideas discussed in this article. The filters based on D_4 were well suited for this example, as the original signal had a linear trend, and the filters effectively separated the linear part of the signal from the noise. Each detail coefficient indicated the extent of the noise in a specific part of the signal. We then reduced the noise through thresholding, in effect smoothing the original signal. In this process, we applied filters at three different resolutions, until the output of the low-pass filter was too short to continue.

X. CONCLUSION

Our discussion in this article has focused on the Haar and Daubechies wavelets and their application to one-dimensional signal processing. As we have seen, the main goal in wavelet research and applications is to find or create a set of basis functions that will capture, in some efficient manner, the essential information contained in a signal or function. Recent work in wavelets by Donoho, Johnson, Coifman, and others has shown that wavelet systems have inherent characteristics that make them almost ideal for a wide range of problems of this type. This article has addressed just the basic ideas behind wavelets. Additional topics for an interested reader include frames, multiwavelets, wavelet packets, biorthogonal wavelets, and filter banks, to name only a few.

Many results in wavelet theory (the important relationship in (16), for example) are generally derived through Fourier analysis. The lack of space and the expository nature of an encyclopedia article motivates the decision to omit such discussion. However, there is such a close relationship between Fourier analysis and wavelets that it is impossible to thoroughly understand the latter without some knowledge of the former.

Wavelets have become an important tool in many disciplines of science. In areas as diverse as radiology, electrical engineering, materials science, seismology, statistics, and speech analysis, scientists have found the key ideas in this subject—multiresolution, localization, and adaptability to specific requirements—to be very important and useful in their work.

SEE ALSO THE FOLLOWING ARTICLES

FOURIER SERIES • IMAGE PROCESSING • SIGNAL PROCESSING, DIGITAL • WAVELETS, ADVANCED

BIBLIOGRAPHY

Aboufadel, E., and Schlicker, S. (1999). "Discovering Wavelets," John Wiley & Sons, New York.

Brislawn, C. (1995). "Fingerprints go digital," *Notices A.M.S.* **42,** 1278–1283.

Daubechies, I. (1992). "Ten Lectures on Wavelets," SIAM, Philadelphia.

Graps, A. (1995). "An introduction to wavelets," *IEEE Comp. Sci. Eng.* **2**(2). (Available at `www.amara.com/IEEEwave/IEEEwavelet.html`)

Hubbard, B. (1996). "The World According to Wavelets," A. K. Peters, Inc., Wellesley, MA.

Jawerth, B., and Sweldens, W. (1994). "An overview of wavelet based multiresolution analyses," *SIAM Rev.* **36,** 377–412.

Kobayashi, M. (ed.). (1998). "Wavelets and Their Applications: Case Studies," SIAM, Philadelphia.

Mallat, S. (1999). "A Wavelet Tour of Signal Processing," 2nd ed., Academic Press, London.

Meyer, Y. (1993). "Wavelets: Algorithms and Applications," SIAM, Philadelphia.

Mulcahy, C. (1996). "Plotting and scheming with wavelets," *Math. Mag.* **69,** 323–343.

Nguyen, T., and Strang, G. (1996). "Wavelets and Filter Banks," Wellesley-Cambridge Press, Wellesley, MA.

Strang, G. (1989). "Wavelets and dilation equations: a brief introduction," *SIAM Rev.* **31,** 614–627.

Strang, G. (1993). "Wavelet transforms versus Fourier transforms." *Bull. A.M.S.* **28,** 288–305.

Strang, G. (1994). "Wavelets," *Am. Scientist* **82,** 250–255.

Stollnitz, E., DeRose, T., and Salesin, D. (1996). "Wavelets for Computer Graphics," Morgan Kaufmann, San Francisco.

Sweldens, W. (1997). "Wavelet cascade applet: mathematical background," (Available at `cm.bell-labs.com/cm/ms/who/wim/cascade/math.html`)

Walter, G. (1994). "Wavelets and Other Orthogonal Systems with Applications," CRC Press, Boca Raton, FL.