科爱传播
KEAI COMMUNICATIONS
精编百科

·导读版·

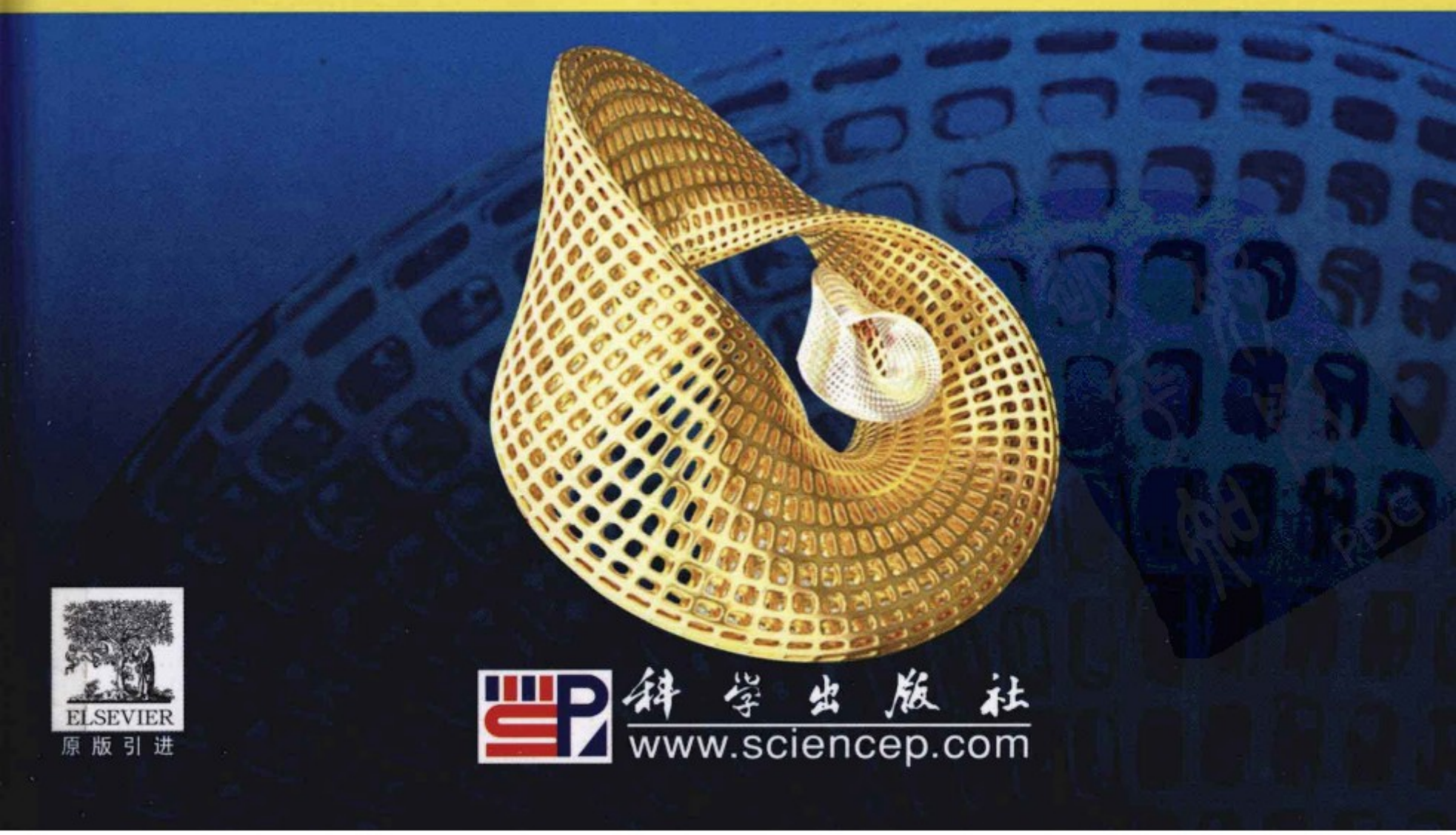
Encyclopedia of Mathematical Physics

Gauge Theory; Quantum Field Theory

数学物理学百科全书 4

规范理论；量子场论

Jean-Pierre Françoise, Gregory L. Naber, Tsou Sheung Tsun



ELSEVIER
原版引进

科学出版社
www.sciencep.com

(O-3129.0101)

科爱传播
KE AI COMMUNICATIONS
精编百科

·导读版·

本卷前言、序、目录均已译成中文，正文部分保留英文原版。另附吴岳良院士所作中文导读一篇。

Encyclopedia of Mathematical Physics

Gauge Theory; Quantum Field Theory

数学物理学百科全书 ❹

规范理论；量子场论

本套书特色

- 编纂队伍阵容强大——来自30个国家的400多位物理学家和数学家，历时4年，倾力奉献。包括诺贝尔物理学奖获得者杨振宁教授和英国牛津大学Roger Penrose教授等。
- 按学科分支重新编排——共分12卷：数学物理导言1卷（含中文翻译），物理学方面7卷（卷2～卷8），数学方面4卷（卷9～卷12）。
- 内容新颖权威——400多篇图文并茂的综述性文章，内容全面系统、领域涵盖广泛，参考文献丰富，可全面了解数学物理基础知识、发展前沿以及核心课题。
- 适用范围广泛——适于物理学和数学领域的所有高等院校的广大师生和科研院所的研究人员及研究生参考使用。

分卷内容

卷1 数学物理学导言(双语版)
卷2 经典力学；流体动力学（导读版）
卷3 可积系统；经典，共形与拓扑场论(导读版)
▶ 卷4 规范理论；量子场论（导读版）
卷5 广义相对论；量子引力；弦论和M-理论(导读版)
卷6 凝聚态物质与光学；量子信息与量子计算；量子力学(导读版)
卷7 无序系统；动力系统(导读版)
卷8 平衡态统计力学；非平衡态统计力学(导读版)
卷9 代数技巧；李群和李代数；离散数学；量子群；随机方法(导读版)
卷10 复几何；微分几何；低维几何；非交换几何(导读版)
卷11 代数拓扑；辛几何与拓扑；常微分和偏微分方程(导读版)
卷12 泛函分析和算子代数；量子化方法和路径积分；变分技术(导读版)

电话：010-64006871
网址：www.kbooks.cn

ISBN 978-7-03-021641-0
9 787030 216410 >
定价：138.00 元

销售分类建议：物理学；数学
本版本只限于在中华人民共和国境内销售

Encyclopedia of Mathematical Physics

数学物理学百科全书④

Gauge Theory; Quantum Field Theory

规范理论；量子场论

Jean-Pierre Françoise
Gregory L. Naber
Tsou Sheung Tsun

科学出版社
北京

图字：01-2008-1082 号

This is an annotated version of
Encyclopedia of Mathematical Physics
Jean-Pierre Françoise , Gregory L Naber , Tsou Sheung Tsun
Copyright © 2007 Elsevier Inc.
ISBN-13：978-0-1251-2660-1
ISBN-10 ：0-12-512660-3
All rights reserved.
No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

AUTHORIZED EDITION FOR SALE IN P. R. CHINA ONLY
本版本只限于在中华人民共和国境内销售

图书在版编目(CIP)数据

数学物理学百科全书，4. 规范理论；量子场论＝Encyclopedia of Mathematical Physics. 4，Gauge Theory；Quantum Field Theory：英文/（法）费朗克斯（Françoise，J. P.）等编著. —影印本. —北京：科学出版社，2008

ISBN 978-7-03-021641-0

Ⅰ. 数… Ⅱ. 费… Ⅲ. ①数学-百科全书-英文②物理学-百科全书-英文③规范场-量子论-英文 Ⅳ. 0-61

中国版本图书馆 CIP 数据核字（2008）第 050929 号

责任编辑：孙红梅 霍志国/责任印制：钱玉芬/封面设计：耕者设计工作室

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京佳信达艺术印刷有限公司印刷
科学出版社发行 各地新华书店经销
*
2008 年 6 月第 一 版 开本：787×1092 1/16
2008 年 6 月第一次印刷 印张：38 1/4
印数：1—2 000 字数：904 000

定价：138.00 元

如有印装质量问题，我社负责调换

《数学物理学百科全书》导读版编委会

主编

吴岳良　院士　中国科学院理论物理研究所
刘克峰　教授　浙江大学数学系　美国加州大学洛杉矶分校数学系

编委（按姓氏笔画）

庄鹏飞　教　授　清华大学物理系
朱重远　研究员　中国科学院理论物理研究所
张元仲　研究员　中国科学院理论物理研究所
沈一兵　教　授　浙江大学数学系
陈晓松　研究员　中国科学院理论物理研究所
陈熙谋　教　授　北京大学物理学院
周向宇　研究员　中国科学院数学与系统科学研究院
姚景齐　研究员　中国科学院数学与系统科学研究院
段海豹　研究员　中国科学院数学与系统科学研究院
曹则贤　研究员　中国科学院物理研究所

目　　录

规范理论

微分形式的阿贝尔和非阿贝尔规范理论　3
阿贝尔希格斯涡旋　13
AdS/CFT 对应　23
Aharonov-Bohm 效应　33
反常　41
BRST 量子化　49
Chern-Simons 模型：严格结果　56
引力和非阿贝尔规范理论中的狄拉克场　61
Donaldson-Witten 理论　69
有效场论　77
电磁对偶　86
电弱理论　94
精确重正化群　102
弦的规范理论　112
规范理论：数学应用　118
瞬子：拓扑理论　132
大 N 与拓扑弦　139
格点规范理论　145
圈空间测度　152
非对易几何与标准模型　157
规范理论的非微扰和拓扑理论　164
微扰重正化理论与 BRST　175
量子色动力学　182
量子电动力学及其精密验证　192
重正化：通论　201
Seiberg-Witten 理论　210
粒子物理的标准模型　220
超引力　227
超对称粒子模型　234
场论中的对称性破缺　240
扭量理论：某些应用　247
二维模型　256

量子场论

量子场论：导论 273
量子场论的代数方法 278
公理化量子场论 285
Batalin-Vilkovisky 量子化 292
外场中的玻色子与费米子 299
约束系统 308
结构量子场论 314
流代数 329
狄拉克算符与狄拉克场 335
色散关系 348
欧几里德场论 363
量子场论中的 Gerbes 373
完整量子场 381
重正化量子场论的 Hopf 代数结构 386
不定度规 395
可积性与量子场论 402
量子场论中的算符乘积展开 412
不定度规量子场：非平庸模型 417
微扰理论及其技巧 423
有拓扑缺陷的量子场 437
弯曲时空中的量子场论 446
量子相变 457
重正化：统计力学与凝聚态物质 464
散射，渐进完备性与束缚态 472
相对论量子场论中的散射：基本概念和工具 485
相对论量子场论中的散射：分析程序 495
对称与守恒律 506
量子场论的对称性：代数理论 513
低时空维度量子场论的对称性 518
热量子场论 526
Tomita-Takesaki 模块理论 535
拓扑缺陷及其同伦分类 542

总目录 551

（穆良柱　译　陈熙谋　校）

目　录

Gauge Theory

Abelian and Nonabelian Gauge Theories Using Differential Forms ····· 3
Abelian Higgs Vortices ····· 13
AdS/CFT Correspondence ····· 23
Aharonov-Bohm Effect ····· 33
Anomalies ····· 41
BRST Quantization ····· 49
Chern-Simons Models: Rigorous Results ····· 56
Dirac Fields in Gravitation and Nonabelian Gauge Theory ····· 61
Donaldson-Witten Theory ····· 69
Effective Field Theories ····· 77
Electric-Magnetic Duality ····· 86
Electroweak Theory ····· 94
Exact Renormalization Group ····· 102
Gauge Theories from Strings ····· 112
Gauge Theory: Mathematical Applications ····· 118
Instantons: Topological Aspects ····· 132
Large-N and Topological Strings ····· 139
Lattice Gauge Theory ····· 145
Measure on Loop Spaces ····· 152
Noncommutative Geometry and the Standard Model ····· 157
Nonperturbative and Topological Aspects of Gauge Theory ····· 164
Perturbative Renormalization Theory and BRST ····· 175
Quantum Chromodynamics ····· 182
Quantum Electrodynamics and Its Precision Tests ····· 192
Renormalization: General Theory ····· 201
Seiberg-Witten Theory ····· 210
Standard Model of Particle Physics ····· 220
Supergravity ····· 227
Supersymmetric Particle Models ····· 234
Symmetry Breaking in Field Theory ····· 240
Twistor Theory: Some Applications ····· 247
Two-Dimensional Models ····· 256

Quantum Field Theory

Quantum Field Theory: A Brief Introduction ····· 273

Algebraic Approach to Quantum Field Theory ····· 278
Axiomatic Quantum Field Theory ····· 285
Batalin-Vilkovisky Quantization ····· 292
Bosons and Fermions in External Fields ····· 299
Constriained Systems ····· 308
Constructive Quantum Field Theory ····· 314
Current Algebra ····· 329
Dirac Operator and Dirac Field ····· 335
Dispersion Relations ····· 348
Euclidean Field Theory ····· 363
Gerbes in Quantum Field Theory ····· 373
Holonomic Quantum Fields ····· 381
Hopf Algebra Structure of Renormalizable Quantum Field Theory ····· 386
Indefinite Metric ····· 395
Integrability and Quantum Field Theory ····· 402
Operator Product Expansion in Quantum Field Theory ····· 412
Quantum Fields with Indefinite Metric: Non-Trivial Models ····· 417
Perturbation Theory and Its Techniques ····· 423
Quantum Fields with Topological Defects ····· 437
Quantum Field Theory in Curved Spacetime ····· 446
Quantum Phase Transitions ····· 457
Renormalization: Statistical Mechanics and Condensed Matter ····· 464
Scattering, Asymptotic Completeness and Bound States ····· 472
Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools ····· 485
Scattering in Relativistic Quantum Field Theory: The Analytic Program ····· 495
Symmetries and Conservation Laws ····· 506
Symmetries in Quantum Field Theory: Algebraic Aspects ····· 513
Symmetries in Quantum Field Theory of Lower Spacetime Dimensions ····· 518
Thermal Quantum Field Theory ····· 526
Tomita-Takesaki Modular Theory ····· 535
Topological Defects and Their Homotopy Classification ····· 542

总目录 ····· 551

Gauge Theory

PDG

Abelian and Nonabelian Gauge Theories Using Differential Forms

A C Hirshfeld, Universität Dortmund, Dortmund, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Quantum electrodynamics is the theory of the electromagnetic interactions of photons and electrons. When attempting to generalize this theory to other interactions it turns out to be necessary to identify its essential components. The essential properties of electrodynamics are contained in its formulation as an "abelian gauge theory." The generalization to include other interactions is then reduced to incorporating the structure of nonabelian groups. This becomes particularly clear when we formulate the theory in the language of differential forms.

Here we first present the formulation of electrodynamics using differential forms. The electromagnetic fields are introduced via the Lorentz force equation. They are recognized as the components of a differential 2-form. This form fulfills two differential conditions, which are equivalent to Maxwell's equations. These are expressed with the help of a differential operator and its Hermitian conjugate, the codifferential operator. We consider the effects of charge conservation and introduce electromagnetic potentials, which are defined up to gauge transformations. We finally consider Weyl's argument for the existence of the electromagnetic interaction as a consequence of the local phase invariance of the electron wave function.

We then go on to present the nonabelian generalization. The gauge bosons appear in a theory with fermions by requiring invariance of the theory with respect to local gauge transformations. When the fermions group into symmetry multiplets this gives rise to a gauge group SU(N) involving N^2-1 gauge bosons mediating the interaction, where N is the dimension of the Lie algebra. The interaction arises through the necessity of replacing the usual derivatives by covariant derivatives, which transform in a natural way in order to preserve the gauge invariance. The covariant derivatives involve the gauge potentials, whose transformation properties are dictated by those of the covariant derivative. Whereas for an abelian gauge theory such as electromagnetism scalar-valued p-forms are sufficient (actually only $p=1,2$), a nonabelian gauge theory involves the use of Lie-algebra-valued p-forms. These are introduced and used to construct the Yang–Mills action, which involves the field strength tensor which is determined from the gauge potentials. This action leads to the Yang–Mills equations for the gauge potentials, which are the nonabelian generalizations of the Maxwell equations.

Relativistic Kinematics

The trajectory of a mass point is described as $x^\mu(\tau)$, where τ is the invariant proper time interval:

$$d\tau^2 = dt^2 - d\boldsymbol{x}\cdot d\boldsymbol{x} = dt^2(1-\boldsymbol{v}^2) \qquad [1]$$

with $\boldsymbol{v}=d\boldsymbol{x}/dt$. With the abbreviation $\gamma=(1-\boldsymbol{v}^2)^{-1/2}$ this yields $d\tau=(1/\gamma)dt$.

The 4-velocity of a point is defined as $u^\mu = dx^\mu/d\tau=\gamma(dx^\mu/dt)$. The quantity

$$u^2 = g_{\mu\nu}u^\mu u^\nu = \frac{dx^\mu dx_\mu}{d\tau^2} = 1 \qquad [2]$$

is a relativistic invariant. Here

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \qquad [3]$$

is the metric of Minkowski space.

The 4-momentum of a particle is $p^\mu = m_0u^\mu = (m_0\gamma, m_0\gamma\boldsymbol{v})$, and $p^\mu p_\mu = m_0^2$. The 4-force is

$$f^\mu = \frac{dp^\mu}{d\tau} = \gamma\frac{dp^\mu}{dt} = \gamma\left(\frac{dp^0}{dt}, f\right) \qquad [4]$$

with the 3-force

$$f = \frac{d(m_0\gamma\boldsymbol{v})}{dt} \qquad [5]$$

Notes:

Differentiate $p^2 = m_0^2$ with respect to τ, this yields

$$2p^\mu f_\mu = 2m_0\gamma^2\left(\frac{dp^0}{dt} - f\cdot v\right) = 0 \qquad [6]$$

or

$$\frac{dp^0}{dt} = f\cdot v = f\cdot\frac{dx}{dt} \qquad [7]$$

This says that

$$dp^0 = f\cdot dx = dW \qquad [8]$$

where W is the work done and p^0 is the energy.

For a charged particle, the Lorentz force is

$$f = q(E + v\times B) \qquad [9]$$

where q is the charge of the particle, E is the electric, and B the magnetic field strength. Since $f\cdot v = qE\cdot v$, we have the four-dimensional form of the Lorentz force:

$$f^\mu = q\gamma(E\cdot v, E + v\times B) \qquad [10]$$

The Lorentz Force Equation with Differential Forms

We write the Lorentz force equation as an equation for a differential form $f = f_\mu dx^\mu$, with $f_\mu = g_{\mu\nu}f^\nu$. The velocity-dependent Lorentz force is

$$f = -qi_u F \qquad [11]$$

with

$$u = \gamma\left(\frac{\partial}{\partial t} + v^x\frac{\partial}{\partial x} + v^y\frac{\partial}{\partial y} + v^z\frac{\partial}{\partial z}\right) \qquad [12]$$

the 4-velocity and F the electromagnetic field strength:

$$F = \mathcal{E}\wedge dt + \mathcal{B} \qquad [13]$$

where $\mathcal{E}$ is a 1-form in three dimensions,

$$\mathcal{E} = E_x dx + E_y dy + E_z dz \qquad [14]$$

and $\mathcal{B}$ is a 2-form in three dimensions,

$$\mathcal{B} = B_x dy\wedge dz + B_y dz\wedge dx + B_z dx\wedge dy \qquad [15]$$

The symbol i_u indicates a contraction of a 2-form with a vector, which is defined as

$$i_u F(v) = F(u, v) \qquad [16]$$

for an arbitrary vector v. The contraction of a 2-form with a vector yields a 1-form.

It is easily seen that a 2-form can be expressed in terms of a polar vector and an axial vector: if it is to be invariant with respect to parity transformations with

$$t\to t, \qquad x\to -x, \qquad y\to -y, \qquad z\to -z \qquad [17]$$

the fields in eqn [13] must transform as

$$E\to -E, \qquad B\to B \qquad [18]$$

Now we check the validity of eqn [11]. We have

$$\begin{aligned} f &= -qi_u F \\ &= q\gamma(v\cdot E)dt - q\gamma[(E^x + (v\times B)^x)dx \\ &\quad + (E^y + (v\times B)^y)dy + (E^z + (v\times B)^z)dz] \end{aligned} \qquad [19]$$

in agreement with eqn [10]. We remember to change the signs in $E_x = -E^x$, $B_x = -B^x$, etc.

The Codifferential Operator

The space of p-forms on an n-dimensional manifold is an

$$\binom{n}{p} = \binom{n}{n-p} = \frac{n!}{(n-p)!p!} \qquad [20]$$

dimensional vector space. The space of p-forms is thus isomorphic to the space of $(n-p)$-forms. The Hodge dual operator maps the p-forms into the $(n-p)$-forms, and is defined by

$$\alpha\wedge *\beta = \langle\alpha,\beta\rangle dx^1\wedge\cdots\wedge dx^n \qquad [21]$$

Here $\langle\alpha,\beta\rangle$ is the scalar product of two p-forms:

$$\langle\alpha,\beta\rangle = \alpha_{i_1\cdots i_p}\beta^{i_1\cdots s i_p} \qquad [22]$$

where $\alpha_{i_1\cdots s i_p}$ are the coefficients of the form α,

$$\alpha = \alpha_{i_1\cdots i_p}dx^{i_1}\wedge\cdots\wedge dx^{i_p} \qquad [23]$$

$\beta_{j_1\cdots s j_p}$ are the coefficients of the form β,

$$\beta = \beta_{j_1\cdots j_p}dx^{j_1}\wedge\cdots\wedge dx^{j_p} \qquad [24]$$

and

$$\beta^{i_1\cdots i_p} = g^{i_1 j_1}\cdots g^{i_p j_p}\beta_{j_1\cdots j_p} \qquad [25]$$

The indices satisfy $i_1 < \cdots < i_p$ and $j_1 < \cdots < j_p$.

The basis elements are orthogonal with respect to this scalar product, and

$$\begin{aligned} &\langle dx^{i_1}\wedge\cdots\wedge dx^{i_p}, dx^{i_1}\wedge\cdots\wedge dx^{i_p}\rangle \\ &\quad = g_{i_1 i_1}\cdots g_{i_p i_p} \end{aligned} \qquad [26]$$

The Hodge dual has the property that

$$\begin{aligned} &*\left(dx^{\sigma(1)}\wedge\cdots\wedge dx^{\sigma(p)}\right)\\ &\quad= g_{\sigma(1)\sigma(1)}\cdots g_{\sigma(p)\sigma(p)}(\operatorname{sign}\sigma)\\ &\qquad\times\left(dx^{\sigma(p+1)}\wedge\cdots\wedge dx^{\sigma(n)}\right) \end{aligned} \tag{27}$$

where σ is a permutation of the indices $(1,\ldots,n)$, $\sigma(1)<\cdots<\sigma(p)$, and $\sigma(p+1)<\cdots<\sigma(n)$. We also have

$$\begin{aligned} &*\left(dx^{\sigma(p+1)}\wedge\cdots\wedge dx^{\sigma(n)}\right)\\ &\quad= g_{\sigma(p+1)\sigma(p+1)}\cdots g_{\sigma(n)\sigma(n)}(-1)^{p(n-p)}(\operatorname{sign}\sigma)\\ &\qquad\times\left(dx^{\sigma(1)}\wedge\cdots\wedge dx^{\sigma(p)}\right) \end{aligned} \tag{28}$$

We therefore find that the application of the Hodge dual to a p-form twice yields

$$\begin{aligned} &**\left(dx^{\sigma(1)}\wedge\cdots\wedge dx^{\sigma(p)}\right)\\ &\quad= g_{\sigma(1)\sigma(1)}\cdots g_{\sigma(p)\sigma(p)}(\operatorname{sign}\sigma)*\left(dx^{\sigma(p+1)}\wedge\cdots\wedge dx^{\sigma(n)}\right)\\ &\quad= g_{\sigma(1)\sigma(1)}\cdots g_{\sigma(n)\sigma(n)}(-1)^{p(n-p)}dx^{\sigma(1)}\wedge\cdots\wedge dx^{\sigma(p)} \end{aligned} \tag{29}$$

or

$$** = (-1)^{p(n-p)}(-1)^{\operatorname{Ind} g}\mathrm{Id} \tag{30}$$

where $\operatorname{Ind} g$ is the number of times (-1) occurs along the diagonal of g.

Now let α be a $(p-1)$-form, and β a p-form. Then $d*\beta$ is an $(n-p+1)$-form, and

$$\begin{aligned} d(\alpha\wedge*\beta) &= d\alpha\wedge*\beta+(-1)^{p-1}\alpha\wedge d*\beta\\ &= d\alpha\wedge*\beta+(-1)^{(p-1)}(-1)^{(n-p+1)(p-1)}\\ &\quad\times(-1)^{\operatorname{Ind} g}\alpha\wedge(**)d*\beta\\ &= d\alpha\wedge*\beta+(-1)^{n(p-1)}(-1)^{\operatorname{Ind} g}\\ &\quad\times\alpha\wedge*(*d*\beta) \end{aligned} \tag{31}$$

We then have

$$(d\alpha,\beta)-(\alpha,d^*\beta)=\int_M d(\alpha\wedge*\beta) \tag{32}$$

with

$$d^* = -(-1)^{n(p-1)}(-1)^{\operatorname{Ind} g}*d* \tag{33}$$

We are here using the scalar product of two p-forms

$$(\alpha,\beta):=\int_M(\alpha\wedge*\beta) \tag{34}$$

With the help of Stokes' theorem the last integral in eqn [32] may be turned into a surface term at infinity, which vanishes for α and β with compact support. d^* is the adjoint operator to d with respect to the scalar product $(,)$. Whereas the differential operator d maps p-forms into $(p+1)$-forms, the codifferential operator d^* maps p-forms into $(p-1)$-forms.

The relation $d^2=0$ leads to

$$(d^*)^2\propto(*d*)(*d*)\propto *d^2* = 0 \tag{35}$$

This fact plays an essential role in connection with the conservation laws.

Finally, we want to obtain a coordinate expression for $d^*\beta$. Indeed $d^*\beta=-\operatorname{Div}\beta$ for

$$(\operatorname{Div}\beta)_K=\frac{\partial\beta^j_K}{\partial x^j} \tag{36}$$

where K is the multi-index of the coeffecients in $\beta=\beta_{\underline{K}}dx^K$, and $\underline{K}$ indicates that $K=(k_1,\ldots,k_p)$ is in the order $k_1<\cdots<k_p$. We will show that $(\alpha,d^*\beta)=(\alpha,-\operatorname{Div}\beta)$ for an arbitrary $(p-1)$-form α. It is a fact that

$$(\alpha,d^*\beta)=(d\alpha,\beta)=\int(d\alpha)_{\underline{I}}\beta^I*1 \tag{37}$$

Now we have the coordinate expressions

$$d\alpha=(d\alpha_{\underline{L}})\wedge dx^L \tag{38}$$

and $(dx^L)_K=\delta^L_K$. It follows that

$$(d\alpha)_{\underline{I}}=(d\alpha_{\underline{L}}\wedge dx^L)_{\underline{I}}=\delta^{j\underline{K}}_{\underline{I}}\frac{\partial\alpha_{\underline{L}}}{\partial x^j}\delta^L_K \tag{39}$$

or

$$(d\alpha)_{\underline{I}}=\delta^{j\underline{K}}_{\underline{I}}\frac{\partial\alpha_K}{\partial x^j} \tag{40}$$

Here we use

$$(\alpha\wedge\beta)_I=\delta^{KL}_I\alpha_K\beta_L \tag{41}$$

where

$$\delta^{KL}_I=\begin{cases}1 & \text{if (KL) is an even permutation of I}\\ -1 & \text{if (KL) is an odd permutation of I}\\ 0 & \text{otherwise}\end{cases} \tag{42}$$

Use of the Leibnitz rule yields

$$\begin{aligned} \int(d\alpha)_{\underline{I}}\beta^I*1 &= \int\delta^{j\underline{K}}_{\underline{I}}\frac{\partial\alpha_K}{\partial x^j}\beta^I*1\\ &= \int\frac{\partial(\delta^{j\underline{K}}_{\underline{I}}\alpha_K\beta^I)}{\partial x^j}*1\\ &\quad-\int\alpha_{\underline{K}}\delta^{j\underline{K}}_{\underline{I}}\frac{\partial\beta^I}{\partial x^j}*1 \end{aligned} \tag{43}$$

The first term corresponds to a surface integration and we can neglect it. We then have $\delta^{jK}_{\underline{I}}\beta^I = \beta^{jK}$ from the antisymmetry of β, so that

$$(\alpha, d^*\beta) = -\int \alpha_{\underline{K}} \frac{\partial \beta^{jK}}{\partial x^j} * 1 = (\alpha, -\mathrm{Div}\beta) \qquad [44]$$

The Maxwell Equations

The Maxwell equations become remarkably concise when expressed in terms of differential forms, namely

$$dF = 0, \qquad d^*F = -j \qquad [45]$$

where F is the field strength and j is the current density. We wish to demonstrate this. We use a $(3+1)$-separation of the exterior derivative into a timelike and a spacelike part:

$$d = \boldsymbol{d} + dt \wedge \frac{\partial}{\partial t} \qquad [46]$$

We then get

$$dF = \left(d\mathcal{E} + \frac{\partial \mathcal{B}}{\partial t}\right) \wedge dt + d\mathcal{B} = 0 \qquad [47]$$

By comparing coefficients, we arrive at

$$d\mathcal{E} = -\frac{\partial \mathcal{B}}{dt}, \qquad d\mathcal{B} = 0 \qquad [48]$$

In vector notation

$$\mathrm{curl}\,\boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \qquad \mathrm{div}\,\boldsymbol{B} = 0 \qquad [49]$$

the usual form of the homogeneous Maxwell equations.

By direct application of the formula [27], one finds

$$*F = -*\mathcal{B} \wedge dt + *\mathcal{E} \qquad [50]$$

where $*$ means the Hodge dual in three space dimensions. One finds

$$d*F = \boldsymbol{d}*\mathcal{E} - \left(\boldsymbol{d}*\mathcal{B} - \frac{\partial *\mathcal{E}}{\partial t}\right) \wedge dt \qquad [51]$$

Therefore,

$$\begin{aligned} d*F = &-(\mathrm{div}\,\boldsymbol{E})dx \wedge dy \wedge dz \\ &+ \left((\mathrm{curl}\,\boldsymbol{B})^x - \frac{\partial E^x}{dt}\right) dy \wedge dz \wedge dt \\ &+ \left((\mathrm{curl}\,\boldsymbol{B})^y - \frac{\partial E^y}{dt}\right) dz \wedge dx \wedge dt \\ &+ \left((\mathrm{curl}\,\boldsymbol{B})^z - \frac{\partial E^z}{dt}\right) dx \wedge dy \wedge dt \end{aligned} \qquad [52]$$

We apply again the Hodge dual:

$$\begin{aligned} *d*F = &-(\mathrm{div}\,\boldsymbol{E})dt + \left((\mathrm{curl}\,\boldsymbol{B})^x - \frac{\partial E^x}{\partial t}\right) dx \\ &+ \left((\mathrm{curl}\,\boldsymbol{B})^y - \frac{\partial E^y}{\partial t}\right) dy \\ &+ \left((\mathrm{curl}\,\boldsymbol{B})^z - \frac{\partial E^z}{\partial t}\right) dz \end{aligned} \qquad [53]$$

In Minkowski space the expression $*d*$ equals the codifferential. Therefore, the equation $d^*F = *d*F = -j$ holds, with j given by $j^\mu = (\rho, J)$, which is equivalent to

$$\mathrm{div}\,\boldsymbol{E} = \rho, \qquad \mathrm{curl}\,\boldsymbol{B} - \frac{\partial \boldsymbol{E}}{\partial t} = \boldsymbol{J} \qquad [54]$$

the inhomogeneous Maxwell equations.

Current Conservation

The electromagnetic 4-current is

$$j^\mu = \rho_0 u^\mu = (\rho_0\gamma, \rho_0\gamma\boldsymbol{v}) = (\rho, \boldsymbol{J}) \qquad [55]$$

where ρ is the charge density and $\boldsymbol{J}$ the current density. This corresponds to a 1-form

$$j = \rho dt - J^x dx - J^y dy - J^z dz \qquad [56]$$

The Hodge dual is $*j = \sigma^3 - j^2 \wedge dt$, with the 3-form $\sigma^3 = \rho dx \wedge dy \wedge dz$, and the 2-form

$$j^2 = -J^x dy \wedge dz - J^y dz \wedge dx - J^z dx \wedge dy \qquad [57]$$

From the Maxwell equation $d^*F = -j$, it follows that

$$(d^*)^2 F = -d^*j = 0 \qquad [58]$$

that is

$$\begin{aligned} *d(*j) &= *d(\sigma^3 - j^2 \wedge dt) = *(d\sigma^3 - dj^2 \wedge dt) \\ &= * \left(\frac{\partial \rho}{\partial t} + \mathrm{div}\boldsymbol{J}\right) dt \wedge dx \wedge dy \wedge dz \\ &= \frac{\partial \rho}{\partial t} + \mathrm{div}\boldsymbol{J} = 0 \end{aligned} \qquad [59]$$

This is the "continuity equation."

The total charge inside a volume V is $Q = \int_V \rho dV$, therefore

$$-\frac{dQ}{dt} = -\frac{d}{dt}\int_V \rho\, dV = \int_{\partial V} \boldsymbol{J} \cdot \boldsymbol{n}\, dS \qquad [60]$$

where ∂V is the surface which encloses the volume V, dS is the surface element, and $\boldsymbol{n}$ is the normal vector to this surface. This is current conservation.

The Gauge Potential

The "Poincaré lemma" tells us that $dF = 0$ implies $F = dA$, with the 4-potential A:

$$A = \phi dt + \boldsymbol{A} \qquad [61]$$

and the vector potential $\boldsymbol{A} = A_x dx + A_y dy + A_z dz$. From

$$F = \mathcal{E} \wedge dt + \mathcal{B} = \left(d + dt \wedge \frac{\partial}{\partial t}\right) A$$
$$= d\phi \wedge dt + d\boldsymbol{A} + dt \wedge \frac{\partial \boldsymbol{A}}{\partial t} \qquad [62]$$

it follows by comparing coefficients that

$$\mathcal{E} = d\phi - \frac{\partial \boldsymbol{A}}{\partial t}, \qquad \mathcal{B} = d\boldsymbol{A} \qquad [63]$$

In vector notation this is

$$\boldsymbol{E} = \mathrm{grad}\phi - \frac{\partial \boldsymbol{A}}{\partial t}, \qquad \boldsymbol{B} = \mathrm{curl}\,\boldsymbol{A} \qquad [64]$$

The 4-potential is determined up to a gauge function Λ:

$$A' = A + d\Lambda \qquad [65]$$

This gauge freedom has no influence on the observable quantities $\boldsymbol{E}$ and $\boldsymbol{B}$:

$$F' = dA' = dA + d^2\Lambda = dA = F \qquad [66]$$

The Laplace operator is $\triangle = (d^* + d)^2 = dd^* + d^*d$, so when the 4-potential A fulfills the condition $d^*A = 0$, we have

$$\triangle A = d^*dA = d^*F = -j \qquad [67]$$

the "classical wave equation." The condition $d^*A = 0$ is called the "Lorentz gauge condition." This condition can always be fulfilled by using the gauge freedom: $d^*(A + d\Lambda) = 0$ is fulfilled when $d^*d\Lambda = \triangle\Lambda = -d^*A$, where we have used the fact that $d^*\Lambda = 0$ for functions. That is to say, $d^*A = 0$ is fulfilled when Λ is a solution of the inhomogeneous wave equation.

Gauge Invariance

In quantum mechanics, the electron is described by a wave function which is determined up to a free phase. Indeed, at every point in space this phase can be chosen arbitrarily:

$$\begin{aligned} \psi(x) &\to \psi'(x) = \exp\{i\alpha(x)\}\psi(x) \\ \bar{\psi}(x) &\to \bar{\psi}'(x) = \bar{\psi}(x)\exp\{-i\alpha(x)\} \end{aligned} \qquad [68]$$

with the only condition being that $\alpha(x)$ is a continuous function. The gauge transformation is of the form $g = \exp\{i\alpha(x)\}$, with g an element of the abelian gauge group $G = U(1)$. The free action is

$$S_0 = \int \mathcal{L}_0 \, d^4x \qquad [69]$$

with

$$\mathcal{L}_0 = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi \qquad [70]$$

the "Lagrange density." This action is not invariant under gauge transformations:

$$\mathcal{L}_0 \to \mathcal{L}_0' = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi - (\partial_\mu \alpha)\bar{\psi}\gamma^\mu \psi \qquad [71]$$

The undesired term can be compensated by the introduction of a gauge potential ω in a covariant derivative of ψ,

$$D\psi = (d + \omega)\psi \qquad [72]$$

which has the desired transformation property $D\psi \to \exp\{i\alpha\}D\psi$ when besides the transformation $\psi(x) \to \exp\{i\alpha(x)\}\psi(x)$ of the matter field the gauge potential simultaneously transforms according to the gauge transformation $\omega \to \omega - id\alpha$. The new Lagrange density is

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi = \mathcal{L}_0 + i\omega_\mu \bar{\psi}(x)\gamma^\mu \psi(x) \qquad [73]$$

The substitution $\partial_\mu \to D_\mu$ is known to physicists; with $\omega = -iqA$ it is the ansatz of minimal coupling for taking into account electromagnetic effects: $\partial_\mu \to \partial_\mu - iqA_\mu$. The Lagrange density becomes in this notation $\mathcal{L} = \mathcal{L}_0 - A_\mu J^\mu$, where $J^\mu = -q\bar{\psi}\gamma^\mu \psi$.

The Lagrange density must now be completed by a kinetic term for the gauge potential and we get the complete electromagnetic Lagrange density

$$\mathcal{L} = \mathcal{L}_0 - A_\mu J^\mu - \tfrac{1}{4} F_{\mu\nu} F^{\mu\nu} \qquad [74]$$

with $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. In the action this corresponds to

$$S = S_0 - \int_M A_\mu J^\mu \mathrm{vol}^4 - \frac{1}{4}\int_M F_{\mu\nu} F^{\mu\nu} \mathrm{vol}^4 \qquad [75]$$

We get the field equations for the potential A by demanding that the variation of the action vanishes:

$$\delta S[A] = -\int_M \delta A_\mu J^\mu \mathrm{vol}^4 - \frac{1}{4}\delta \int_M F_{\mu\nu} F^{\mu\nu} \mathrm{vol}^4 \qquad [76]$$

We write now

$$\int_M \delta A_\mu J^\mu \mathrm{vol}^4 = (\delta A, j) \qquad [77]$$

and

$$\begin{aligned}\frac{1}{4}\delta\int_M F_{\mu\nu}F^{\mu\nu}\mathrm{vol}^4 \\ =\frac{1}{2}\delta\int_M F\wedge *F=\frac{1}{2}\delta(F,F) \\ =(\delta dA,F)=(d\delta A,F)=(\delta A,d^*F)\end{aligned} \quad [78]$$

where we have exchanged the action of δ and d. Since this holds for arbitrary variations δA we find

$$d^*F=-j \quad [79]$$

the inhomogeneous Maxwell equation.

Nonabelian Gauge Theories

In SU(N) gauge theory the elementary particles are taken to be members of symmetry multiplets. For example, in electroweak theory the left-handed electron and the neutrino are members of an SU(2) doublet:

$$\psi=\begin{pmatrix}e^-\\ \nu\end{pmatrix} \quad [80]$$

A gauge transformation is

$$\psi'(x)=g^{-1}(x)\psi(x),\qquad \bar{\psi}'(x)=\bar{\psi}(x)g(x) \quad [81]$$

with

$$g(x)=\exp\{\Lambda(x)\} \quad [82]$$

where $g(x)$ is an element of the Lie group SU(2) and Λ is an element of the Lie algebra su(2). The Lie algebra is a vector space, and its elements may be expanded in terms of a basis:

$$\Lambda(x)=\Lambda^a(x)T_a \quad [83]$$

For su(2) the basis elements are traceless and anti-Hermitian (see below), they are conventionally expressed in terms of the Pauli matrices,

$$T_a=\frac{\sigma_a}{2\mathrm{i}} \quad [84]$$

with

$$\begin{aligned}\sigma_1=\begin{pmatrix}0&1\\1&0\end{pmatrix},\qquad \sigma_2=\begin{pmatrix}0&-\mathrm{i}\\ \mathrm{i}&0\end{pmatrix} \\ \sigma_3=\begin{pmatrix}1&0\\0&-1\end{pmatrix}\end{aligned} \quad [85]$$

They are conventionally normalized according to

$$\mathrm{tr}(T_aT_b)=-\tfrac{1}{2}\delta_{ab} \quad [86]$$

The Dirac Lagrangian is not invariant with respect to local gauge transformations:

$$\begin{aligned}\mathcal{L}_0&=\bar{\psi}(\mathrm{i}\gamma^\mu\partial_\mu-m)\psi\rightarrow\mathcal{L}_0' \\ &=\mathcal{L}_0+\mathrm{i}\bar{\psi}\gamma^\mu(g\partial_\mu g^{-1})\psi\end{aligned} \quad [87]$$

We introduce the gauge potential

$$\omega_\mu(x)=\omega_\mu^a(x)T_a \quad [88]$$

with a gauge transformation

$$\omega_\mu\rightarrow\omega_\mu'=g^{-1}\omega_\mu g+g^{-1}\partial_\mu g \quad [89]$$

The Lagrange density is modified through a covariant derivative:

$$\partial_\mu\rightarrow D_\mu=\partial_\mu+\omega_\mu \quad [90]$$

The covariant derivative D_μ transforms according to

$$D_\mu\rightarrow D_\mu'=g^{-1}D_\mu g \quad [91]$$

and thus the modified Lagrange density

$$\mathcal{L}=\bar{\psi}(\mathrm{i}\gamma^\mu D_\mu-m)\psi=\mathcal{L}_0+\mathrm{i}\bar{\psi}\gamma^\mu\omega_\mu\psi \quad [92]$$

is invariant with respect to local gauge transformations.

The extra term in the Langrange density is conventionally written

$$-J_a^\mu A_\mu^a \quad [93]$$

with

$$A_\mu^a=-\mathrm{i}q\omega_\mu^a \quad [94]$$

and

$$J_a^\mu=\bar{\psi}\gamma^\mu T_a\psi \quad [95]$$

In mathematical terminology ω is called a connection. The quantity A is the physicist's gauge potential. The connection is anti-Hermitian and the gauge potential Hermitian. The gauge potential also includes the coupling constant q. We will refer to both ω and A as the gauge potential, where the relation between them is given by eqn [94].

We can write the gauge potential as $A=A_\mu^a dx^\mu T_a$ or, in the SU(2) case, as

$$A_\mu=A_\mu^1T_1+A_\mu^2T_2+A_\mu^3T_3 \quad [96]$$

where we see explicitly that it involves three vector fields, which couple to the electroweak currents [95] with the single coupling constant q, and which will become after symmetry breaking the three vector bosons W_+, W_-, Z_0 of the electroweak gauge theory. Actually, a mix of the neutral gauge boson and the photon will combine to yield the Z_0 boson, while the orthogonal mixture gives rise to the electromagnetic interaction, in an SU(2) $\times$ U(1) theory. At this stage,

the gauge bosons are all massless, their masses are generated by the "Higgs' mechanism."

Lie-Algebra-Valued *p*-Forms

To describe nonabelian fields, we need Lie-algebra-valued p-forms:

$$\phi = T_a \phi^a \qquad [97]$$

where T_a is a generator of the Lie algebra, the index a runs over the number of generators of the Lie algebra, and the ϕ^a are the usual scalar-valued p-forms. The composition in a Lie algebra is a Lie bracket, which is defined for two Lie-algebra-valued p-forms by

$$[\phi, \psi] := [T_a, T_b]\phi^a \wedge \psi^b \qquad [98]$$

The Lie bracket in the algebra is

$$[T_a, T_b] = f_{ab}^c T_c \qquad [99]$$

where f_{bc}^a are the structure constants. It follows from this that

$$[\psi, \phi] = [T_a, T_b]\psi^a \wedge \phi^b = -[T_b, T_a]\psi^a \wedge \phi^b \qquad [100]$$

or

$$[\psi, \phi] = (-1)^{pq+1}[\phi, \psi] \qquad [101]$$

when ϕ is a p-form and ψ is a q-form. In the special case that T_a is a matrix, also the product $T_a T_b$ is defined, and from this the product of two Lie-algebra-valued p-forms

$$\phi \wedge \psi = T_a \phi^a \wedge T_b \psi^b = T_a T_b \phi^a \wedge \psi^b \qquad [102]$$

Now the Lie bracket is a commutator:

$$[T_a, T_b] = T_a T_b - T_b T_a \qquad [103]$$

and

$$\begin{aligned}[\phi, \psi] &= [T_a, T_b]\phi^a \wedge \psi^b \\ &= T_a \phi^a \wedge T_b \psi^b - (-1)^{pq} T_b \psi^b \wedge T_a \phi^a \\ &= \phi \wedge \psi - (-1)^{pq} \psi \wedge \phi \end{aligned} \qquad [104]$$

From this relation it follows that for ϕ and ψ odd p-forms

$$[\phi, \psi] = \phi \wedge \psi + \psi \wedge \phi \qquad [105]$$

For ϕ an odd p-form

$$[\phi, \phi] = \phi \wedge \phi + \phi \wedge \phi = 2(\phi \wedge \phi) \qquad [106]$$

The Gauge Potential and the Field Strength

The generalization of the abelian relationship between the gauge potential and the field strength, $F = dA$, is

$$\theta = d\omega + \tfrac{1}{2}[\omega, \omega] = d\theta + \omega \wedge \omega \qquad [107]$$

where because ω is a 1-form we can use eqn [106]. The mathematician refers to θ as the curvature. The physicist writes, in analogy to eqn [94],

$$F = -\mathrm{i}\, q\theta = \tfrac{1}{2} F_{\mu\nu}^a dx^\mu \wedge dx^\nu T_a \qquad [108]$$

One obtains for the components

$$F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a - \mathrm{i}q f_{bc}^a A_\mu^b A_\nu^c \qquad [109]$$

A generalization of the gauge transformation of A, that is, $A' = A + d\Lambda$, is eqn [89]:

$$\omega' = g^{-1}\omega g + g^{-1} dg \qquad [110]$$

A quantity ϕ with the transformation property

$$\phi' = g^{-1}\phi g \qquad [111]$$

is called a "tensorial" quantity. The gauge potential ω is according to this definition nontensorial. Nevertheless the field strength is tensorial. Indeed

$$\begin{aligned}\theta' &= d(g^{-1}\omega g) + (dg^{-1}) \wedge dg \\ &\quad + \tfrac{1}{2}[g^{-1}\omega g + g^{-1} dg, g^{-1}\omega g + g^{-1} dg] \\ &= (dg^{-1}) \wedge \omega g + g^{-1} d\omega g - g^{-1}\omega \wedge dg + (dg^{-1}) \wedge dg \\ &\quad + \tfrac{1}{2} g^{-1}[\omega, \omega] g + \tfrac{1}{2}[g^{-1}\omega g, g^{-1} dg] \\ &\quad + \tfrac{1}{2}[g^{-1} dg, g^{-1}\omega g] + \tfrac{1}{2}[g^{-1} dg, g^{-1} dg] \\ &= g^{-1}\theta g + (dg^{-1}) \wedge \omega g - g^{-1}\omega \wedge dg + (dg^{-1}) \wedge dg \\ &\quad + g^{-1}\omega \wedge dg + g^{-1} dg \wedge g^{-1}\omega g + g^{-1} dg \wedge g^{-1} dg \\ &= g^{-1}\theta g \end{aligned} \qquad [112]$$

where we have used the derivation of the relation $g^{-1}g = \mathrm{Id}$ to get

$$dg^{-1} = -g^{-1} dg\, g^{-1} \qquad [113]$$

In the abelian case, we had $dF = 0$. The nonabelian analog is

$$\begin{aligned} d\theta &= d\omega \wedge \omega - \omega \wedge d\omega \\ &= (\theta - \omega \wedge \omega) \wedge \omega - \omega \wedge (\theta - \omega \wedge \omega) \\ &= \theta \wedge \omega - \omega \wedge \theta \end{aligned} \qquad [114]$$

or

$$d\theta + \omega \wedge \theta - \theta \wedge \omega = 0 \qquad [115]$$

the Bianchi identity. It can also be written as

$$d\theta + \omega \wedge \theta - \theta \wedge \omega = d\theta + [\omega, \theta] = 0 \qquad [116]$$

because from eqn [104]

$$\omega \wedge \theta + (-1)^{2\cdot 1}\theta \wedge \omega = [\omega, \theta] \qquad [117]$$

The covariant derivative D is defined as

$$D\phi := d\phi + [\omega, \phi] \qquad [118]$$

for ϕ a tensorial quantity. The covariant derivative takes tensorial p-forms into tensorial $(p+1)$-forms:

$$\begin{aligned} D'\phi' &= d(g^{-1}\phi g) + [g^{-1}\omega g + g^{-1}dg, g^{-1}\phi g] \\ &= dg^{-1} \wedge \phi g + g^{-1}d\phi g + (-1)^p g^{-1}\phi \wedge dg \\ &\quad + [g^{-1}\omega g, g^{-1}\phi g] + [g^{-1}dg, g^{-1}\phi g] \\ &= g^{-1}D\phi g + dg^{-1} \wedge \phi g + (-1)^p g^{-1}\phi \wedge dg \\ &\quad + g^{-1}dgg^{-1} \wedge \phi g - (-1)^p g^{-1}\phi \wedge dg \\ &= g^{-1}D\phi g \end{aligned} \qquad [119]$$

We have thereby verified the transformation property of eqn [91].

The Gauge Group

From the gauge transformation $\psi' = g\psi$ the requirement $|\psi'|^2 = |\psi|^2$ leads to $g^\dagger g = 1$. That means that g belongs to the unitary Lie group $G = \mathrm{U}(n)$, whose elements fulfill $g^\dagger = \bar{g}^{\mathrm{T}} = g^{-1}$. For elements of the Lie algebra $\mathcal{G} = \mathrm{u}(n)$ this implies

$$(e^X)^\dagger = e^{\bar{X}^{\mathrm{T}}} = e^{-X} \qquad [120]$$

or

$$X^\dagger = \bar{X}^{\mathrm{T}} = -X \qquad [121]$$

where $\bar{X}$ is complex conjugation and X^{T} means transposition.

For elements of the Lie algebra we can define a scalar product (the Killing metric)

$$\langle X, Y \rangle := -\mathrm{tr}\,(XY) = -X^\alpha{}_\beta X^\beta{}_\alpha \qquad [122]$$

The scalar product is real:

$$\langle \bar{X}, \bar{Y} \rangle = -\bar{X}^\beta{}_\alpha \bar{Y}^\alpha{}_\beta = -X^\alpha{}_\beta X^\beta{}_\alpha = \langle X, Y \rangle \qquad [123]$$

symmetric:

$$\langle X, Y \rangle = -\mathrm{tr}(X, Y) = -\mathrm{tr}(Y, X) = \langle Y, X \rangle \qquad [124]$$

and positive definite:

$$\langle X, X \rangle = -X^\alpha{}_\beta X^\beta{}_\alpha = X^\alpha{}_\beta \bar{X}^\alpha{}_\beta = |X^\alpha{}_\beta|^2 \qquad [125]$$

The scalar product is invariant under the action of G on $\mathcal{G}$: for $g \in G$

$$\begin{aligned} \langle gXg^{-1}, gYg^{-1} \rangle &= -\mathrm{tr}\,(gXYg^{-1}) \\ &= -\mathrm{tr}(X, Y) = \langle X, Y \rangle \end{aligned} \qquad [126]$$

or for $X, Y, Z \in \mathcal{G}$

$$\langle e^{tX} Y e^{-tX}, e^{tX} Z e^{-tX} \rangle = \langle Y, Z \rangle \qquad [127]$$

We take the derivative of this equation with respect to t at the value $t = 0$ and get:

$$\langle [X, Y], Z \rangle + \langle Y, [X, Z] \rangle = 0 \qquad [128]$$

We define an action of the algebra $\mathcal{G}$ on itself: $ad(X): \mathcal{G} \to \mathcal{G}$

$$ad(X)Y = [X, Y] \qquad [129]$$

We can then formulate our conclusion as follows: the action of $\mathcal{G}$ on itself is anti-Hermitian:

$$\langle ad(X)Y, Z \rangle = -\langle Y, ad(X)Z \rangle \qquad [130]$$

or

$$[ad(X)]^\dagger = -ad(X) \qquad [131]$$

From $g^\dagger g = 1$ we have $|\det(g)|^2 = 1$. For the gauge group $G = \mathrm{SU}(N)$ we require in addition $\det(g) = 1$. Since

$$\det(g) = \det(\exp(X)) = \exp(\mathrm{tr}(X)) \qquad [132]$$

the elements $X \in \mathrm{su}(N)$ must be traceless. A basis of the vector space of traceless, anti-Hermitian (2×2) matrices is given by the Pauli matrices, eqn [85].

The Yang–Mills Action

The SU(2) Yang–Mills action is, in analogy to the abelian case,

$$\begin{aligned} S &= -\frac{1}{4q^2}\int_M F^a_{\mu\nu} F^{a\mu\nu} \mathrm{vol}^4 = \frac{1}{2q^2}\int_M \mathrm{tr}(F_{\mu\nu}F^{\mu\nu})\mathrm{vol}^4 \\ &= \frac{1}{2q^2}\int_M \mathrm{tr}(F \wedge *F) \end{aligned} \qquad [133]$$

We have included the trace in our definition of the scalar product:

$$(\phi, \psi) := -\int_M \mathrm{tr} < \phi_I \psi^I > \mathrm{vol}^n = -\int_M \mathrm{tr}(\phi \wedge *\psi) \qquad [134]$$

We then write eqn [133] as

$$S[\omega] = \tfrac{1}{2}(\theta, \theta) \qquad [135]$$

taking into account the relation between θ and the field strength F, and indicating the dependence on

the gauge potential. Since θ is tensorial the action is invariant.

Now we calculate the variation von $S[\omega]$ with respect to a variation of the gauge potential:

$$\begin{aligned}\delta S[\omega] &= \frac{d}{dt} S[\omega(t)]|_{t=0} = \frac{1}{2}\delta(\theta,\theta) \\ &= \frac{1}{2}((\delta\theta,\theta)+(\theta,\delta\theta)) \\ &= (\delta\theta,\theta) = \left(\delta\left(d\omega+\frac{1}{2}[\omega,\omega]\right),\theta\right) \\ &= \left(\delta d\omega+\frac{1}{2}[\delta\omega,\omega]+\frac{1}{2}[\omega,\delta\omega],\theta\right) \\ &= (d\delta\omega+[\omega,\delta\omega],\theta) \qquad [136]\end{aligned}$$

where we have exchanged the order of δ and d. We remark that although ω is not a tensorial section, $\delta\omega$ is: for $\omega_1' = g^{-1}\omega_1 g + g^{-1}dg$ and $\omega_2' = g^{-1}\omega_2 g + g^{-1}dg$ is

$$\delta\omega = \omega_1' - \omega_2' = g^{-1}(\omega_1-\omega_2)g \qquad [137]$$

The quantity θ is in any case tensorial. Therefore, the covariant derivative is defined, and we have

$$D\delta\omega = d\delta\omega + [\omega,\delta\omega] \qquad [138]$$

and

$$D\theta = d\theta + [\omega,\theta] \qquad [139]$$

In general, the action of the covariant derivative on tensorial quantities can be written as $D = d + ad(\omega)$, where $ad(X)$ is the representation of the Lie algebra on itself introduced in the previous section. We now have

$$\delta S[\omega] = (D\delta\omega,\theta) = (\delta\omega, D^*\theta) = 0 \qquad [140]$$

for an arbitrary variation $\delta\omega$. Therefore, $D^*\theta = 0$.

We have obtained

$$D^*\theta = 0 \qquad [141]$$

the "Yang–Mills equations," and

$$D\theta = 0 \qquad [142]$$

the "Bianchi identites." These are the generalizations of the Maxwell equations $d^*F = 0$ and $dF = 0$ in the absence of external sources. For the general case of interacting fermions, we write out the full action, in analogy to eqn [74], and obtain, in analogy to eqns [79] and [58],

$$D^*\theta = -J, \qquad D^*J = 0 \qquad [143]$$

We shall now derive, again for the pure gauge sector, coordinate expressions for the Yang–Mills equations. Consider the expression

$$\begin{aligned}\delta S[\omega] &= (D\delta\omega,\theta) = (\delta\omega, D^*\theta) \\ &= (d\delta\omega+[\omega,\delta\omega],\theta)\,. \qquad [144]\end{aligned}$$

The first term in the last expression is

$$(d\delta\omega,\theta) = (\delta\omega, d^*\theta) = -\mathrm{tr}\int_M \delta\omega_\nu \{d^*\theta\}^\nu \mathrm{vol}^4 \qquad [145]$$

The second term can be computed using

$$\begin{aligned}[\omega,\delta\omega]_{\mu\nu} &= \{\omega\wedge\delta\omega+\delta\omega\wedge\omega\}(\partial_\mu,\partial_\nu) \\ &= \omega_\mu\delta\omega_\nu - \omega_\nu\delta\omega_\mu + \delta\omega_\mu\omega_\nu - \delta\omega_\nu\omega_\mu \qquad [146]\end{aligned}$$

and hence

$$[\omega,\delta\omega]_{\mu\nu}\theta^{\mu\nu} = 2[\omega_\mu,\delta\omega_\nu]\theta^{\mu\nu} \qquad [147]$$

because θ is antisymmetric, $\theta^{\mu\nu} = -\theta^{\nu\mu}$. Thus,

$$\begin{aligned}([\omega,\delta\omega],\theta) &= -\int_M \mathrm{tr}([\omega,\delta\omega]\wedge *\theta) \\ &= -\frac{1}{2}\int_M \mathrm{tr}([\omega,\delta\omega]_{\mu\nu}\theta^{\mu\nu})\mathrm{vol}^4 \\ &= -\int_M \mathrm{tr}([\omega_\mu,\delta\omega_\nu]\theta^{\mu\nu})\mathrm{vol}^4 \\ &= \int_M \langle[\omega_\mu,\delta\omega_\nu],\theta^{\mu\nu}\rangle\mathrm{vol}^4 \qquad [148]\end{aligned}$$

where $\langle\,,\rangle$ is the scalar product in $\mathcal{G}$. From eqn [128] this equals

$$\begin{aligned}&-\int_M \langle\delta\omega_\nu,[\omega_\mu,\theta^{\mu\nu}]\rangle\mathrm{vol}^4 \\ &= \int_M \mathrm{tr}(\delta\omega_\nu[\omega_\mu,\theta^{\mu\nu}])\mathrm{vol}^4 \qquad [149]\end{aligned}$$

Combining this with eqn [144] gives

$$\begin{aligned}(\delta\omega, D^*\theta) &= -\int_M \mathrm{tr}(\delta\omega_\nu\{(d^*\theta)^\nu - [\omega_\mu,\theta^{\mu\nu}]\})\mathrm{vol}^4 \\ &= (\delta\omega,\{(d^*\theta)^\nu - [\omega_\mu,\theta^{\mu\nu}]\}) \qquad [150]\end{aligned}$$

We can now insert the coordinate expression for

$$(d\theta)^\nu = -\partial_\mu\theta^{\mu\nu} \qquad [151]$$

Finally, the coordinate expressions of the Yang–Mills equations $D^*\theta = 0$ are

$$(D^*\theta)^\nu = -\{\partial_\mu\theta^{\mu\nu} + [\omega_\mu,\theta^{\mu\nu}]\} = 0 \qquad [152]$$

The Analogy with Electromagnetism

The Yang–Mills equation and the Bianchi identity in the absence of external sources are

$$\partial_\nu F^{\mu\nu} - iq[A_\nu, F^{\mu\nu}] = 0 \qquad [153]$$

and

$$\begin{aligned}&\partial_\mu F_{\nu\tau} + \partial_\tau F_{\mu\nu} + \partial_\nu F_{\tau\mu} - iq\{[A_\mu,F_{\nu\tau}] \\ &\quad + [A_\tau,F_{\mu\nu}] + [A_\nu,F_{\tau\mu}]\} = 0 \qquad [154]\end{aligned}$$

We shall write these equations in terms of the fields

$$F^{i0} = E^i, \quad i = 1, 2, 3 \qquad [155]$$

$$F^{12} = B^3, \qquad F^{31} = B^2, \qquad F^{12} = B^3 \qquad [156]$$

where the **E** and **B** vectors may be thought of as "electric" and "magnetic" fields, even though they have Lie-algebra indices, $F^{i0} = (F^a)^{i0} T_a$, etc. In the context of the SU(3) theory, they are referred to as the "chromoelectric" and "chromomagnetic" fields, respectively.

The Yang–Mills equations with $\mu = 0$ are

$$\partial_i F^{i0} - \mathrm{i}q[A_i, F^{i0}] = 0 \qquad [157]$$

with $i = 1, 2, 3$ a spatial index. In vector notation this is

$$\operatorname{div} \boldsymbol{E} = \mathrm{i}q(\boldsymbol{A} \cdot \boldsymbol{E} - \boldsymbol{E} \cdot \boldsymbol{A}) \qquad [158]$$

This is the analog of Gauss's equation. Even though we started out without external sources, $\mathrm{i}q(\boldsymbol{A} \cdot \boldsymbol{E} - \boldsymbol{E} \cdot \boldsymbol{A})$ plays the role of a "charge density." The Yang–Mills field $\boldsymbol{E}$ and the potential $\boldsymbol{A}$ combine to act as a source for the Yang–Mills field. This is an essential feature of nonabelian gauge theories in which they differ from the abelian case, due to the fact that the commutator $[\boldsymbol{A}, \boldsymbol{E}]$ is nonvanishing.

Now consider the Yang–Mills equations with a spatial index $\mu = i$:

$$\partial_0 F^{i0} + \partial_j F^{ij} - \mathrm{i}q[A_0, F^{i0}] - \mathrm{i}q[A_j F^{ij}] = 0 \qquad [159]$$

In vector notation this is

$$\operatorname{curl} \boldsymbol{B} = \frac{\partial \boldsymbol{E}}{\partial t} - \mathrm{i}q(A_0 \boldsymbol{E} - \boldsymbol{E} A_0) + \mathrm{i}q(\boldsymbol{A} \times \boldsymbol{B} + \boldsymbol{B} \times \boldsymbol{A}) \qquad [160]$$

replacing the Ampere–Maxwell law. Note that there are two extra contributions to the "current" other than the displacement current.

The analogs of the laws of Faraday and of the absence of magnetic monopoles are derived similarly from the Bianchi identities. The results are

$$\operatorname{curl} \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = \mathrm{i}q\{(\boldsymbol{A} \times \boldsymbol{E} + \boldsymbol{E} \times \boldsymbol{A}) + (A_0 \boldsymbol{B} - \boldsymbol{B} A_0)\} \qquad [161]$$

and

$$\operatorname{div} \boldsymbol{B} = \mathrm{i}q(\boldsymbol{A} \cdot \boldsymbol{B} - \boldsymbol{B} \cdot \boldsymbol{A}) \qquad [162]$$

Further Remarks

The foundations of the mathematics of differential forms were laid down by Poincaré (1953). They were applied to the description of electrodynamics already by Cartan (1923). A modern presentation of differential forms and the manifolds on which they are defined is given in Abraham *et al.* (1983). A recent treatment of electrodynamics in this approach is Hehl and Obukhov (2003). Weyl's argument is in his paper of 1929.

Nonabelian gauge theories today explain the electromagnetic, the strong and weak nuclear interactions. The original paper is that of Yang and Mills (1954). Glashow, Salam, and Weinberg (1980) saw the way to apply it to the weak interactions by using spontaneous symmetry breaking to generate the masses through the use of the Higgs' (1964) mechanism. t'Hooft and Veltman (1972) showed that the resulting quantum field theory was renormalizable. The strong interactions were recognized as the nonabelian gauge theory with gauge group SU(3) by Gell-Mann (1972). For a modern treatment which puts nonabelian gauge theories in the context of differential geometry, see Frankel (1987).

See also: Dirac Fields in Gravitation and Nonabelian Gauge Theory; Electroweak Theory; Measure on Loop Spaces; Nonperturbative and Topological Aspects of Gauge Theory; Quantum Electrodynamics and its Precision Tests.

Further Reading

Abraham A, Marsden J, and Ratiu T (1983) *Manifolds, Tensor Analysis, and Applications.* MA: Addison-Wesley.

Cartan É (1923) *On manifolds with an Affine Connection and the Theory of General Relativity.* English translation of the French original 1923/1924 (Bibliopolis, Napoli 1986).

Frankel T (1987) *The Geometry of Physics, An Introduction.* Cambridge University Press.

Gell-Mann M (1972) Quarks: developments in the quark theory of hadrons. *Acta Physica Austriaca Suppl.* IV: 733.

Glashow SL (1980) Towards a unified theory: threads in a tapestry. *Reviews of Modern Physics* 52: 539.

Hehl FW and Obukhov YN (2003) *Foundations of Classical Electrodynamics.* Boston: Birkhäuser.

Higgs PW (1964) Broken symmetries and the masses of gauge bosons. *Physical Review Letters* 13: 508.

t'Hooft G and Veltman M (1972) Regularization and renormalization of gauge fields. *Nuclear Physics B* 44: 189.

Poincaré H (1953) *Oeuvre.* Paris: Gauthier-Villars.

Salam A (1980) Gauge unification of fundamental forces. *Reviews of Modern Physics* 52: 525.

Weinberg SM (1980) Conceptual foundations of the unified theory of weak and electromagnetic interactions. *Reviews of Modern Physics* 52: 515.

Weyl H (1929) Elektron und gravitation. *Zeitschrift fuer Physik* 56: 330.

Yang CN and Mills RL (1954) Construction of isotopic spin and isotopic gauge invariance. *Physical Review* 96: 191.

Abelian Higgs Vortices

J M Speight, University of Leeds, Leeds, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

For the purpose of this article, vortices are topological solitons arising in field theories in $(2+1)$-dimensional spacetime when a complex-valued field ϕ is allowed to acquire winding at infinity, meaning that the phase of $\phi(t,\boldsymbol{x})$, as $\boldsymbol{x}$ traverses a large circle in the spatial plane, changes by $2\pi n$, where n is a nonzero integer. Such winding cannot be removed by any continuous deformation of ϕ (hence "topological") and traps a considerable amount of energy which tends to coalesce into smooth, stable lumps with highly particle-like characteristics (hence "solitons"). Clearly, the universe is $(3+1)$ dimensional. Nonetheless, planar field theories are of physical interest for two main reasons. First, the theory may arise by dimensional reduction of a $(3+1)$-dimensional model under the assumption of translation invariance in one direction. Vortices are then transverse slices through straight tube-like objects variously interpreted as magnetic flux tubes in a superconductor or cosmic strings. Second, a crucial ingredient of the standard model of particle physics is spontaneous breaking of gauge symmetry by a Higgs field. As well as endowing the fundamental gauge bosons and chiral fermions with mass, this mechanism can potentially generate various types of topological solitons (monopoles, strings, and domain walls) whose structure and interactions one would like to understand. Vortices in $(2+1)$ dimensions are interesting in this regard because they arise in the simplest field theory exhibiting the Higgs mechanism, the abelian Higgs model (AHM). They are thus a useful theoretical laboratory in which to test ideas which may ultimately find application in more realistic theories. This article describes the properties of abelian Higgs vortices and explains how, using a mixture of numerical and analytical techniques, a good understanding of their dynamical interactions has been obtained.

The Abelian Higgs Model

Throughout this article spacetime will be $\mathbb{R}^{2+1}$ endowed with the Minkowski metric with signature $(+,-,-)$, and Cartesian coordinates $x^\mu, \mu = 0,1,2$, with $x^0 = t$ (the speed of light $c=1$). A spacetime point will be denoted x, its spatial part by $\boldsymbol{x}=(x^1,x^2)$. Latin indices $j,k,\ldots$ range over 1, 2, and repeated indices (Latin or Greek) are summed over. We sometimes use polar coordinates in the spatial plane, $\boldsymbol{x}=r(\cos\theta,\sin\theta)$, and sometimes a complex coordinate $z=x^1+\mathrm{i}x^2=re^{\mathrm{i}\theta}$. Occasionally, it is convenient to think of $\mathbb{R}^{2+1}$ as a subspace of $\mathbb{R}^{3+1}$ and denote by $\boldsymbol{k}$ the unit vector in the (fictitious) third spatial direction. The complex scalar Higgs field is denoted ϕ, and the electromagnetic gauge potential A_μ, best thought of as the components of a 1-form $A=A_\mu\,\mathrm{d}x^\mu$. $F_{\mu\nu}=\partial_\mu A_\nu-\partial_\nu A_\mu$ is the field strength tensor which, in $\mathbb{R}^{2+1}$, has only three independent components, identified with the magnetic field $B=F_{12}$ and electric field $(E_1,E_2)=(F_{01},F_{02})$. The gauge-covariant derivative is $D_\mu\phi=\partial_\mu\phi-ieA_\mu\phi$, e being the electric charge of the Higgs. Under a U(1) gauge transformation,

$$\phi\mapsto e^{\mathrm{i}\Lambda}\phi, \qquad A_\mu\mapsto A_\mu+e^{-1}\partial_\mu\Lambda \tag{1}$$

$\Lambda:\mathbb{R}^{2+1}\to\mathbb{R}$ being any smooth function, $F_{\mu\nu}$ and $|\phi|$ remain invariant, while $D_\mu\phi\mapsto e^{\mathrm{i}\Lambda}D_\mu\phi$. Only gauge-invariant quantities are physically observable (classically).

With these conventions, the AHM has Lagrangian density

$$\mathcal{L}=-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}+\frac{\alpha}{2}D_\mu\phi\overline{D^\mu\phi}-\frac{\lambda}{8}(\nu^2-|\phi|^2)^2 \tag{2}$$

which is manifestly gauge invariant. By rescaling ϕ, A_μ, x and the unit of action, we can (and henceforth will) assume that $e=\nu=\alpha=1$. The only parameter which cannot be scaled away is $\lambda>0$. Its value greatly influences the model's behavior.

The field equations, obtained by demanding that $\phi(x), A_\mu(x)$ be a local extremal of the action $S=\int\mathcal{L}d^3x$, are

$$\begin{aligned} D_\mu D^\mu\phi+\frac{\lambda}{2}(1-|\phi|^2)\phi&=0\\ \partial^\mu F_{\mu\nu}+\frac{\mathrm{i}}{2}(\phi\overline{D_\nu\phi}-\bar\phi D_\nu\phi)&=0\end{aligned} \tag{3}$$

This is a coupled set of nonlinear second-order PDEs. Of particular interest are solutions which have finite total energy. Energy is not a Lorentz-invariant quantity. To define it we must choose an inertial frame and, having broken Lorentz invariance, it is convenient to work in a temporal gauge, for which $A_0\equiv 0$ (which may be obtained by a gauge transformation with $\Lambda(t,\boldsymbol{x})=\int_0^t A_0(t',\boldsymbol{x})\,\mathrm{d}t'$, after which only time-independent gauge transformations are permitted). The potential energy of a field is then

$$\begin{aligned} E&=\frac{1}{2}\int\left(B^2+D_i\phi\overline{D_i\phi}+\frac{\lambda}{4}(1-|\phi|^2)^2\right)\mathrm{d}x^1\mathrm{d}x^2\\ &=E_{\mathrm{mag}}+E_{\mathrm{grad}}+E_{\mathrm{self}}\end{aligned} \tag{4}$$

while its kinetic energy is

$$E_{\text{kin}} = \frac{1}{2}\int \left(|\partial_0 \boldsymbol{A}|^2 + \partial_0\phi\overline{\partial_0\phi}\right) \mathrm{d}x^1\,\mathrm{d}x^2 \qquad [5]$$

If $\phi, \boldsymbol{A}$ satisfy the field equations then the total energy $E_{\text{tot}} = E_{\text{kin}} + E$ is independent of t. By Derrick's theorem, static solutions have $E_{\text{mag}} \equiv E_{\text{self}}$ (Manton and Sutcliffe 2004, pp. 82–87).

Configurations with finite energy have quantized total magnetic flux. To see this, note that E finite implies $|\phi| \to 1$ as $r \to \infty$, so $\phi \sim \mathrm{e}^{\mathrm{i}\chi(r,\theta)}$ at large r for some real (in general, multivalued) function χ. The winding number of ϕ is its winding around a circle of large radius R, that is, the integer $n = (\chi(R, 2\pi) - \chi(R, 0))/2\pi$. Although the phase of ϕ is clearly gauge dependent, n is not, because to change this, a gauge transformation $\mathrm{e}^{\mathrm{i}\Lambda}: \mathbb{R}^2 \to \mathrm{U}(1)$ would itself need nonzero winding around the circle, contradicting smoothness of $\mathrm{e}^{\mathrm{i}\Lambda}$. The model is invariant under spatial reflexions, under which $n \mapsto -n$, so we will assume (unless noted otherwise) that $n \geq 0$. Finiteness of E also implies that $D\phi = \mathrm{d}\phi - \mathrm{i}A\phi \to 0$, so $A \sim -\mathrm{i}\mathrm{d}\phi/\phi \sim \mathrm{d}\chi$ as $r \to \infty$ (note $\phi \neq 0$ for large r). Hence, the total magnetic flux is

$$\int_{\mathbb{R}^2} B\,\mathrm{d}^2\boldsymbol{x} = \lim_{R\to\infty}\oint_{S_R} A = \lim_{R\to\infty}\int_0^{2\pi} \partial_\theta\chi\,\mathrm{d}\theta = 2\pi n \qquad [6]$$

where $S_R = \{\boldsymbol{x}: |\boldsymbol{x}| = R\}$ and we have used Stokes's theorem. The above argument uses only generic properties of E, namely that finite E_{self} requires $|\phi|$ to assume a nonzero constant value as $r \to \infty$. So flux quantization is a robust feature of this type of model. As presented, the argument is somewhat formal, but it can be made mathematically rigorous at the cost of gauge-fixing technicalities (Manton and Sutcliffe 2004, pp. 164–166). Note that if $n \neq 0$ then, by continuity, $\phi(\boldsymbol{x})$ must vanish at some $\boldsymbol{x} \in \mathbb{R}^2$, and one expects a lump of energy density to be associated with each such $\boldsymbol{x}$ since $\phi = 0$ maximizes the integrand of E_{self}.

Radially Symmetric Vortices

The model supports static solutions within the radially symmetric ansatz $\phi = \sigma(r)\mathrm{e}^{\mathrm{i}n\theta}, A = a(r)\,\mathrm{d}\theta$, which reduces the field equations to a coupled pair of nonlinear ODEs:

$$\begin{aligned} \frac{\mathrm{d}^2\sigma}{\mathrm{d}r^2} + \frac{1}{r}\frac{\mathrm{d}\sigma}{\mathrm{d}r} - \frac{1}{r^2}(n-a)^2\sigma + \frac{\lambda}{2}(1-\sigma^2)\sigma &= 0 \\ \frac{\mathrm{d}^2 a}{\mathrm{d}r^2} - \frac{1}{r}\frac{\mathrm{d}a}{\mathrm{d}r} + (n-a)\sigma^2 &= 0 \end{aligned} \qquad [7]$$

Finite energy requires $\lim_{r\to\infty}\sigma(r) = 1$, $\lim_{r\to\infty} a(r) = n$ while smoothness requires $\sigma(r) \sim \text{const}_1 r^n, a(r) \sim \text{const}_2 r^2$ as $r \to 0$. It is known that solutions to this system, which we shall call n-vortices, exist for all n, λ, though no explicit formulas for them are known. They may be found numerically, and are depicted in **Figure 1**. Note that σ and a always rise monotonically to their vacuum values, and B always falls monotonically to 0, as r increases. These solutions have their magnetic flux concentrated in a single, symmetric lump, a flux tube in the $\mathbb{R}^{3+1}$ picture. In contrast, the total energy density (integrand of E in [4]) is nonmonotonic for $n \geq 2$, being peaked on a ring whose radius grows with n. This is a common feature of planar solitons.

The large r asymptotics of n-vortices are well understood. For $\lambda \leq 4$ one may linearize [7] about $\sigma = 1, a = n$, yielding

$$\sigma(r) \sim 1 + \frac{q_n}{2\pi}K_0(\sqrt{\lambda}r) \qquad [8]$$

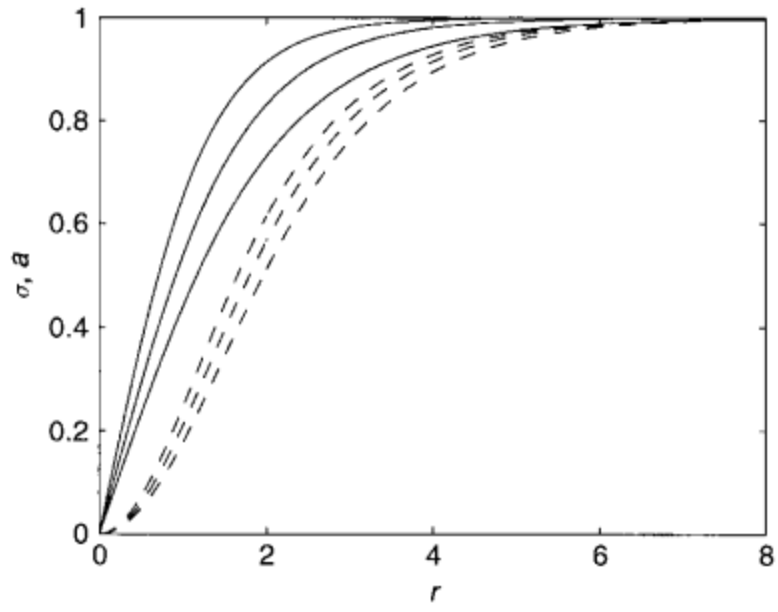
$$a(r) \sim n + \frac{m_n}{2\pi}rK_1(r) \qquad [9]$$

where q_n, m_n are unknown constants and K_α denotes the modified Bessel's function. For $\lambda > 4$ linearization is no longer well justified, and the asymptotic behaviour of σ (though not a) is quite different (Manton and Sutcliffe 2004, pp. 174–175). We shall not consider this rather extreme regime further. Note that

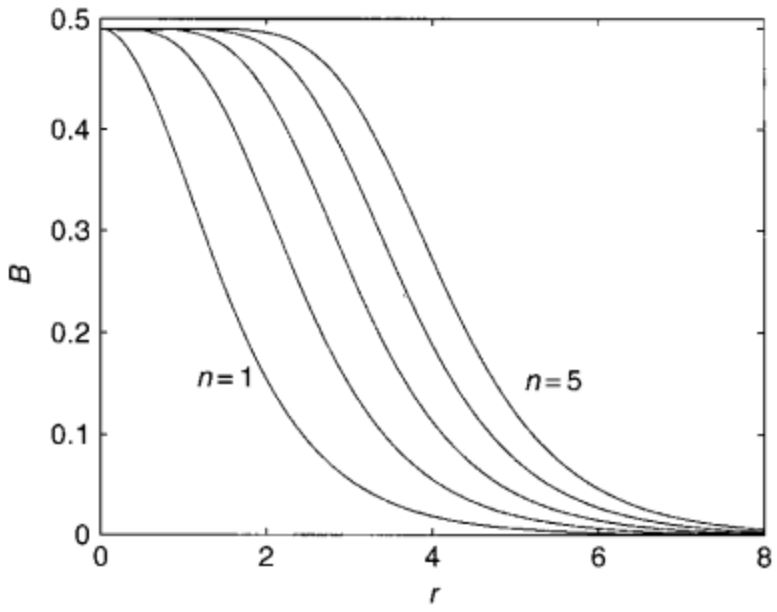
$$K_\alpha(r) \sim \sqrt{\frac{\pi}{2r}}\mathrm{e}^{-r} \quad \text{as } r \to \infty \qquad [10]$$

for all α, so both σ and a approach their vacuum values exponentially fast, but with different decay lengths: $1/\sqrt{\lambda}$ for σ, 1 for a. This can be seen in **Figure 1a**. The constants q_n and m_n depend on λ and must be inferred by comparing the numerical solutions with [8], [9]; $q = q_1$ and $m = m_1$ will receive a physical interpretation shortly.

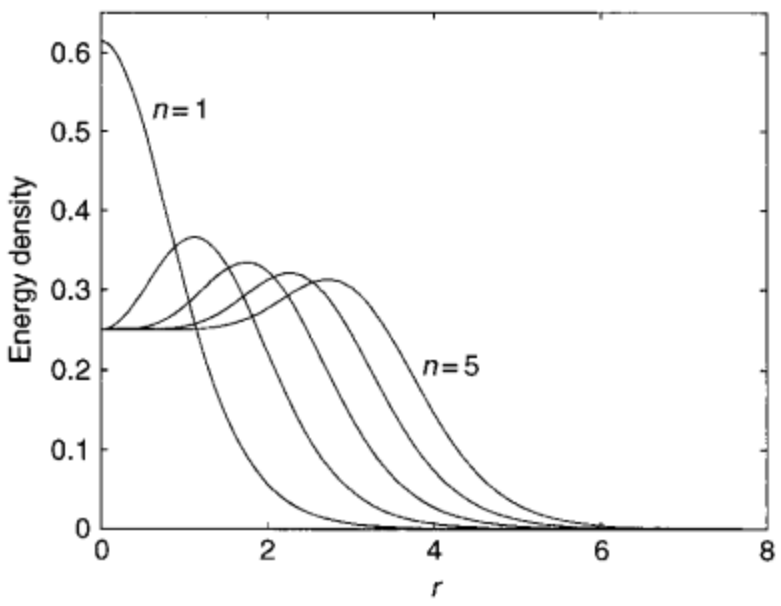
The 1-vortex (henceforth just "vortex") is stable for all λ, but n-vortices with $n \geq 2$ are unstable to break up into n separate vortices if $\lambda > 1$. We shall say that the AHM is type I if $\lambda < 1$, type II if $\lambda > 1$, and critically coupled if $\lambda = 1$, based on this distinction. Let E_n denote the energy of an n-vortex. **Figure 2** shows the energy per vortex E_n/n plotted against n for $\lambda = 0.5, 1$, and 2. It decreases with n for $\lambda = 0.5$, indicating that it is energetically favorable for isolated vortices to coalesce into higher winding lumps. For $\lambda = 2$, by contrast, E_n/n increases with n indicating that it is energetically favorable for n-vortices to fission into their constituent vortex parts. The case $\lambda = 1$ balances between these behaviors: E_n/n is independent of n. In fact, the energy of a collection of vortices is independent of their positions in this case.



(a)



(b)



(c)

Figure 1 Static, radially symmetric n-vortices: (a) the 1-vortex profile functions $\sigma(r)$ (solid curve) and $a(r)$ (dashed curve) for $\lambda = 2, 1,$ and 1/2, left to right; (b) the magnetic field B; and (c) the energy density of n-vortices, $n = 1$ to 5, left to right, for $\lambda = 1$.

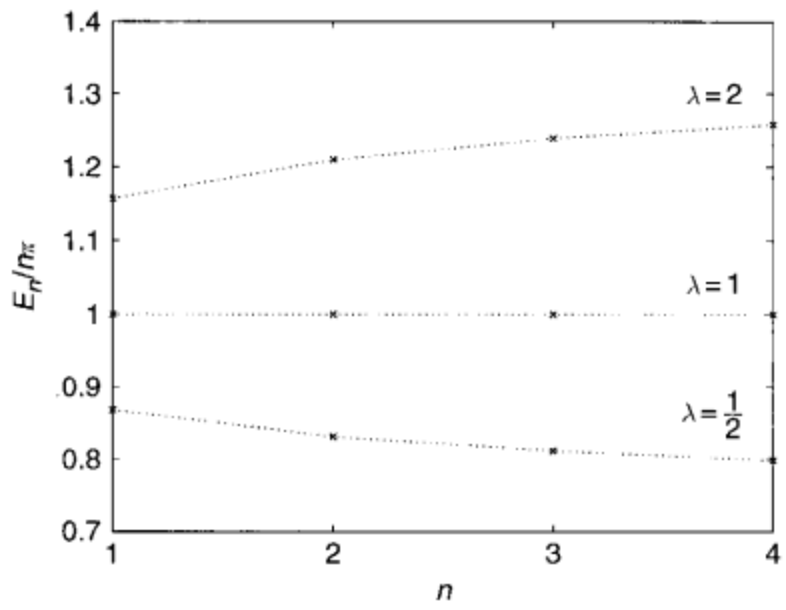


Figure 2 The energy per unit winding E_n/n of radially symmetric n-vortices for $\lambda = 1/2, 1,$ and 2.

Interaction Energy

A precise understanding of the type I/II dichotomy can be obtained using the 2-vortex interaction energy $E_{\text{int}}(s)$ introduced by Jacobs and Rebbi. This is defined to be the minimum of E over all $n = 2$ configurations for which $\phi(\boldsymbol{x}) = 0$ at some pair of points $\boldsymbol{x}_1, \boldsymbol{x}_2$ distance s apart. One interprets $\boldsymbol{x}_1, \boldsymbol{x}_2$ as the vortex positions. E_{int} can only depend on their separation $s = |\boldsymbol{x}_1 - \boldsymbol{x}_2|$, by translation and rotation invariance. **Figure 3** presents graphs of $E_{\text{int}}(s)$ generated by a lattice minimization algorithm. For $\lambda < 1$, vortices uniformly attract one another, so a vortex pair has least energy when coincident. For $\lambda > 1$, vortices uniformly repel, always lowering their energy by moving further apart. The graph for $\lambda = 1$ would be a horizontal line, $E_{\text{int}}(s) = 2\pi$.

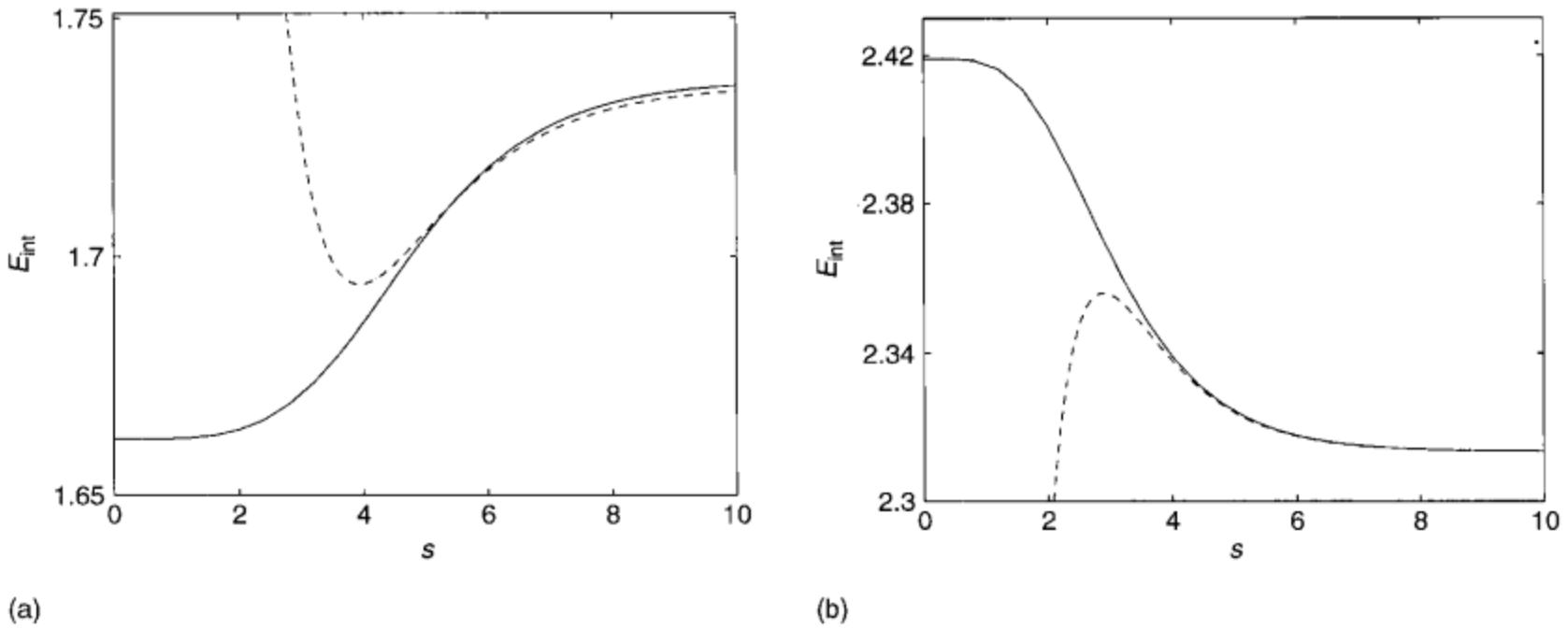


Figure 3 The 2-vortex interaction energy $E_{int}(s)$ as a function of vortex separation (solid curve), in comparison with its asymptotic form $E_{int}^{\infty}(s)$ (dashed curve) for (a) $\lambda = 1/2$ and (b) $\lambda = 2$.

The large s behavior of $E_{int}(s)$ is known, and can be understood in two ways (Manton and Sutcliffe 2004, pp. 177–181). Speight, adapting ideas of Manton on asymptotic monopole interactions, observed that, in the real ϕ gauge ($\phi \mapsto e^{-i\theta}\phi$, $A \mapsto A - d\theta$), the difference between the vortex and the vacuum $\phi = 1, A = 0$ at large r,

$$\psi = \phi - 1 \sim \frac{q}{2\pi} K_0(\sqrt{\lambda} r) \qquad [11]$$

$$(A_0, \boldsymbol{A}) \sim \frac{m}{2\pi}(0, \boldsymbol{k} \times \nabla K_0(r)) \qquad [12]$$

is identical to the solution of a linear Klein–Gordon–Proca theory,

$$(\partial_\mu \partial^\mu + \lambda)\psi = \kappa, \qquad (\partial_\mu \partial^\mu + 1)A_\nu = j_\nu \qquad [13]$$

in the presence of a composite point source,

$$\kappa = q\delta(\boldsymbol{x}), \qquad (j_0, \boldsymbol{j}) = m(0, \boldsymbol{k} \times \nabla\delta(\boldsymbol{x})) \qquad [14]$$

located at the vortex position. Viewed from afar, therefore, a vortex looks like a point particle carrying both a scalar monopole charge q and a magnetic dipole moment m, a "point vortex," inducing a real scalar field of mass $\sqrt{\lambda}$ (the Higgs particle) and a vector boson field of mass 1 (the "photon"). If physics is to be model independent, therefore, the interaction energy of a pair of well-separated vortices should approach that of the corresponding pair of point vortices as the separation grows. Computing the latter is an easy exercise in classical linear field theory, yielding

$$E_{int}(s) \sim E_{int}^{\infty}(s) = 2E_1 - \frac{q^2}{2\pi} K_0(\sqrt{\lambda} s) + \frac{m^2}{2\pi} K_0(s) \qquad [15]$$

Bettencourt and Rivers obtained the same formula by a more direct superposition ansatz approach, though they did not give the constants q, m a physical interpretation.

The force between a well-separated vortex pair, $-E_{int}'(s)$, consists of the mutual attraction of identical scalar monopoles, of range $1/\sqrt{\lambda}$, and the mutual repulsion of identical magnetic dipoles, of range 1. If $\lambda < 1$, scalar attraction dominates at large s so vortices attract. If $\lambda > 1$, magnetic repulsion dominates and they repel. If $\lambda = 1$ then $q \equiv m$, as we shall see, so the forces cancel exactly. **Figure 3** shows both E_{int} and E_{int}^{∞} for $\lambda = 0.5, 2$. The agreement is good for s large, but breaks down for $s < 4$, as one expects. Vortices are not point particles, as in the linear model, and when they lie close together the overlap of their cores produces significant effects.

The same method predicts the interaction energy between an n_1-vortex and an n_2-vortex at large separation. We just replace $2E_1$ by $E_{n_1} + E_{n_2}$, q^2 by $q_{n_1} q_{n_2}$, and m^2 by $m_{n_1} m_{n_2}$. In particular, an antivortex ((−1)-vortex) has $E_{-1} = E_1, q_{-1} = q_1 = q$, and $m_{-1} = -m_1 = -m$, so the interaction energy for a vortex–antivortex pair is

$$E_{int}^{v\bar{v}}(s) \sim 2E_1 - \frac{q^2}{2\pi} K_0(\sqrt{\lambda} r) - \frac{m^2}{2\pi} K_0(r) \qquad [16]$$

which is uniformly attractive. It would be pleasing if q_n, m_n could be deduced easily from q, m. One might guess $q_n = |n|q, m_n = nm$, in analogy with monopoles. Unfortunately, this is false: q_n, m_n grow approximately exponentially with $|n|$.

Vortex Scattering

The AHM being Lorentz invariant, one can obtain time-dependent solutions wherein a single n-vortex travels at constant velocity, with speed $0 < v < 1$ and $E_{\text{tot}} = (1 - v^2)^{-1/2} E_n$, by Lorentz boosting the static solutions described above. Of more dynamical interest are solutions in which two or more vortices undergo relative motion. The simplest problem is vortex scattering. Two vortices, initially well separated, are propelled towards one another. In the center-of-mass (COM) frame they have, as $t \to -\infty$, equal speed v, and approach one another along parallel lines distance b (the impact parameter) apart, see **Figure 4**. If $b = 0$, they approach head-on. Assuming they do not capture one another, they interact and, as $t \to \infty$, recede along parallel straight lines having been deflected through an angle Θ (the scattering angle). If scattering is elastic, the exit lines also lie b apart and each vortex travels at speed v as $t \to \infty$. The dependence of Θ on v, b, and λ has been studied through lattice simulations by several authors, perhaps most comprehensively by Myers, Rebbi, and Strilka (1992). We shall now describe their results.

Note first that vortex scattering is actually inelastic: vortices recede with speed $< v$ because some of their initial kinetic energy is dispersed by the collision as small-amplitude traveling waves ("radiation"). This energy loss can be as high as 80% in very fast collisions at small b. At small v the energy loss is tiny, but can still have important consequences for type I vortices: if v is very small, they start with only just enough energy to escape their mutual attraction. In undergoing a small b collision they can lose enough of this energy to become trapped in an oscillating bound state. In this case they do not truly scatter and Θ is ill-defined. Myers *et al.* find that $v \geq 0.2$ suffices to avoid capture when $\lambda = 1/2$. Since type I vortices attract, one might expect Θ to be always negative, indicating that the vortices deflect towards one another. In fact, as **Figure 5a** shows, this happens only for small v and large b. Another naive expectation is that $\Theta = 0$ or $\Theta = 180°$ when $b = 0$ (either vortices pass through one another or ricochet backwards in a head-on collision). In fact $\Theta = 90°$, the only other possibility allowed by reflexion symmetry of the initial data. **Figure 6** depicts snapshots of such a scattering process at modest v. The vortices deform each other as they get close until, at the moment of coincidence, they are close to the static 2-vortex ring. They then break apart along a line perpendicular to their line of approach. One may consider them to have exchanged half-vortices, so that each emergent vortex is a mixture of the incoming vortices. This rather surprising phenomenon was actually predicted by Ruback in advance of any numerical simulations and turns out to be a generic feature of planar topological solitons.

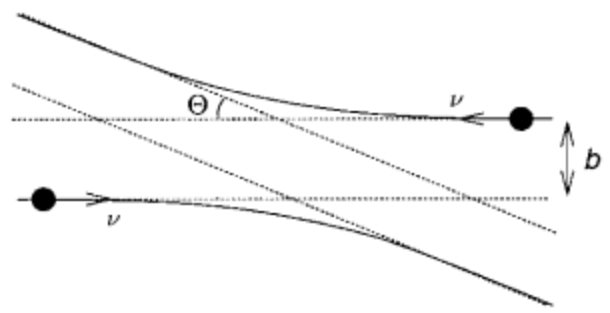


Figure 4 The geometry of vortex scattering.

Consider now the type II case ($\lambda = 2$, **Figure 5b**). Here, $\Theta > 0$ for all v, b as one expects of particles that repel each other. Head-on scattering is more interesting now since two regimes emerge: for $v > v_{\text{crit}} \approx 0.3$, one has the surprising 90° scattering already described, while for $v < v_{\text{crit}}$ the vortices bounce backwards, $\Theta = 180°$. This is easily explained. In order to undergo 90° head-on scattering, the vortices must become coincident (otherwise reflexion symmetry is violated), hence must have initial energy at least E_2. For $v < v_{\text{crit}}$, where

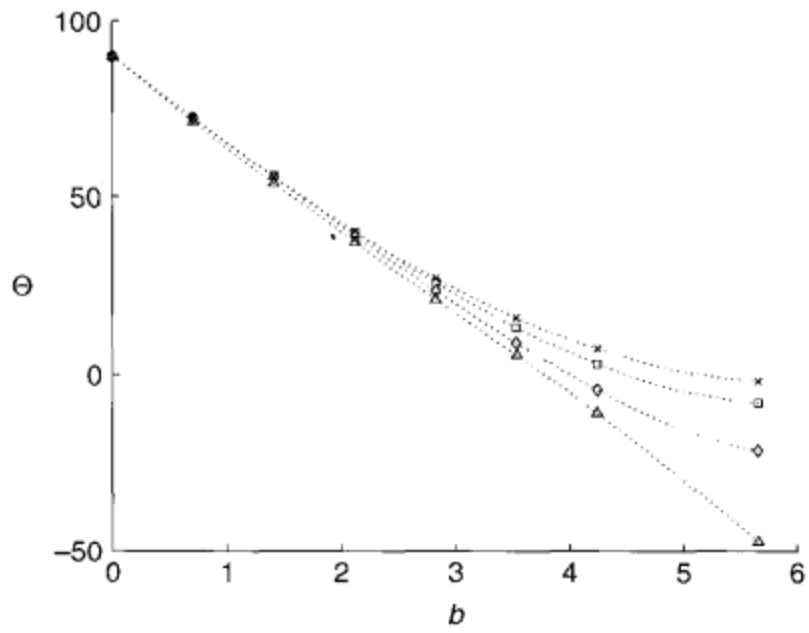
$$\frac{2E_1}{\sqrt{1 - v_{\text{crit}}^2}} = E_2 \qquad [17]$$

they have too little energy, so come to a halt before coincidence, then recede from one another. The solution v_{crit} of [17] depends on λ and is plotted in **Figure 7**. For v slightly above v_{crit}, we see that, in contrast to the type I case, $\Theta(b)$ is not monotonic: maximum deflection occurs at nonzero b.

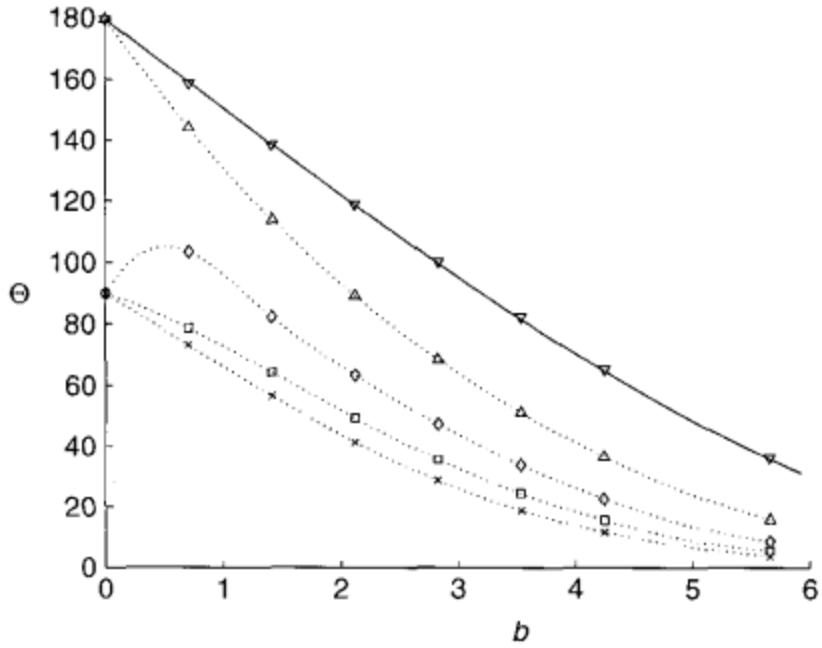
The point vortex formalism yields a simple model of type II vortex scattering which is remarkably successful at small v. One writes down the Lagrangian for two identical (nonrelativistic) point particles of mass E_1 moving along trajectories $\boldsymbol{x}_1(t), \boldsymbol{x}_2(t)$ under the influence of the repulsive potential E_{int}^{∞},

$$L = \tfrac{1}{2} E_1 (|\dot{\boldsymbol{x}}_1|^2 + |\dot{\boldsymbol{x}}_2|^2) - E_{\text{int}}^{\infty}(|\boldsymbol{x}_1 - \boldsymbol{x}_2|) \qquad [18]$$

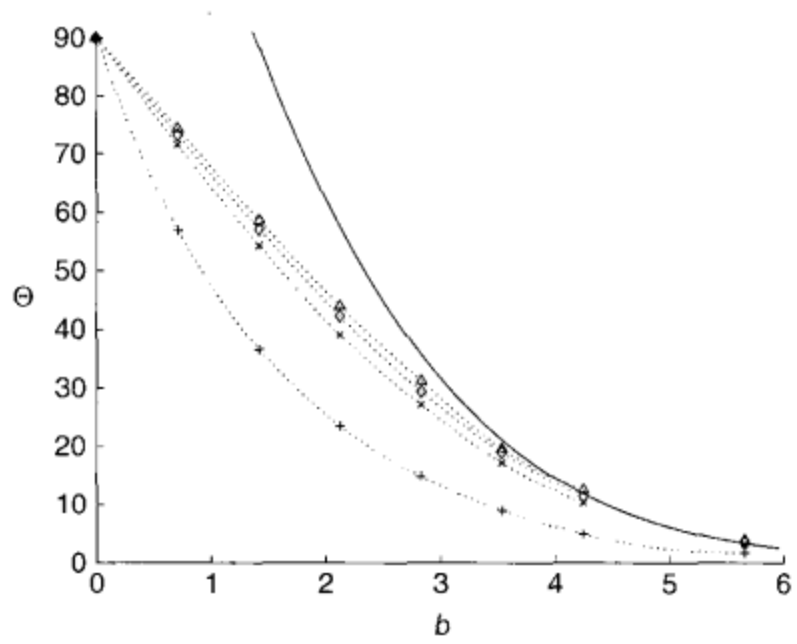
Energy and angular momentum conservation reduce $\Theta(v, b)$ to an integral over one variable ($s = |\boldsymbol{x}_1 - \boldsymbol{x}_2|$) which is easily computed numerically. To illustrate, **Figure 5b** shows the result for $\lambda = 2, v = 0.1$ in comparison with the lattice simulations of



(a)



(b)



(c)

Figure 5 The 2-vortex scattering angle Θ as a function of impact parameter b for $v=0.1$ ($\triangledown$), $v=0.2$ ($\triangle$), $v=0.3$ ($\diamond$), $v=0.4$ ($\approx$), $v=0.5$ ($\times$), and $v=0.9$ ($+$), as computed by Myers *et al.* (1992): (a) $\lambda=1/2$; (b) $\lambda=2$; (c) $\lambda=1$. The dotted curves are merely guides to the eye. The solid curves in (b), (c) were computed using the point vortex model. Note that Myers *et al.* use different normalizations, so $b=\sqrt{2}b_{MRS}$ and $\lambda=\lambda_{MRS}/2$.

Myers *et al.* The agreement is almost perfect. For large v the approximation breaks down not only because relativistic corrections become significant, but also because small b collisions then probe the small $|x_1-x_2|$ region where vortex core overlap effects become important. For the same reason, the point vortex model is less useful for type I scattering. Here there is no repulsion to keep the vortices well separated, so its validity is restricted to the small v, large b regime.

Critical coupling is theoretically the most interesting regime, where most analytic progress has been made. Since $E_{\text{int}} \equiv E_{\text{int}}^{\infty} \equiv 0$, one might expect vortex scattering to be trivial ($\Theta(v,b) \equiv 0$), but this is quite wrong, as shown in **Figure 5c**. In particular, $\Theta(v,0)=90°$ for all v, just as in the large v type I and type II cases. The point is that scalar attraction and magnetic repulsion of vortices are mediated by fields with different Lorentz transformation properties. While they cancel for static vortices, there is no reason to expect them to cancel for vortices in relative motion.

Critical Coupling

The AHM with $\lambda=1$ has many remarkable properties, at which we have so far only hinted. These all stem from Bogomol'nyi's crucial observation (Manton and Sutcliffe 2004, pp. 197–202) that the potential energy in this case can be rewritten as

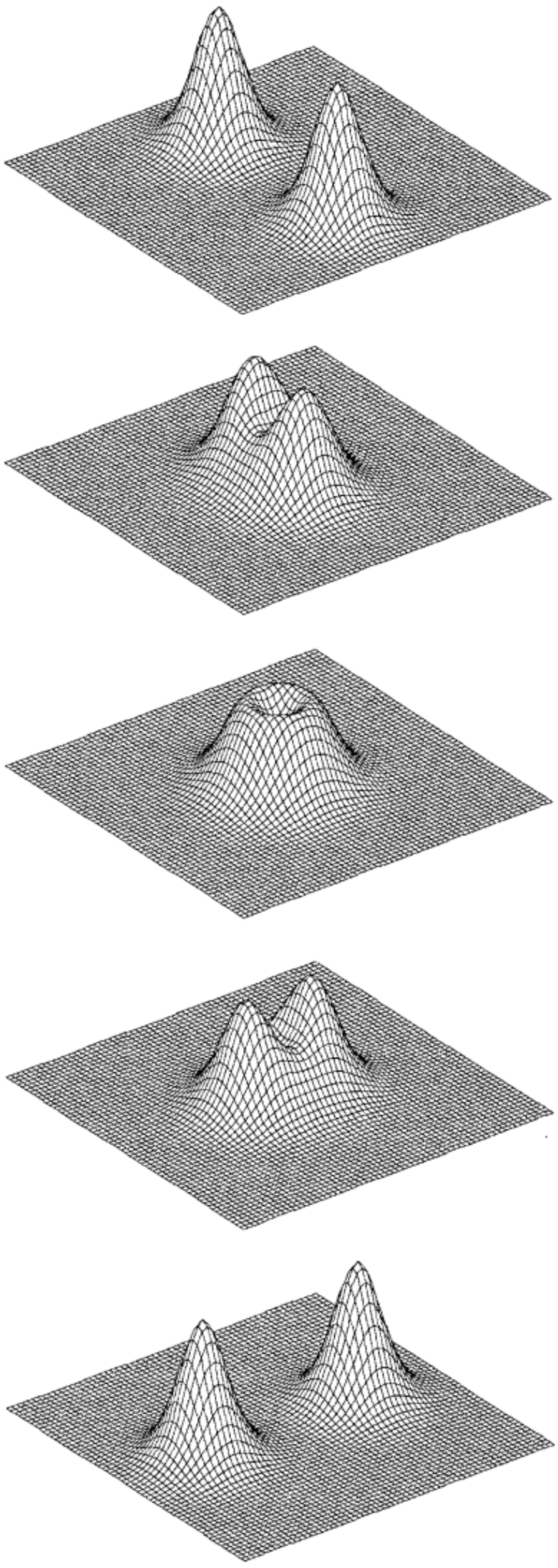


Figure 6 Snapshots of the energy density during a head-on collision of vortices. This 90° scattering phenomenon is a generic feature of planar topological soliton dynamics.

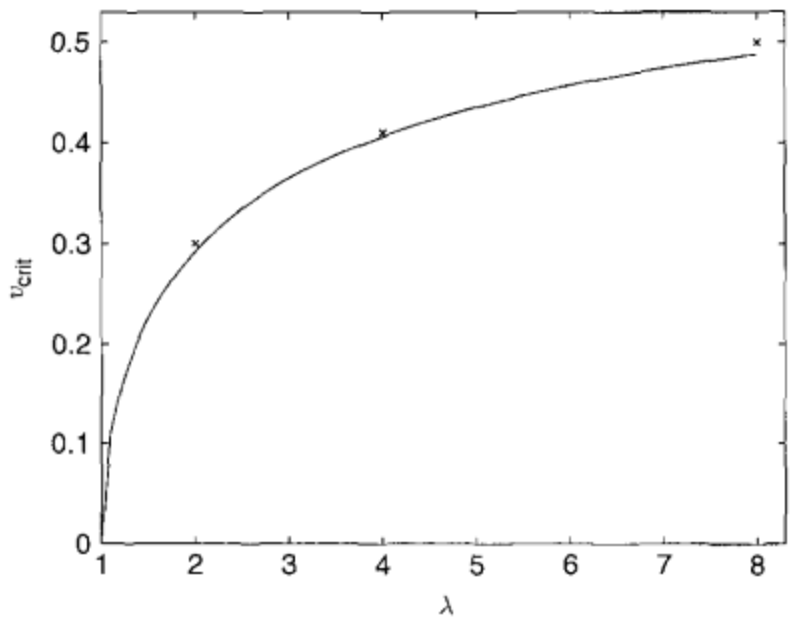


Figure 7 The critical velocity for 90° head-on scattering of type II vortices v_{crit} as a function of λ, as predicted by equation [17] (solid curve), in comparison with the results of Myers *et al.* (1992), (crosses).

$$E = \frac{1}{2}\int \left\{ \left(B - \frac{1}{2}(1 - |\phi|^2) \right)^2 + |D_1\phi + iD_2\phi|^2 + B \right\} d^2\boldsymbol{x} - i\int_{\mathbb{R}^2} d(\bar{\phi}D\phi) \qquad [19]$$

The last integral vanishes by Stokes's theorem, so $E \geq \pi n$ by flux quantization [6], and $E = \pi n$ if and only if

$$(D_1 + iD_2)\phi = 0 \qquad [20]$$

$$\tfrac{1}{2}(1 - |\phi|^2) = B \qquad [21]$$

Note that system [20], [21] is first order, in contrast to the second-order field equations [3]. No explicit solutions of [20], [21] are known. However, Taubes has proved that for each unordered list $[z_1, z_2, \ldots, z_n]$ of n points in $\mathbb{C}$, not necessarily distinct, there exists a solution of [20], [21], unique up to gauge transformations, with $\phi(z_1) = \phi(z_2) = \cdots = \phi(z_n) = 0$ and ϕ nonvanishing elsewhere, the zero at z_r having the same multiplicity as z_r has in the list. Note that the list is unordered: a solution is uniquely determined by the positions and multiplicities of the zeroes of ϕ, but the order in which we label these is irrelevant. The solution minimizes E within the class $\mathcal{C}_n$ of winding n configurations, so is automatically a stable static solution of the model.

Equation [20] applied to the symmetric n-vortex, $\phi = \sigma(r)\, e^{in\theta}$, $A = a(r)\, d\theta$ implies $a(r) = n - r\sigma'(r)/\sigma(r)$. Comparing with [8], [9], it follows that $q_n = m_n$ when $\lambda = 1$ as previously claimed, since $K_1 = -K_0'$. Tong has conjectured, based on a string duality argument, that $q_1 = -2\pi 8^{1/4}$. This is consistent with current numerics but has no direct derivation so far.

Taubes's theorem shows that this n-vortex is just one point, corresponding to the list $[0, 0, \ldots, 0]$, in a $2n$-dimensional space of static multivortex solutions called the moduli space M_n. This space may be visualized as the flat, finite-dimensional valley bottom in $\mathcal{C}_n$ on which E attains its minimum value, πn. Points in M_n are in one-to-one correspondence with distinct unordered lists $[z_1, z_2, \ldots, z_n]$, which are themselves in one-to-one correspondence with points in $\mathbb{C}^n$, as follows. To each list, we assign the unique monic polynomial whose roots are z_r,

$$\begin{aligned} p(z) &= (z - z_1)(z - z_2)\cdots(z - z_n) \\ &= a_0 + a_1 z + \cdots + a_{n-1} z^{n-1} + z^n \end{aligned} \qquad [22]$$

This polynomial is uniquely determined by its coefficients $(a_0, a_1, \ldots, a_{n-1}) \in \mathbb{C}^n$, which give good global coordinates on $\mathsf{M}_n \cong \mathbb{C}^n$. The zeros z_r of ϕ may be used as local coordinates on M_n, away from Δ, the subset of M_n on which two or more of the zeros z_r coincide, but are not good global coordinates.

Let $(\phi, A)_a$ denote the static solution corresponding to $a \in \mathbb{C}^n$. If the zeros z_r are all at least s apart, Taubes showed the solution is just a linear superposition of 1-vortices located at z_r, up to corrections exponentially small in s. Imagine these constituent vortices are pushed with small initial velocities. Then $(\phi(t), A(t))$ must remain close to the valley bottom M_n, since departing from it costs kinetic energy, of which there is little. Manton has suggested, therefore, that the dynamics is well approximated by the constrained variational problem wherein $(\phi(t), A(t)) = (\phi, A)_{a(t)} \in \mathsf{M}_n$ for all t. Since the action $S = \int \mathcal{L}\, \mathrm{d}^3x = \int (E_{\mathrm{kin}} - E)\, \mathrm{d}t$, and $E = \pi n$, constant, on M_n, this constrained problem amounts to Lagrangian mechanics on configuration space M_n with Lagrangian $L = E_{\mathrm{kin}}|_{\mathsf{M}_n}$. Now E_{kin} is real, positive, and quadratic in time derivatives of ϕ, $\boldsymbol{A}$, so

$$L = \tfrac{1}{2}\gamma_{rs}(a)\dot{a}_r\dot{\bar{a}}_s \qquad [23]$$

γ_{rs} forming the entries of a positive-definite $n \times n$ Hermitian matrix $(\gamma_{sr} \equiv \overline{\gamma_{rs}})$. Since $(\phi, A)_a$ is not known explicitly, neither are $\gamma_{rs}(a)$. Observe, however, that L is the Lagrangian for geodesic motion in M_n with respect to the Riemannian metric

$$\gamma = \gamma_{rs}(a)\mathrm{d}a_r\mathrm{d}\bar{a}_s \qquad [24]$$

Manton originally proposed this geodesic approximation for monopoles, but it is now standard for all topological solitons of Bogomol'nyi type (where one has a moduli space of static multisolitons saturating a topological lower bound on E). Note that geodesics are independent of initial speed, which agrees with Myers *et al*: **Figure 5c** shows that $\Theta(v, b)$ is approximately independent of v for $v \leq 0.5$. Further, Stuart (1994) has proved that, for initial speeds of order ϵ, small, the fields stay (pointwise) ϵ^2 close to their geodesic approximant for times of order ϵ^{-1}.

On symmetry grounds, two vortex dynamics in the COM frame reduces to geodesic motion in $\mathsf{M}_2^0 \cong \mathbb{C}$, the subspace of centered 2-vortices ($a_1 = 0$, so $z_1 = -z_2$), with induced metric

$$\gamma^0 = G(|a_0|)\mathrm{d}a_0\mathrm{d}\bar{a}_0 \qquad [25]$$

G being some positive function. Note that $a_0 = z_1 z_2$, so the intervortex distance $|z_1 - z_2| = 2|z_1| = 2|a_0|^{1/2}$. The line $a_0 = \beta \in \mathbb{R}$, traversed with β increasing, say, is geodesic in M_2^0. The vortex positions (roots of $z^2 + a_0$) are $\pm\sqrt{|\beta|}$ for $\beta \leq 0$ and $\pm\mathrm{i}\sqrt{\beta}$ for $\beta > 0$. This describes perfectly the 90° scattering phenomenon: two vortices approach head-on along the x^1 axis, coincide to form a 2-vortex ring, then break apart along the x^2 axis, as in **Figure 6**. This behavior occurs because $a_0 = z_1 z_2$, rather than $z_1 - z_2$, is the correct global coordinate on M_2^0, since vortices are classically indistinguishable.

Samols found a useful formula (Manton and Sutcliffe 2004, pp. 205–215) for γ in terms of the behavior of $|\phi_a|$ close to its zeros, using which he devised an efficient numerical scheme to evaluate $G(|a_0|)$, and computed $\Theta(b)$ in detail, finding excellent agreement with lattice simulations at low speeds. He also studied the quantum scattering of vortices, approximating the quantum state by a wave function Ψ on M_n evolving according to the natural Schrödinger equation for quantum geodesic motion,

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = -\tfrac{1}{2}\hbar^2\Delta_\gamma\Psi \qquad [26]$$

where Δ_γ is the Laplace–Beltrami operator on (M_n, γ). This technique, introduced for monopoles by Gibbons and Manton, is now standard for solitons of Bogomol'nyi type.

By analyzing the forces between moving point vortices at $\lambda = 1$, Manton and Speight (2003) showed that, as the vortex separations become uniformly large, the metric on M_n approaches

$$\begin{aligned} \gamma^\infty = \pi\sum_r \Bigg[&\mathrm{d}z_r\mathrm{d}\bar{z}_r - \frac{q^2}{4\pi}\sum_{s\neq r} K_0(|z_r - z_s|) \\ &\times (\mathrm{d}z_r - \mathrm{d}z_s)(\mathrm{d}\bar{z}_r - \mathrm{d}\bar{z}_s)\Bigg] \end{aligned} \qquad [27]$$

This formula can also be obtained by a method of matched asymptotic expansions. We can use [27] to study 2-vortex scattering for large b, when the

vortices remain well separated. (Note that γ^{∞} is not positive definite if any $|z_r - z_s|$ becomes too small.) The results are good, provided $v \leq 0.5$ and $b \geq 3$ (see **Figure** 5c).

Other Developments

The (critically coupled) AHM on a compact physical space Σ is of considerable theoretical and physical interest. Bradlow showed that $\mathsf{M}_n(\Sigma)$ is empty unless $V = \text{Area}(\Sigma) \geq 4\pi n$, so there is a limit to how many vortices a space of finite area can accommodate (Manton and Sutcliffe 2004, pp. 227–230). Manton has analyzed the thermodynamics of a gas of vortices by studying the statistical mechanics of geodesic flow on $\mathsf{M}_n(\Sigma)$. In this context, spatial compactness is a technical device to allow nonzero vortex density n/V for finite n, without confining the fields to a finite box, which would destroy the Bogomol'nyi properties. In the limit of interest, $n, V \to \infty$ with n/V fixed, the thermodynamical properties turn out to depend on Σ only through V, so $\Sigma = S^2$ and $\Sigma = T^2$ give equivalent results, for example. The equation of state of the gas is (P = pressure, T = temperature)

$$P = \frac{nT}{V - 4\pi n} \qquad [28]$$

which is similar, at low density n/V, to that of a gas of hard disks of area 2π. The crucial step in deriving [28] is to find the volume of $\mathsf{M}_n(\Sigma)$ which, despite there being no formula for γ, may be computed exactly by remarkable indirect arguments (Manton and Sutcliffe 2004, pp. 231–234).

The static AHM coincides with the Ginzburg–Landau model of superconductivity, which has precisely the same type I/II classification. Here the "Higgs" field represents the wave function of a condensate of Cooper pairs, usually (but not always) electrons. There has been a parallel development of the static model by condensed matter theorists, therefore; see Fossheim and Sudbo (2004), for example. In fact the vortex was actually first discovered by Abrikosov in the condensed matter context. One important difference is that type I superconductors do not support vortex solutions in an external magnetic field B_{ext} because the critical $|B_{\text{ext}}|$ required to create a single vortex is greater than the critical $|B_{\text{ext}}|$ required to destroy the condensate completely ($\phi \equiv 0$). Type II superconductors do support vortices, and there are such superconductors with $\lambda \approx 1$, but the vortex dynamics we have described is not relevant to these systems. In this context there is an obvious preferred reference frame (the rest frame of the superconductor) so it is unsurprising that the Lorentz-invariant AHM is inappropriate. Insofar as vortices move at all, they seem to obey a first-order (in time) dynamical system, in contrast to the second-order AHM. Manton has devised a first-order system which may have relevance to superconductivity, by replacing E_{kin} with a Chern–Simons–Schrödinger functional (Manton and Sutcliffe 2004, pp. 193–197). Rather than attracting or repelling, vortices now tend to orbit one another at constant separation. There is again a moduli space approximation to slow vortex dynamics for $\lambda \approx 1$, but it has a Hamiltonian-mechanical rather than Riemannian-geometric flavor.

Finally, an interesting simplification of the AHM, which arises, for example, as a phenomenological model of liquid helium-4, is obtained if we discard the gauge field A_μ, or equivalently set the electric charge of ϕ to $e = 0$. There is now no type I/II classification, since λ may be absorbed by rescaling. The resulting model, which has only global U(1) phase symmetry, supports n-vortices $\phi = \sigma(r)e^{in\theta}$ for all n, but these are not exponentially spatially localized,

$$\sigma(r) = 1 - \frac{n^2}{\lambda r^2} - \frac{n^2(8 + n^2)}{2\lambda^2 r^4} + O(r^{-6}) \qquad [29]$$

and cannot have finite E by Derrick's theorem. They are unstable for $|n| > 1$, and 1-vortices uniformly repel one another. They can be given an interesting first-order dynamics (the Gross–Pitaevski equation).

Abbreviations

A_μ	electromagnetic gauge potential
b	impact parameter
D_μ	gauge-covariant derivative
E	potential energy
E_{kin}	kinetic energy
$F_{\mu\nu}$	electromagnetic field strength tensor
L	Lagrangian
$\mathcal{L}$	Lagrangian density
S	action
ϕ	Higgs field
Θ	scattering angle

See also: Fractional Quantum Hall Effect; Ginzburg–Landau Equation; High T_c Superconductor Theory; Integrable Systems: Overview; Nonperturbative and Topological Aspects of Gauge Theory; Quantum Fields with Topological Defects; Solitons and Other Extended Field Configurations; Symmetry Breaking in Field Theory; Topological Defects and Their Homotopy Classification; Variational Techniques for Ginzburg–Landau Energies.

Further Reading

Atiyah M and Hitchin N (1988) *The Geometry and Dynamics of Magnetic Monopoles*. Princeton: Princeton University Press.

Fossheim K and Sudbo A (2004) *Superconductivity: Physics and Applications*. Hoboken NJ: Wiley.

Jaffe A and Taubes C (1980) *Vortices and Monopoles: Structure of Static Gauge Theories*. Boston: Birkhäuser.

Nakahara M (1990) *Geometry, Topology and Physics*. Bristol: Adam-Hilger.

Manton NS and Speight JM (2003) Asymptotic interactions of critically coupled vortices. *Communications in Mathematical Physics* 236: 535–555.

Manton NS and Sutcliffe PM (2004) *Topological Solitons*. Cambridge: Cambridge University Press.

Myers E, Rebbi C, and Strilka R (1992) Study of the interaction and scattering of vortices in the abelian Higgs (or Ginzburg-Landau) model. *Physical Review* 45: 1355–1364.

Rajaraman R (1989) *Solitons and Instantons*. Amsterdam: North-Holland.

Stuart D (1994) Dynamics of abelian Higgs vortices in the near Bogomolny regime. *Communications in Mathematical Physics* 159: 51–91.

Vilenkin A and Shellard EPS (1994) *Cosmic Strings and Other Topological Defects*. Cambridge: Cambridge University Press.

AdS/CFT Correspondence

C P Herzog, University of California at Santa Barbara, Santa Barbara, CA, USA
I R Klebanov, Princeton University, Princeton, NJ, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The anti-de Sitter/conformal field theory (AdS/CFT) correspondence is a conjectured equivalence between a quantum field theory in d spacetime dimensions with conformal scaling symmetry and a quantum theory of gravity in $(d+1)$-dimensional anti-de Sitter space. The most promising approaches to quantizing gravity involve superstring theories, which are most easily defined in 10 spacetime dimensions, or M-theory which is defined in 11 spacetime dimensions. Hence, the AdS/CFT correspondences based on superstrings typically involve backgrounds of the form $AdS_{d+1} \times Y_{9-d}$ while those based on M-theory involve backgrounds of the form $AdS_{d+1} \times Y_{10-d}$, where Y are compact spaces.

The examples of the AdS/CFT correspondence discussed in this article are dualities between (super)conformal nonabelian gauge theories and superstrings on $AdS_5 \times Y_5$, where Y_5 is a five-dimensional Einstein space (i.e., a space whose Ricci tensor is proportional to the metric, $R_{ij} = 4g_{ij}$). In particular, the most basic (and maximally supersymmetric) such duality relates $\mathcal{N} = 4$ SU(N) super Yang–Mills (SYM) and type IIB superstring in the curved background $AdS_5 \times S^5$.

There exist special limits where this duality is more tractable than in the general case. If we take the large-N limit while keeping the 't Hooft coupling $\lambda = g_{YM}^2 N$ fixed (g_{YM} is the Yang–Mills coupling strength), then each Feynman graph of the gauge theory carries a topological factor N^χ, where χ is the Euler characteristic of the graph. The graphs of spherical topology (often called "planar"), to be identified with string tree diagrams, are weighted by N^2; the graphs of toroidal topology, to be identified with string one-loop diagrams, by N^0, etc. This counting corresponds to the closed-string coupling constant of order N^{-1}. Thus, in the large-N limit the gauge theory becomes "planar," and the dual string theory becomes classical. For small $g_{YM}^2 N$, the gauge theory can be studied perturbatively; in this regime the dual string theory has not been very useful because the background becomes highly curved. The real power of the AdS/CFT duality, which already has made it a very useful tool, lies in the fact that, when the gauge theory becomes strongly coupled, the curvature in the dual description becomes small; therefore, classical supergravity provides a systematic starting point for approximating the string theory.

There is a strong motivation for an improved understanding of dualities of this type. In one direction, generalizations of this duality provide the tantalizing hope of a better understanding of quantum chromodynamics (QCD); QCD is a nonabelian gauge theory that describes the strong interactions of mesons, baryons, and glueballs, and has a conformal symmetry which is broken by quantum effects. In the other direction, AdS/CFT suggests that quantum gravity may be understandable as a gauge theory. Understanding the confinement of quarks and gluons that takes place in low-energy QCD and quantizing gravity are well acknowledged to be two of the most important outstanding problems of theoretical physics.

Some Geometrical Preliminaries

The d-dimensional sphere of radius L, S^d, may be defined by a constraint

$$\sum_{i=1}^{d+1} (X^i)^2 = L^2 \qquad [1]$$

on $d+1$ real coordinates X^i. It is a positively curved maximally symmetric space with symmetry group SO($d+1$). We will denote the round metric on S^d of unit radius by $d\Omega_d^2$.

Notes:

The d-dimensional anti-de Sitter space, AdS_d, may be defined by a constraint

$$(X^0)^2 + (X^d)^2 - \sum_{i=1}^{d-1}(X^i)^2 = L^2 \qquad [2]$$

This constraint shows that the symmetry group of AdS_d is $SO(2, d-1)$. AdS_d is a negatively curved maximally symmetric space, that is, its curvature tensor is related to the metric by

$$R_{abcd} = -\frac{1}{L^2}[g_{ac}g_{bd} - g_{ad}g_{bc}] \qquad [3]$$

Its metric may be written as

$$ds^2_{AdS} = L^2\left(-(y^2+1)dt^2 + \frac{dy^2}{y^2+1} + y^2\, d\Omega^2_{d-2}\right) \qquad [4]$$

where the radial coordinate $y \in [0, \infty)$, and t is defined on a circle of length 2π. This space has closed timelike curves; to eliminate them, we will work with the universal covering space where $t \in (-\infty, \infty)$. The boundary of AdS_d, which plays an important role in the AdS/CFT correspondence, is located at infinite y. There exists a subspace of AdS_d called the Poincaré wedge, with the metric

$$ds^2 = \frac{L^2}{z^2}\left(dz^2 - (dx^0)^2 + \sum_{i=1}^{d-2}(dx^i)^2\right) \qquad [5]$$

where $z \in [0, \infty)$.

A Euclidean continuation of AdS_d is the Lobachevsky space (hyperboloid), L_d. It is obtained by reversing the sign of $(X^d)^2$, dt^2, and $(dx^0)^2$ in [2], [4], and [5], respectively. After this Euclidean continuation, the metrics [4] and [5] become equivalent; both of them cover the entire L_d. Another equivalent way of writing the metric is

$$ds^2_L = L^2\left(d\rho^2 + \sinh^2\rho\, d\Omega^2_{d-1}\right) \qquad [6]$$

which shows that the boundary at infinite ρ has the topology of S^{d-1}. In terms of the Euclideanized metric [5], the boundary consists of the R^{d-1} at $z=0$, and a single point at $z=\infty$.

The Geometry of Dirichlet Branes

Our path toward formulating the AdS_5/CFT_4 correspondence requires introduction of Dirichlet branes, or D-branes for short. They are soliton-like "membranes" of various internal dimensionalities contained in type II superstring theories. A Dirichlet p-brane (or Dp brane) is a $(p+1)$-dimensional hyperplane in $(9+1)$-dimensional spacetime where strings are allowed to end. A D-brane is much like a topological defect: upon touching a D-brane, a closed string can open up and turn into an open string whose ends are free to move along the D-brane. For the endpoints of such a string the $p+1$ longitudinal coordinates satisfy the conventional free (Neumann) boundary conditions, while the $9-p$ coordinates transverse to the Dp brane have the fixed (Dirichlet) boundary conditions, hence the origin of the term "Dirichlet brane." The Dp brane preserves half of the bulk supersymmetries and carries an elementary unit of charge with respect to the $(p+1)$-form gauge potential from the Ramond–Ramond (RR) sector of type II superstring.

For this article, the most important property of D-branes is that they realize gauge theories on their world volume. The massless spectrum of open strings living on a Dp brane is that of a maximally supersymmetric U(1) gauge theory in $p+1$ dimensions. The $9-p$ massless scalar fields present in this supermultiplet are the expected Goldstone modes associated with the transverse oscillations of the Dp brane, while the photons and fermions provide the unique supersymmetric completion. If we consider N parallel D-branes, then there are N^2 different species of open strings because they can begin and end on any of the D-branes. N^2 is the dimension of the adjoint representation of U(N), and indeed we find the maximally supersymmetric U(N) gauge theory in this setting.

The relative separations of the Dp branes in the $9-p$ transverse dimensions are determined by the expectation values of the scalar fields. We will be interested in the case where all scalar expectation values vanish, so that the N Dp branes are stacked on top of each other. If N is large, then this stack is a heavy object embedded into a theory of closed strings which contains gravity. Naturally, this macroscopic object will curve space: it may be described by some classical metric and other background fields including the RR $(p+2)$-form field strength. Thus, we have two very different descriptions of the stack of Dp branes: one in terms of the U(N) supersymmetric gauge theory on its world volume, and the other in terms of the classical RR charged p-brane background of the type II closed superstring theory. The relation between these two descriptions is at the heart of the connections between gauge fields and strings that are the subject of this article.

Coincident D3 Branes

Gauge theories in $3+1$ dimensions play an important role in physics, and as explained above, parallel D3 branes realize a $(3+1)$-dimensional U(N) SYM

theory. Let us compare a stack of D3 branes with the RR-charged black 3-brane classical solution where the metric assumes the form

$$ds^2 = H^{-1/2}(r)\left[-f(r)(dx^0)^2 + (dx^i)^2\right] + H^{1/2}(r)\left[f^{-1}(r)dr^2 + r^2\, d\Omega_5{}^2\right] \qquad [7]$$

where $i = 1, 2, 3$ and

$$H(r) = 1 + \frac{L^4}{r^4}, \qquad f(r) = 1 - \frac{r_0{}^4}{r^4}$$

The solution also contains an RR self-dual 5-form field strength

$$F = dx^0 \wedge dx^1 \wedge dx^2 \wedge dx^3 \wedge d(H^{-1}) + 4L^4\, \mathrm{vol}(S^5) \qquad [8]$$

so that the Einstein equation of type IIB supergravity, $R_{\mu\nu} = F_{\mu\alpha\beta\gamma\delta}F_{\nu}{}^{\alpha\beta\gamma\delta}/96$, is satisfied.

In the extremal limit $r_0 \to 0$, the 3-brane metric becomes

$$ds^2 = \left(1 + \frac{L^4}{r^4}\right)^{-1/2}\left(-(dx^0)^2 + (dx^i)^2\right) + \left(1 + \frac{L^4}{r^4}\right)^{1/2}(dr^2 + r^2\, d\Omega_5^2) \qquad [9]$$

Just like the stack of parallel, ground-state D3 branes, the extremal solution preserves 16 of the 32 supersymmetries present in the type IIB theory. Introducing $z = L^2/r$, one notes that the limiting form of [9] as $r \to 0$ factorizes into the direct product of two smooth spaces, the Poincaré wedge [5] of AdS_5, and S^5, with equal radii of curvature L. The 3-brane geometry may thus be viewed as a semi-infinite throat of radius L which, for $r \gg L$, opens up into flat $(9+1)$-dimensional space. Thus, for L much larger than the string length scale, $\sqrt{\alpha'}$, the entire 3-brane geometry has small curvatures everywhere and is appropriately described by the supergravity approximation to type IIB string theory.

The relation between L and $\sqrt{\alpha'}$ may be found by equating the gravitational tension of the extremal 3-brane classical solution to N times the tension of a single D3 brane:

$$\frac{2}{\kappa^2}L^4\, \mathrm{vol}(S^5) = N\frac{\sqrt{\pi}}{\kappa} \qquad [10]$$

where $\mathrm{vol}(S^5) = \pi^3$ is the volume of a unit 5-sphere, and $\kappa = \sqrt{8\pi G}$ is the ten-dimensional gravitational constant. It follows that

$$L^4 = \frac{\kappa}{2\pi^{5/2}}N = g_{\mathrm{YM}}^2\, N\alpha'^2 \qquad [11]$$

where we used the standard relations $\kappa = 8\pi^{7/2}g_{st}\alpha'^2$ and $g_{\mathrm{YM}}^2 = 4\pi g_{st}$ [10]. Thus, the size of the throat in string units is $\lambda^{1/4}$. This remarkable emergence of the 't Hooft coupling from gravitational considerations is at the heart of the success of the AdS/CFT correspondence. Moreover, the requirement $L \gg \sqrt{\alpha'}$ translates into $\lambda \gg 1$: the gravitational approach is valid when the 't Hooft coupling is very strong and the perturbative field-theoretic methods are not applicable.

Example: Thermal Gauge Theory from Near-Extremal D3 Branes

An important black hole observable is the Bekenstein–Hawking (BH) entropy, which is proportional to the area of the event horizon. For the 3-brane solution [7], the horizon is located at $r = r_0$. For $r_0 > 0$ the 3-brane carries some excess energy E above its extremal value, and the BH entropy is also non-vanishing. The Hawking temperature is then defined by $T^{-1} = \partial S_{\mathrm{BH}}/\partial E$.

Setting $r_0 \ll L$ in [9], we obtain a near-extremal 3-brane geometry, whose Hawking temperature is found to be $T = r_0/(\pi L^2)$. The eight-dimensional "area" of the horizon is

$$A_h = (r_0/L)^3 V_3 L^5\, \mathrm{vol}(S^5) = \pi^6 L^8 T^3 V_3 \qquad [12]$$

where V_3 is the spatial volume of the D3 brane (i.e., the volume of the x^1, x^2, x^3 coordinates). Therefore, the BH entropy is

$$S_{\mathrm{BH}} = \frac{2\pi A_h}{\kappa^2} = \frac{\pi^2}{2}N^2 V_3 T^3 \qquad [13]$$

This gravitational entropy of a near-extremal 3-brane of Hawking temperature T is to be identified with the entropy of $\mathcal{N} = 4$ supersymmetric $U(N)$ gauge theory (which lives on N coincident D3 branes) heated up to the same temperature.

The entropy of a free $U(N)$ $\mathcal{N} = 4$ supermultiplet – which consists of the gauge field, $6N^2$ massless scalars, and $4N^2$ Weyl fermions – can be calculated using the standard statistical mechanics of a massless gas (the blackbody problem), and the answer is

$$S_0 = \frac{2\pi^2}{3}N^2 V_3 T^3 \qquad [14]$$

It is remarkable that the 3-brane geometry captures the T^3 scaling characteristic of a conformal field theory (CFT) (in a CFT this scaling is guaranteed by the extensivity of the entropy and the absence of dimensionful parameters). Also, the N^2 scaling indicates the presence of $O(N^2)$ unconfined degrees

of freedom, which is exactly what we expect in the $\mathcal{N}=4$ supersymmetric U(N) gauge theory. But what is the explanation of the relative factor of 3/4 between S_{BH} and S_0? In fact, this factor is not a contradiction but rather a prediction about the strongly coupled $\mathcal{N}=4$ SYM theory at finite temperature. As we argued above, the supergravity calculation of the BH entropy, [13], is relevant to the $\lambda \to \infty$ limit of the $\mathcal{N}=4$ SU(N) gauge theory, while the free-field calculation, [14], applies to the $\lambda \to 0$ limit. Thus, the relative factor of 3/4 is not a discrepancy: it relates two different limits of the theory. Indeed, on general field-theoretic grounds, we expect that in the 't Hooft large-N limit, the entropy is given by

$$S = \frac{2\pi^2}{3} N^2 f(\lambda) V_3 T^3 \qquad [15]$$

The function f is certainly not constant: perturbative calculations valid for small $\lambda = g_{YM}^2 N$ give

$$f(\lambda) = 1 - \frac{3}{2\pi^2}\lambda + \frac{3+\sqrt{2}}{\pi^3}\lambda^{3/2} + \cdots \qquad [16]$$

Thus, the BH entropy in supergravity, [13], is translated into the prediction that

$$\lim_{\lambda\to\infty} f(\lambda) = \frac{3}{4} \qquad [17]$$

The Essentials of the AdS/CFT Correspondence

The AdS/CFT correspondence asserts a detailed map between the physics of type IIB string theory in the throat of the classical 3-brane geometry, that is, the region $r \ll L$, and the gauge theory living on a stack of D3 branes. As already noted, in this limit $r \ll L$, the extremal D3 brane geometry factors into a direct product of $AdS_5 \times S^5$. Moreover, the gauge theory on this stack of D3 branes is the maximally supersymmetric $\mathcal{N}=4$ SYM.

Since the horizon of the near-extremal 3-brane lies in the region $r \ll L$, the entropy calculation could have been carried out directly in the throat limit, where $H(r)$ is replaced by L^4/r^4. Another way to motivate the identification of the gauge theory with the throat is to think about the absorption of massless particles. In the D-brane description, a particle incident from asymptotic infinity is converted into an excitation of the stack of D-branes, that is, into an excitation of the gauge theory on the world volume. In the supergravity description, a particle incident from the asymptotic (large r) region tunnels into the $r \ll L$ region and produces an excitation of the throat. The fact that the two different descriptions of the absorption process give identical cross sections supports the identification of excitations of $AdS_5 \times S^5$ with the excited states of the $\mathcal{N}=4$ SYM theory.

Maldacena (1998) motivated this correspondence by thinking about the low-energy ($\alpha' \to 0$) limit of the string theory. On the D3 brane side, in this low-energy limit, the interaction between the D3 branes and the closed strings propagating in the bulk vanishes, leaving a pure $\mathcal{N}=4$ SYM theory on the D3 branes decoupled from type IIB superstrings in flat space. Around the classical 3-brane solutions, there are two types of low-energy excitations. The first type propagate in the bulk region, $r \gg L$, and have a cross section for absorption by the throat which vanishes as the cube of their energy. The second type are localized in the throat, $r \leq L$, and find it harder to tunnel into the asymptotically flat region as their energy is taken smaller. Thus, both the D3 branes and the classical 3-brane solution have two decoupled components in the low-energy limit, and in both cases, one of these components is type IIB superstrings in flat space. Maldacena conjectured an equivalence between the other two components.

Immediate support for this identification comes from symmetry considerations. The isometry group of AdS_5 is SO(2,4), and this is also the conformal group in $3+1$ dimensions. In addition, we have the isometries of S^5 which form $SU(4) \sim SO(6)$. This group is identical to the R-symmetry of the $\mathcal{N}=4$ SYM theory. After including the fermionic generators required by supersymmetry, the full isometry supergroup of the $AdS_5 \times S^5$ background is SU(2,2|4), which is identical to the $\mathcal{N}=4$ superconformal symmetry. We will see that, in theories with reduced supersymmetry, the S^5 factor is replaced by other compact Einstein spaces Y_5, but AdS_5 is the "universal" factor present in the dual description of any large-N CFT and makes the SO(2,4) conformal symmetry a geometric one.

The correspondence extends beyond the supergravity limit, and we must think of $AdS_5 \times Y_5$ as a background of string theory. Indeed, type IIB strings are dual to the electric flux lines in the gauge theory, providing a string-theoretic setup for calculating correlation functions of Wilson loops. Furthermore, if $N \to \infty$ while $g_{YM}^2 N$ is held fixed and finite, then there are string scale corrections to the supergravity limit (Maldacena 1998, Gubser *et al.* 1998, Witten 1998) which proceed in powers of $\alpha'/L^2 = (g_{YM}^2 N)^{-1/2}$. For finite N, there are also

string loop corrections in powers of $\kappa^2/L^8 \sim N^{-2}$. As expected, with $N \to \infty$ we can take the classical limit of the string theory on $AdS_5 \times Y_5$. However, in order to understand the large-N gauge theory with finite 't Hooft coupling, we should think of $AdS_5 \times Y_5$ as the target space of a two-dimensional sigma model describing the classical string physics.

Correlation Functions and the Bulk/Boundary Correspondence

A basic premise of the AdS/CFT correspondence is the existence of a one-to-one map between gauge-invariant operators in the CFT and fields (or extended objects) in AdS. Gubser *et al.* (1998) and Witten (1998) formulated precise methods for calculating correlation functions of various operators in a CFT using its dual formulation. A physical motivation for these methods comes from earlier calculations of absorption by 3-branes. When a wave is absorbed, it tunnels from asymptotic infinity into the throat region, and then continues to propagate toward smaller r. Let us separate the 3-brane geometry into two regions: $r \gtrsim L$ and $r \lesssim L$. For $r \lesssim L$ the metric is approximately that of $AdS_5 \times S^5$, while for $r \gtrsim L$ it becomes very different and eventually approaches the flat metric. Signals coming in from large r (small $z = L^2/r$) may be considered as disturbing the "boundary" of AdS_5 at $r \sim L$, and then propagating into the bulk of AdS_5. Discarding the $r \gtrsim L$ part of the 3-brane metric, the gauge theory correlation functions are related to the response of the string theory to boundary conditions at $r \sim L$. It is therefore natural to identify the generating functional of correlation functions in the gauge theory with the string theory path integral subject to the boundary conditions that $\phi(x, z) = \phi_0(x)$ at $z = L$ (at $z = \infty$ all fluctuations are required to vanish). In calculating correlation functions in a CFT, we will carry out the standard Euclidean continuation; then on the string theory side, we will work with L_5, which is the Euclidean version of AdS_5.

More explicitly, we identify a gauge theory quantity W with a string-theory quantity Z_{string}:

$$W[\phi_0(x)] = Z_{\text{string}}[\phi_0(x)] \qquad [18]$$

W generates the connected Euclidean Green's functions of a gauge-theory operator $\mathcal{O}$,

$$W[\phi_0(x)] = \left\langle \exp \int d^4x \phi_0 \mathcal{O} \right\rangle \qquad [19]$$

Z_{string} is the string theory path integral calculated as a functional of ϕ_0, the boundary condition on the field ϕ related to $\mathcal{O}$ by the AdS/CFT duality. In the large-N limit, the string theory becomes classical which implies

$$Z_{\text{string}} \sim e^{-I[\phi_0(x)]} \qquad [20]$$

where $I[\phi_0(x)]$ is the extremum of the classical string action calculated as a functional of ϕ_0. If we are further interested in correlation functions at very large 't Hooft coupling, then the problem of extremizing the classical string action reduces to solving the equations of motion in type IIB supergravity whose form is known explicitly. A simple example of such a calculation is presented in the next subsection.

Our reasoning suggests that from the point of view of the metric [5], the boundary conditions are imposed not quite at $z = 0$, which is the true boundary of L_5, but at some finite value $z = \epsilon$. It does not matter which value it is since the metric [5] is unchanged by an overall rescaling of the coordinates (z, x); thus, such a rescaling can take $z = L$ into $z = \epsilon$ for any ϵ. The physical meaning of this cutoff is that it acts as a UV regulator in the gauge theory. Indeed, the radial coordinate z is to be considered as the effective energy scale of the gauge theory, and decreasing z corresponds to increasing the energy. A safe method for performing calculations of correlation functions, therefore, is to keep the cutoff on the z-coordinate at intermediate stages and remove it only at the end.

Two-Point Functions and Operator Dimensions

In the following, we present a brief discussion of two-point functions of scalar operators in CFT_d. The corresponding field in L_{d+1} is a scalar field of mass m whose Euclidean action is proportional to

$$\frac{1}{2} \int d^d x \, dz \, z^{-d+1} \left[(\partial_z \phi)^2 + \sum_{a=1}^{d} (\partial_a \phi)^2 + \frac{m^2 L^2}{z^2} \phi^2 \right] \qquad [21]$$

In calculating correlation functions of vertex operators from the AdS/CFT correspondence, the first problem is to reconstruct an on-shell field in L_{d+1} from its boundary behavior. The near-boundary, that is, small z, behavior of the classical solution is

$$\phi(z, x) \to z^{d-\Delta} [\phi_0(x) + O(z^2)] + z^{\Delta} [A(x) + O(z^2)] \qquad [22]$$

where Δ is one of the roots of

$$\Delta(\Delta - d) = m^2 L^2 \qquad [23]$$

$\phi_0(x)$ is regarded as a "source" in [19] that couples to the dual gauge-invariant operator $\mathcal{O}$ of dimension Δ, while $A(x)$ is related to the expectation value,

$$A(x) = \frac{1}{2\Delta - d}\langle \mathcal{O}(x)\rangle \qquad [24]$$

It is possible to regularize the Euclidean action to obtain the following value as a functional of the source:

$$I[\phi_0(x)] = -(\Delta - (d/2))\pi^{-d/2}\frac{\Gamma(\Delta)}{\Gamma(\Delta - (d/2))} \times \int \mathrm{d}^d x \int \mathrm{d}^d x' \frac{\phi_0(x)\phi_0(x')}{|x - x'|^{2\Delta}} \qquad [25]$$

Varying twice with respect to ϕ_0, we find that the two-point function of the corresponding operator is

$$\langle \mathcal{O}(x)\mathcal{O}(x')\rangle = \frac{(2\Delta - d)\Gamma(\Delta)}{\pi^{d/2}\Gamma(\Delta - (d/2))}\frac{1}{|x - x'|^{2\Delta}} \qquad [26]$$

Which of the two roots, Δ_+ or Δ_-, of [23]

$$\Delta_\pm = \frac{d}{2} \pm \sqrt{\frac{d^2}{4} + m^2 L^2} \qquad [27]$$

should we choose for the operator dimension? For positive m^2, Δ_+ is certainly the right choice: here the other root, Δ_-, is negative. However, it turns out that for

$$-\frac{d^2}{4} < m^2 L^2 < -\frac{d^2}{4} + 1 \qquad [28]$$

both roots of [23] may be chosen. Thus, there are two possible CFTs corresponding to the same classical AdS action: in one of them the corresponding operator has dimension Δ_+, while in the other the dimension is Δ_-. We note that Δ_- is bounded from below by $(d-2)/2$, which is precisely the unitarity bound on dimensions of scalar operators in d-dimensional field theory! Thus, the ability to choose dimension Δ_- is crucial for consistency of the AdS/CFT duality.

Whether string theory on $\mathrm{AdS}_5 \times Y_5$ contains fields with m^2 in the range [28] depends on Y_5. The example discussed in the next section, $Y_5 = T^{1,1}$, turns out to contain such fields, and the possibility of having dimension Δ_-, [27], is crucial for consistency of the AdS/CFT duality in that case. However, for $Y_5 = S^5$, which is dual to the $\mathcal{N} = 4$ large-N SYM theory, there are no such fields and all scalar dimensions are given by [27].

The operators in the $\mathcal{N} = 4$ large-N SYM theory naturally break up into two classes: those that correspond to the Kaluza–Klein states of supergravity and those that correspond to massive string states. Since the radius of the S^5 is L, the masses of the Kaluza–Klein states are proportional to $1/L$. Thus, the dimensions of the corresponding operators are independent of L and therefore also of λ. On the gauge-theory side, this independence is explained by the fact that the supersymmetry protects the dimensions of certain operators from being renormalized: they are completely determined by the representation under the superconformal symmetry. All families of the Kaluza–Klein states, which correspond to such protected operators, were classified long ago. Correlation functions of such operators in the strong 't Hooft coupling limit may be obtained from the dependence of the supergravity action on the boundary values of corresponding Kaluza–Klein fields, as in [19]. A variety of explicit calculations have been performed for two-, three-, and even four-point functions. The four-point functions are particularly interesting because their dependence on operator positions is not determined by the conformal invariance.

On the other hand, the masses of string excitations are $m^2 = 4n/\alpha'$, where n is an integer. For the corresponding operators the formula [27] predicts that the dimensions do depend on the 't Hooft coupling and, in fact, blow up for large $\lambda = g_{\mathrm{YM}}^2 N$ as $2\lambda^{1/4}\sqrt{n}$.

Calculation of Wilson Loops

The Wilson loop operator of a nonabelian gauge theory

$$W(\mathcal{C}) = \mathrm{tr}\left[P \exp\left(\mathrm{i}\oint_{\mathcal{C}} A\right)\right] \qquad [29]$$

involves the path-ordered integral of the gauge connection A along a contour $\mathcal{C}$. For $\mathcal{N} = 4$ SYM, one typically uses a generalization of this loop operator which incorporates other fields in the $\mathcal{N} = 4$ multiplet, the adjoint scalars and fermions. Using a rectangular contour, we can calculate the quark–antiquark potential from the expectation value $\langle W(\mathcal{C})\rangle$. One thinks of the quarks located a distance L apart for a time T, yielding

$$\langle W\rangle \sim \mathrm{e}^{-TV(L)} \qquad [30]$$

where $V(L)$ is the potential.

According to Maldacena, and Rey and Yee, the AdS/CFT correspondence relates the Wilson loop expectation value to a sum over string world sheets ending on the boundary of $L_5(z = 0)$ along the contour $\mathcal{C}$:

$$\langle W\rangle \sim \int \mathrm{e}^{-S} \qquad [31]$$

where S is the action functional of the string world sheet. In the large 't Hooft coupling limit $\lambda \to \infty$, this path integral may be evaluated using a saddle-point approximation. The leading answer is $\sim e^{-S_0}$, where S_0 is the action for the classical solution, which is proportional to the minimal area of the string world sheet in L_5 subject to the boundary conditions. The area as currently defined is actually divergent, and to regularize it one must position the contour at $z = \epsilon$ (this is the same type of regulator as used in the definition of correlation functions).

Consider a circular Wilson loop of radius a. The action of the corresponding classical string world sheet is

$$S_0 = \sqrt{\lambda}\left(\frac{a}{\epsilon} - 1\right) \qquad [32]$$

Subtracting the linearly divergent term, which is proportional to the length of the contour, one finds

$$\ln\langle W\rangle = \sqrt{\lambda} + O(\ln\lambda) \qquad [33]$$

a result which has been duplicated in field theory by summing certain classes of rainbow Feynman diagrams in $\mathcal{N} = 4$ SYM. From these sums, one finds

$$\langle W\rangle_{\text{rainbow}} = \frac{2}{\sqrt{\lambda}} I_1\left(\sqrt{\lambda}\right) \qquad [34]$$

where I_1 is a Bessel function. This formula is one of the few available proposals for extrapolation of an observable from small to large coupling. At large λ,

$$\langle W\rangle_{\text{rainbow}} \sim \sqrt{\frac{2}{\pi}} \frac{e^{\sqrt{\lambda}}}{\lambda^{3/4}} \qquad [35]$$

in agreement with the geometric prediction.

The quark–antiquark potential is extracted from a rectangular Wilson loop of width L and length T. After regularizing the divergent contribution to the energy, one finds the attractive potential

$$V(L) = -\frac{4\pi^2\sqrt{\lambda}}{\Gamma(1/4)^4 L} \qquad [36]$$

The Coulombic $1/L$ dependence is required by the conformal invariance of the theory. The fact that the potential scales as the square root of the 't Hooft coupling indicates some screening of the charges at large coupling.

Conformal Field Theories and Einstein Manifolds

Interesting generalizations of the duality between $AdS_5 \times S^5$ and $\mathcal{N} = 4$ SYM with less supersymmetry and more complicated gauge groups can be

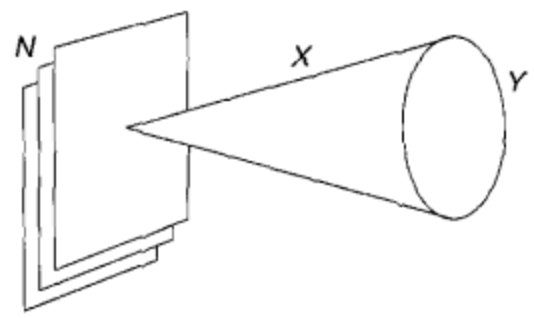


Figure 1 D3 branes placed at the tip of a Ricci-flat cone X.

produced by placing D3 branes at the tip of a Ricci-flat six-dimensional cone X (see **Figure 1**). The cone metric may be cast in the form

$$ds_X^2 = dr^2 + r^2\, ds_Y^2 \qquad [37]$$

where Y is the level surface of X. In particular, Y is a positively curved Einstein manifold, that is, one for which $R_{ij} = 4g_{ij}$. In order to preserve the $\mathcal{N} = 1$ supersymmetry, X must be a Calabi–Yau space; then Y is defined to be Sasaki–Einstein.

The D3 branes appear as a point in X and span the transverse Minkowski space $\mathbb{R}^{3,1}$. The ten-dimensional metric they produce assumes the form [9], but with the sphere metric $d\Omega_5^2$ replaced by the metric on Y, ds_Y^2. The equality of tensions [10] now requires that

$$L^4 = \frac{\sqrt{\pi}\kappa N}{2\,\text{vol}(Y)} = 4\pi g_s N \alpha'^2 \frac{\pi^3}{\text{vol}(Y)} \qquad [38]$$

In the near-horizon limit, $r \to 0$, the geometry factors into $AdS_5 \times Y$. Because the D3 branes are located at a singularity, the gauge theory becomes much more complicated, typically involving a product of several $SU(N)$ factors coupled to matter in bifundamental representations, often described using a quiver diagram (see **Figure 2** for an example).

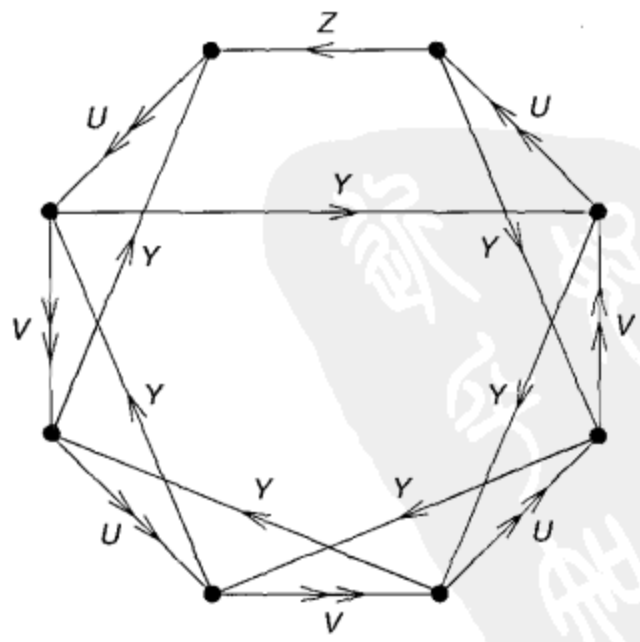


Figure 2 The quiver for $Y^{4,3}$. Each node corresponds to an $SU(N)$ gauge group and each arrow to a bifundamental chiral superfield.

The simplest examples of X are orbifolds $\mathbb{C}^3/\Gamma$, where Γ is a discrete subgroup of SO(6). Indeed, if $\Gamma \subset \mathrm{SU}(3)$, then $\mathcal{N}=1$ supersymmetry is preserved. The level surface of such an X is $Y=S^5/\Gamma$. In this case, the product structure of the gauge theory can be motivated by thinking about image stacks of D3 branes from the action of Γ.

The next simplest example of a Calabi–Yau cone X is the conifold which may be described by the following equation in four complex variables:

$$\sum_{a=1}^{4} z_a{}^2 = 0 \qquad [39]$$

Since this equation is symmetric under an overall rescaling of the coordinates, this space is a cone. The level surface Y of the conifold is a coset manifold $T^{1,1}=(\mathrm{SU}(2)\times\mathrm{SU}(2))/\mathrm{U}(1)$. This space has the $\mathrm{SO}(4)\sim\mathrm{SU}(2)\times\mathrm{SU}(2)$ symmetry which rotates the z's, and also the U(1) R-symmetry under $z_a \to \mathrm{e}^{\mathrm{i}\theta} z_a$. The metric on $T^{1,1}$ is known explicitly; it assumes the form of an S^1 bundle over $S^2\times S^2$.

The supersymmetric field theory on the D3 branes probing the conifold singularity is $\mathrm{SU}(N)\times\mathrm{SU}(N)$ gauge theory coupled to two chiral superfields, A_i, in the $(N, \overline{N})$ representation and two chiral superfields, B_j, in the $(\overline{N}, N)$ representation. The A's transform as a doublet under one of the global SU(2)'s, while the B's transform as a doublet under the other SU(2). Cancelation of the anomaly in the U(1) R-symmetry requires that the A's and the B's each have R-charge 1/2. For consistency of the duality, it is necessary that we add an exactly marginal superpotential which preserves the $\mathrm{SU}(2)\times\mathrm{SU}(2)\times\mathrm{U}(1)_{\mathrm{R}}$ symmetry of the theory. Since a marginal superpotential has R-charge equal to 2 it must be quartic, and the symmetries fix it uniquely up to overall normalization:

$$W = \epsilon^{ij}\epsilon^{kl}\,\mathrm{tr}\,A_iB_kA_jB_l \qquad [40]$$

There are in fact infinite families of Calabi–Yau cones X, but there are two problems one faces in studying these generalized AdS/CFT correspondences. The first is geometric: the cones X are not all well understood and only for relatively few do we have explicit metrics. However, it is often possible to calculate important quantities such as the vol(Y) without knowing the metric. The second problem is gauge theoretic: although many techniques exist, there is no completely general procedure for constructing the gauge theory on a stack of D-branes at an arbitrary singularity.

Let us mention two important classes of Calabi–Yau cones X. The first class consists of cones over the so-called $Y^{p,q}$ Sasaki–Einstein spaces. Here, p and q are integers with $p \geq q$. Gauntlett *et al.* (2004) discovered metrics on all the $Y^{p,q}$, and the quiver gauge theories that live on the D-branes probing the singularity are now known. Making contact with the simpler examples discussed above, the $Y^{p,0}$ are orbifolds of $T^{1,1}$ while the $Y^{p,p}$ are orbifolds of S^5.

In the second class of cones X, a del Pezzo surface shrinks to zero size at the tip of the cone. A del Pezzo surface is an algebraic surface of complex dimension 2 with positive first Chern class. One simple del Pezzo surface is a complex projective space of dimension 2, $\mathbb{P}^2$, which gives rise to the $\mathcal{N}=1$ preserving $S^5/\mathbb{Z}_3$ orbifold. Another simple case is $\mathbb{P}^1\times\mathbb{P}^1$, which leads to $T^{1,1}/\mathbb{Z}_2$. The remaining del Pezzos surfaces B_k are $\mathbb{P}^2$ blown up at k points, $1\leq k\leq 8$. The cone where B_1 shrinks to zero size has level surface $Y^{2,1}$. Gauge theories for all the del Pezzos have been constructed. Except for the three del Pezzos just discussed, and possibly also for B_6, metrics on the cones over these del Pezzos are not known. Nevertheless, it is known that for $3\leq k\leq 8$, the volume of the Sasaki–Einstein manifold Y associated with B_k is $\pi^3(9-k)/27$.

The Central Charge

The central charge provides one of the most amazing ways to check the generalized AdS/CFT correspondences. The central charge c and conformal anomaly a can be defined as coefficients of certain curvature invariants in the trace of the stress energy tensor of the conformal gauge theory:

$$\langle T^{\alpha}_{\alpha}\rangle = -aE_4 - cI_4 \qquad [41]$$

(The curvature invariants E_4 and I_4 are quadratic in the Riemann tensor and vanish for Minkowski space.) As discussed above, correlators such as $\langle T_{\mu\nu}\rangle$ can be calculated from supergravity, and one finds

$$a = c = \frac{\pi^3 N^2}{4\,\mathrm{vol}(Y)} \qquad [42]$$

On the gauge-theory side of the correspondence, anomalies completely determine a and c:

$$a = \tfrac{3}{32}(3\,\mathrm{tr}\,R^3 - \mathrm{tr}\,R)$$
$$c = \tfrac{1}{32}(9\,\mathrm{tr}\,R^3 - 5\,\mathrm{tr}\,R) \qquad [43]$$

The trace notation implies a sum over the R-charges of all of the fermions in the gauge theory. (From the geometric knowledge that $a=c$, we can conclude that $\mathrm{tr}\,R=0$.)

The R-charges can be determined using the principle of a-maximization. For a superconformal gauge theory, the R-charges of the fermions maximize a subject to the constraints that the

Novikov–Shifman–Vainshtein–Zakharov (NSVZ) beta function of each gauge group vanishes and the R-charge of each superpotential term is 2.

For the $Y^{p,q}$ spaces mentioned above, one finds that

$$\mathrm{vol}(Y^{p,q}) = \frac{q^2\left(2p + \sqrt{4p^2 - 3q^2}\right)}{3p^2\left(3q^2 - 2p^2 + p\sqrt{4p^2 - 3q^2}\right)}\pi^3 \qquad [44]$$

The gauge theory consists of $p - q$ fields Z, $p + q$ fields Y, $2p$ fields U, and $2q$ fields V. These fields all transform in the bifundamental representation of a pair of SU(N) gauge groups (the quiver diagram for $Y^{4,3}$ is given in **Figure 2**). The NSVZ beta function and superpotential constraints determine the R-charges up to two free parameters x and y. Let x be the R-charge of Z and y the R-charge of Y. Then the U have R-charge $1 - (1/2)(x + y)$ and the V have R-charge $1 + (1/2)(x - y)$.

The technique of a maximization leads to the result

$$x = \frac{1}{3q^2}\left(-4p^2 + 2pq + 3q^2 + (2p - q)\sqrt{4p^2 - 3q^2}\right)$$

$$y = \frac{1}{3q^2}\left(-4p^2 - 2pq + 3q^2 + (2p + q)\sqrt{4p^2 - 3q^2}\right)$$

Thus, as calculated by Benvenuti *et al.* (2004) and Bertolini *et al.* (2004)

$$a(Y^{p,q}) = \frac{\pi^3 N^2}{4\,\mathrm{vol}(Y^{p,q})} \qquad [45]$$

in remarkable agreement with the prediction [42] of the AdS/CFT duality.

A Path to a Confining Theory

There exists an interesting way of breaking the conformal invariance for spaces Y whose topology includes an S^2 factor (examples of such spaces include $T^{1,1}$ and $Y^{p,q}$, which are topologically $S^2 \times S^3$). At the tip of the cone over Y, one may add M wrapped D5 branes to the N D3 branes. The gauge theory on such a combined stack is no longer conformal; it exhibits a novel pattern of quasiperiodic renormalization group flow, called a duality cascade.

To date, the most extensive study of a theory of this type has been carried out for the conifold, where one finds an $\mathcal{N} = 1$ supersymmetric SU(N) × SU($N + M$) theory coupled to chiral superfields A_1, A_2 in the $(\mathbf{N}, \overline{\mathbf{N+M}})$ representation, and B_1, B_2 in the $(\overline{\mathbf{N}}, \mathbf{N+M})$ representation. D5 branes source RR 3-form flux; hence, the supergravity dual of this theory has to include M units of this flux. Klebanov and Strassler (2000) found an exact nonsingular supergravity solution incorporating the 3-form and the 5-form RR field strengths, and their back-reaction on the geometry. This back-reaction creates a "geometric transition" to the deformed conifold

$$\sum_{a=1}^{4} z_a^2 = \epsilon^2 \qquad [46]$$

and introduces a "warp factor" so that the full ten-dimensional geometry has the form

$$ds_{10}{}^2 = h^{-1/2}(\tau)(-(dx^0)^2 + (dx^i)^2) + h^{1/2}(\tau)\, d\tilde{s}_6{}^2 \qquad [47]$$

where $d\tilde{s}_6{}^2$ is the Calabi–Yau metric of the deformed conifold, which is known explicitly.

The field-theoretic interpretation of this solution is unconventional. After a finite amount of RG flow, the SU($N + M$) group undergoes a Seiberg duality transformation. After this transformation, and an interchange of the two gauge groups, the new gauge theory is SU($\tilde{N}$) × SU($\tilde{N} + M$) with the same matter and superpotential, and with $\tilde{N} = N - M$. The self-similar structure of the gauge theory under the Seiberg duality is the crucial fact that allows this pattern to repeat many times. If $N = (k + 1)M$, where k is an integer, then the duality cascade stops after k steps, and we find SU(M) × SU($2M$) gauge theory. This IR gauge theory exhibits a multitude of interesting effects visible in the dual supergravity background. One of them is confinement, which follows from the fact that the warp factor h is finite and nonvanishing at the smallest radial coordinate, $\tau = 0$. The methods presented in the section "Calculation of Wilson loops," then imply that the quark–antiquark potential grows linearly at large distances. Other notable IR effects are chiral symmetry breaking and the Goldstone mechanism. Particularly interesting is the appearance of an entire "baryonic branch" of the moduli space in the gauge theory, whose existence has been demonstrated also in the dual supergravity language.

Conclusions

This article tries to present a logical path from studying gravitational properties of D-branes to the formulation of an exact duality between conformal field theories and string theory in anti-de Sitter backgrounds, and also sketches some methods for breaking the conformal symmetry. Due to space limitations, many aspects and applications of the AdS/CFT correspondence have been omitted. At the moment, practical applications of this duality are limited mainly to very strongly coupled, large-N gauge theories, where the dual string description is well approximated by classical supergravity. To understand the implications of the duality for more general parameters, it is necessary to find better

methods for attacking the world sheet approach to string theories in anti-de Sitter backgrounds with RR background fields turned on. When such methods are found, it is likely that the material presented here will have turned out to be just a tiny tip of a monumental iceberg of dualities between fields and strings.

Acknowledgments

The authors are very grateful to all their collaborators on gauge/string duality for their valuable input over many years. The research of I R Klebanov is supported in part by the National Science Foundation (NSF) grant no. PHY-0243680. The research of C P Herzog is supported in part by the NSF under grant no. PHY99-07949. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

See also: Brane Construction of Gauge Theories; Branes and Black Hole Statistical Mechanics; Einstein Equations: Exact Solutions; Gauge Theories from Strings; Large-*N* and Topological Strings; Large-*N* Dualities; Mirror Symmetry: A Geometric Survey; Quantum Chromodynamics; Quantum Field Theory in Curved Spacetime; Superstring Theories.

Further Reading

Aharony O, Gubser SS, Maldacena JM, Ooguri H, and Oz Y (2000) Large N field theories, string theory and gravity. *Physics Reports* 323: 183 (arXiv:hep-th/9905111).

Benvenuti S, Franco S, Hanany A, Martelli D, and Sparks J (2005) An infinite family of superconformal quiver gauge theories with Sasaki–Einstein duals. *JHEP* 0506: 064 (arXiv:hep-th/0411264).

Bertolini M, Bigazzi F, and Cotrone AL (2004) New checks and subtleties for AdS/CFT and a-maximization. *JHEP* 0412: 024 (arXiv:hep-th/0411249).

Bigazzi F, Cotrone AL, Petrini M, and Zaffaroni A (2002) Supergravity duals of supersymmetric four dimensional gauge theories. *Rivista del Nuovo Cimento* 25N12: 1 (arXiv:hep-th/0303191).

D'Hoker E and Freedman DZ (2002) Supersymmetric gauge theories and the AdS/CFT correspondence, arXiv:hep-th/0201253.

Gauntlett J, Martelli D, Sparks J, and Waldram D (2004) Sasaki–Einstein metrics on $S^2 \times S^3$. *Advances in Theoretical Mathematics in Physics* 8: 711 (arXiv:hep-th/0403002).

Gubser SS, Klebanov IR, and Polyakov AM (1998) Gauge theory correlators from noncritical string theory. *Physics Letters B* 428: 105 (hep-th/9802109).

Herzog CP, Klebanov IR, and Ouyang P (2002) D-branes on the conifold and $N=1$ gauge/gravity dualities, arXiv:hep-th/0205100.

Klebanov IR (2000) TASI lectures: introduction to the AdS/CFT correspondence, arXiv:hep-th/0009139.

Klebanov IR and Strassler MJ (2000) Supergravity and a confining gauge theory: Duality cascades and χ-resolution of naked singularities. *JHEP* 0008: 052 (arXiv:hep-th/0007191).

Maldacena J (1998) The large N limit of superconformal field theories and supergravity. *Advances in Theoretical and Mathematical Physics* 2: 231 (hep-th/9711200).

Maldacena JM (1998) Wilson loops in large N field theories. *Physics Review Letters* 80: 4859 (arXiv:hep-th/9803002).

Polchinski J (1998) *String Theory*. Cambridge: Cambridge University Press.

Polyakov AM (1999) The wall of the cave. *International Journal of Modern Physics A* 14: 645.

Rey SJ and Yee JT (2001) Macroscopic strings as heavy quarks in large N gauge theory and anti-de Sitter supergravity. *European Physics Journal C* 22: 379 (arXiv:hep-th/9803001).

Semenoff GW and Zarembo K (2002) Wilson loops in SYM theory: from weak to strong coupling. *Nuclear Physics Proceeding Supplements* 108: 106 (arXiv:hep-th/0202156).

Strassler MJ The duality cascade, TASI 2003 lectures, arXiv:hep-th/0505153.

Witten E (1998) Anti-de Sitter space and holography. *Advances in Theoretical and Mathematical Physics* 2: 253 (hep-th/9802150).

Aharonov–Bohm Effect

M Socolovsky, Universidad Nacional Autónoma de México, México DF, México

© 2006 Elsevier Ltd. All rights reserved.

Introduction

In classical electrodynamics, the interaction of charged particles with the electromagnetic field is local, through the pointlike coupling of the electric charge of the particles with the electric and magnetic fields, $\boldsymbol{E}$ and $\boldsymbol{B}$, respectively. This is mathematically expressed by the Lorentz-force law. The scalar and vector potentials, φ and $\boldsymbol{A}$, which are the time and space components of the relativistic 4-potential A_μ, are considered auxiliary quantities in terms of which the field strengths $\boldsymbol{E}$ and $\boldsymbol{B}$, the observables, are expressed in a gauge-invariant manner. The homogeneous or first pair of Maxwell equations are a direct consequence of the definition of the field strengths in terms of A_μ. The inhomogeneous or second pair of Maxwell equations, which involve the charges and currents present in the problem, are also usually written in terms of $\boldsymbol{E}$ and $\boldsymbol{B}$; however when writing them in terms of A_μ, the number of degrees of freedom of the electromagnetic field is explicitly reduced from six to four; and finally, with two additional gauge transformations, one ends with the two physical degrees of freedom of the electromagnetic field.

In quantum mechanics, however, both the Schrödinger equation and the path-integral approaches for scalar and unpolarized charged particles in the presence of electromagnetic fields, are written in terms of the potential and not of the field strengths. Even in the case of the Schrödinger–Pauli equation for spin 1/2 electrons with magnetic moment $\boldsymbol{\mu}$ interacting with a magnetic field $\boldsymbol{B}$, one knows that the coupling $-\boldsymbol{\mu}\cdot\boldsymbol{B}$ is the nonrelativistic limit of the Dirac equation, which depends on A_μ but not on $\boldsymbol{E}$ and $\boldsymbol{B}$. Since gauge invariance also holds in the quantum domain, it was thought that $\boldsymbol{A}$ and φ were mere auxiliary quantities, like in the classical case.

Aharonov and Bohm, in 1959, predicted a quantum interference effect due to the motion of charged particles in regions where $\boldsymbol{B}(\boldsymbol{E})$ vanishes, but not $\boldsymbol{A}(\varphi)$, leading to a nonlocal gauge-invariant effect depending on the flux of the magnetic field in the inaccessible region, in the magnetic case, and on the difference of the integrals over time of time-varying potentials, in the electric case. (The magnetic effect was already noticed 10 years before by Ehrenberg and Siday in a paper on the refractive index of electrons.)

In the context of the Schrödinger equation, one can show that due to gauge invariance, if ψ_0 is a solution to the equation in the absence of an electromagnetic potential, then the product of $\psi_0(\boldsymbol{x})$ times the integral of A_μ over a path joining an arbitrary reference point $\boldsymbol{x}_0$ to $\boldsymbol{x}$ is also a solution, if the integral is path independent. However, it is the path integral of Feynman which in the formulas for propagators of charged particles in the presence of electromagnetic fields clearly shows that *the action of these fields on charged particles is nonlocal, and it is given by the celebrated non-integrable (path-dependent) phase factor of Wu and Yang* (1975). Moreover, this fact provides an additional proof of the nonlocal character of quantum mechanics: to surround fluxes, or to develop a potential difference, the particle has to travel simultaneously at least through two paths.

Thus, the fact that the Aharonov–Bohm (A–B) effect was verified experimentally, by Chambers and others, demonstrates the necessity of introducing the (gauge-dependent) potential A_μ in describing the electromagnetic interactions of the quantum particle. This is widely regarded as the single most important piece of evidence for electromagnetism being a gauge theory. Moreover, it shows, to paraphrase Yang, that the field underdescribes the physical theory, while the potential overdescribes it, and it is the phase factor which describes it exactly.

The content of this article is essentially twofold. The first four sections are mainly physical, where we describe the magnetic A–B effect using the Schrödinger equation and the Feynman path integral. The fifth section is geometrical and is the longest of the article. We describe the effect in the context of fiber bundles and connections, namely as a result of the coupling of the wave function (section of an associated bundle) to a nontrivial flat connection (non-pure gauge vector potential with zero magnetic field) in a trivial bundle (the A–B bundle) with topologically nontrivial (non-simply-connected) base space. We discuss the moduli space of flat connections and the holonomy groups giving the phase shifts of the interference patterns. Finally, in the last section, we briefly comment on the nonabelian A–B effect.

Electromagnetic Fields in Classical Physics

In classical physics, the motion of charged particles in the presence of electromagnetic fields is governed by the equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{p} = q\left(\boldsymbol{E} + \frac{\boldsymbol{v}}{c} \times \boldsymbol{B}\right) \qquad [1]$$

where

$$\boldsymbol{p} = \frac{m\boldsymbol{v}}{\sqrt{1 - (v^2/c^2)}}$$

is the mechanical momentum of the particle with electric charge q, mass m, and velocity $\boldsymbol{v} = \dot{\boldsymbol{x}}$ (c is the velocity of light in vacuum, and for $|\boldsymbol{v}| \ll c$ the left-hand side (LHS) of [1] is approximately $m\boldsymbol{v}$); the right-hand side (RHS) is the Lorentz force, where $\boldsymbol{E}$ and $\boldsymbol{B}$ are, respectively, the electric and magnetic fields at the spacetime point $(t, \boldsymbol{x})$ where the particle is located. Equation [1] is easily derived from the Euler–Lagrange equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \boldsymbol{v}}\right) - \frac{\partial L}{\partial \boldsymbol{x}} = 0 \qquad [2]$$

with the Lagrangian L given by the sum of the free Lagrangian for the particle,

$$L_0 = -mc^2\sqrt{1 - \frac{v^2}{c^2}} \qquad [3]$$

and the Lagrangian describing the particle–field interaction,

$$L_{\mathrm{int}} = \frac{q}{c}\boldsymbol{A} \cdot \boldsymbol{v} - q\varphi \qquad [4]$$

In [4], $\boldsymbol{A}$ and φ are, respectively, the vector potential and the scalar potential, which together form the 4-potential $A_\mu = (A_0, -\boldsymbol{A}) = (\varphi, -A^i)$, $i = 1, 2, 3$, in terms of which the electric and magnetic field strengths are given by

$$\boldsymbol{E} = -\frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{A} - \nabla\varphi \qquad [5a]$$

$$\boldsymbol{B} = \nabla \times \boldsymbol{A} \qquad [5b]$$

The classical action corresponding to a given path of the particle is

$$\begin{aligned} S &= \int_{t_1}^{t_2} \mathrm{d}t\, L = \int_{t_1}^{t_2} \mathrm{d}t (L_0 + L_{\mathrm{int}}) \\ &= \int_{t_1}^{t_2} \mathrm{d}t\, L_0 + \int_{t_1}^{t_2} \mathrm{d}t\, L_{\mathrm{int}} \equiv S_0 + S_{\mathrm{int}} \end{aligned} \qquad [6]$$

$\boldsymbol{E}$, $\boldsymbol{B}$, and S are invariant under the gauge transformation

$$\boldsymbol{A} \to \boldsymbol{A}' = \boldsymbol{A} - \nabla\Lambda \qquad [7a]$$

$$\varphi \to \varphi' = \varphi + \frac{1}{c}\frac{\partial}{\partial t}\Lambda \qquad [7b]$$

where Λ is a real-valued differentiable scalar function (at least of class C^2) on spacetime. That is, if $\boldsymbol{E}'$, $\boldsymbol{B}'$, and S'_{int} are defined in terms of $\boldsymbol{A}'$ and φ' as $\boldsymbol{E}$, $\boldsymbol{B}$, and S_{int} are defined in terms of $\boldsymbol{A}$ and φ, then $\boldsymbol{E}' = \boldsymbol{E}$, $\boldsymbol{B}' = \boldsymbol{B}$, and $S'_{\mathrm{int}} = S_{\mathrm{int}}$. This fact leads to the concept that, classically, the observables $\boldsymbol{E}$ and $\boldsymbol{B}$ are the physical quantities, while A_μ is only an auxiliary quantity. Also, and most important in the present context, eqn [1] states that the motion of the particles is determined by the values or state of the field strengths in an infinitesimal neighborhood of the particles, that is, classically, $\boldsymbol{E}$ and $\boldsymbol{B}$ act locally. If one defines the differential 1-form $A \equiv A_\mu \mathrm{d}x^\mu$ (with $\mathrm{d}x^0 = c\,\mathrm{d}t$), then the components of the differential 2-form $F = \mathrm{d}A = (1/2)(\partial_\mu A_\nu - \partial_\nu A_\mu)\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \equiv (1/2)F_{\mu\nu}\,\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$ are precisely the electric and magnetic fields:

$$F_{\mu\nu} = \begin{pmatrix} 0 & E^1 & E^2 & E^3 \\ -E^1 & 0 & -B^3 & B^2 \\ -E^2 & B^3 & 0 & -B^1 \\ -E^3 & -B^2 & B^1 & 0 \end{pmatrix} \qquad [8]$$

At the level of A,

$$\mathrm{d}F = \mathrm{d}^2 A = 0 \qquad [9]$$

is an identity, but at the level of $\boldsymbol{E}$ and $\boldsymbol{B}$, [9] amounts to the homogeneous (or first pair of) Maxwell equations obeyed by the field strengths:

$$\nabla \cdot \boldsymbol{B} = 0 \qquad [10a]$$

$$\nabla \times \boldsymbol{E} + \frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{B} = 0 \qquad [10b]$$

Therefore, these equations have a geometrical origin. The second pair of Maxwell equations is dynamical, and is obtained from the field action (in the Heaviside system of units)

$$S_{\mathrm{field}} = -\frac{1}{4c}\int \mathrm{d}^4 x F_{\mu\nu}F^{\mu\nu} \qquad [11]$$

which leads to

$$\nabla \cdot \boldsymbol{E} = 4\pi\rho \qquad [12a]$$

$$\nabla \times \boldsymbol{B} - \frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{E} = \frac{4\pi \boldsymbol{j}}{c} \qquad [12b]$$

where $(\rho, -\boldsymbol{j}) = (j^0, -\boldsymbol{j})$ is the 4-current satisfying, as a consequence of [12a] and [12b], the conservation law

$$\partial_\mu j^\mu = 0 \qquad [13]$$

For a pointlike particle, $\rho(t, \boldsymbol{x}) = q\delta^3(\boldsymbol{x} - \boldsymbol{x}(t))$ and $\boldsymbol{j} = \rho\boldsymbol{v}$.

Electromagnetic Fields in Quantum Physics

In quantum physics, the motion of charged particles in external electromagnetic fields is governed by the Schrödinger equation or, equivalently, by the Feynman path integral. In both cases, however, it is the 4-potential A_μ which appears in the equations, and not the field strengths. For simplicity, we consider here scalar (spinless) charged particles or unpolarized electrons (spin-(1/2)particles), both of which, in the nonrelativistic approximation, can be described quantum mechanically by a complex wave function $\psi(t,\boldsymbol{x})$.

To derive the Schrödinger equation, one starts from the classical Hamiltonian

$$H = \boldsymbol{P}\cdot\boldsymbol{v} - L - mc^2 = \frac{1}{2}\left(\boldsymbol{P} - \frac{q}{c}\boldsymbol{A}\right)^2 + q\varphi \qquad [14]$$

where

$$\boldsymbol{P} = \frac{\partial}{\partial \boldsymbol{v}}L = \boldsymbol{p} + \frac{q}{c}\boldsymbol{A}$$

is the canonical momentum of the particle, and we have subtracted its rest energy. The replacements $\boldsymbol{P} \to -\mathrm{i}\hbar\nabla$ and $H \to \mathrm{i}\hbar\partial/\partial t$ lead to

$$\begin{aligned}\mathrm{i}\hbar\frac{\partial}{\partial t}\psi &= \left(\frac{1}{2m}\left(\mathrm{i}\hbar\nabla + \frac{q}{c}\boldsymbol{A}\right)^2 + q\varphi\right)\psi \\ &= \left(-\frac{\hbar^2}{2m}\nabla^2 + \frac{q^2}{2mc^2}A^2\right. \\ &\quad \left. + \frac{\mathrm{i}\hbar q}{2mc}\nabla\cdot\boldsymbol{A} + \frac{\mathrm{i}\hbar q}{mc}\boldsymbol{A}\cdot\nabla + q\varphi\right)\psi \qquad [15]\end{aligned}$$

The gauge transformation [7a] and [7b] is a symmetry of this equation, if simultaneously to the change of the 4-potential, the wave function transforms as follows:

$$\psi(t,\boldsymbol{x}) \to \psi'(t,\boldsymbol{x}) = \mathrm{e}^{-(\mathrm{i}q/\hbar c)\Lambda}\,\psi(t,\boldsymbol{x}) \qquad [7c]$$

So, $\boldsymbol{A}'$ and ψ' obey [15]. At each $(t,\boldsymbol{x})$, $\mathrm{e}^{-(\mathrm{i}q/\hbar c)\Lambda}$ belongs to U(1), the unit circle in the complex plane.

In the path-integral approach, the kernel $K(t',\boldsymbol{x}';t,\boldsymbol{x})$, which gives the probability amplitude for the propagation of the particle from the spacetime point $(t,\boldsymbol{x})$ to the spacetime point $(t',\boldsymbol{x}')$ $(t<t')$, is given by

$$\begin{aligned}&K(t',\boldsymbol{x}';t,\boldsymbol{x}) \\ &= \int_{\boldsymbol{x}(t)=\boldsymbol{x}}^{\boldsymbol{x}(t')=\boldsymbol{x}'} D\boldsymbol{x}(\tau)\exp\left(\frac{\mathrm{i}}{\hbar}(S_0 + S_{\mathrm{int}})\right) \\ &= \int_{\boldsymbol{x}(t)=\boldsymbol{x}}^{\boldsymbol{x}(t')=\boldsymbol{x}'} D\boldsymbol{x}(\tau)\exp\left(\frac{\mathrm{i}}{\hbar}\int_t^{t'}\mathrm{d}\tau\left(\frac{1}{2}m\dot{\boldsymbol{x}}^2\right.\right. \\ &\quad \left.\left. + \frac{q}{c}\boldsymbol{A}\cdot\boldsymbol{v} - q\varphi\right)\right) \\ &= \int_{\boldsymbol{x}(t)=\boldsymbol{x}}^{\boldsymbol{x}(t')=\boldsymbol{x}'} D\boldsymbol{x}(\tau)\exp\left(\frac{\mathrm{i}}{\hbar}\int_t^{t'}\mathrm{d}\tau\,\frac{1}{2}m\dot{\boldsymbol{x}}^2\right) \\ &\quad \times\exp\left(\frac{\mathrm{i}q}{\hbar c}\int_t^{t'}(\boldsymbol{A}\cdot\mathrm{d}\boldsymbol{x} - \varphi\,\mathrm{d}x^0)\right) \\ &= \int_{\boldsymbol{x}(t)=\boldsymbol{x}}^{\boldsymbol{x}(t')=\boldsymbol{x}'} D\boldsymbol{x}(\tau)\exp\left(\frac{\mathrm{i}}{\hbar}\int_t^{t'}\mathrm{d}\tau\,\frac{1}{2}m\dot{\boldsymbol{x}}^2\right) \\ &\quad \times\exp\left(\frac{\mathrm{i}q}{\hbar c}\int_t^{t'}\mathrm{d}x^\mu A_\mu\right) \qquad [16]\end{aligned}$$

where the integral $\int D\boldsymbol{x}(\tau)\ldots$ is over all continuous spacetime paths $(\tau,\boldsymbol{x}(\tau))$ which join $(t,\boldsymbol{x})$ with $(t',\boldsymbol{x}')$. If one knows the wave function at $(t,\boldsymbol{x})$, then the wave function at $(t',\boldsymbol{x}')$ is given by

$$\psi(t',\boldsymbol{x}') = \int \mathrm{d}^3\boldsymbol{x}\; K(t',\boldsymbol{x}';t,\boldsymbol{x})\psi(t,\boldsymbol{x}) \qquad [17]$$

An important point is the natural appearance in the integrand of the functional integral of the factor

$$\mathrm{e}^{(\mathrm{i}q/\hbar c)\int_\gamma A}$$

for each path γ joining $(t,\boldsymbol{x})$ with $(t',\boldsymbol{x}')$.

A Solution to the Schrödinger Equation

In what follows, we shall restrict ourselves to static magnetic fields; then in the previous formulas, we set $\varphi = 0$ and $\boldsymbol{A}(t,\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{x})$. It is then easy to show that if $\boldsymbol{x}_0$ is an arbitrary reference point and the integral $\int_{\boldsymbol{x}_0}^{\boldsymbol{x}}\boldsymbol{A}(\boldsymbol{x}')\cdot\mathrm{d}\boldsymbol{x}'$ is independent of the integration path from $\boldsymbol{x}_0$ to $\boldsymbol{x}$, that is, it is a well-defined function f of $\boldsymbol{x}$, and if ψ_0 is a solution of the free Schrödinger equation, that is,

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\psi_0 = -\frac{\hbar^2}{2m}\nabla^2\psi_0 \qquad [18]$$

then

$$\psi(t,\boldsymbol{x}) = \exp\left(\frac{\mathrm{i}q}{\hbar c}\int_{\boldsymbol{x}_0}^{\boldsymbol{x}}\boldsymbol{A}(\boldsymbol{x}')\cdot\mathrm{d}\boldsymbol{x}'\right)\psi_0(t,\boldsymbol{x}) \qquad [19]$$

is a solution of [15]. In fact, replacing [19] in [15], the LHS gives

$$\exp\left(\frac{\mathrm{i}q}{\hbar c}f(\boldsymbol{x})\right)\mathrm{i}\hbar\frac{\partial}{\partial t}\psi_0$$

while for the RHS one has

$$\exp\left(\frac{\mathrm{i}q}{\hbar c}f(\boldsymbol{x})\right)\left(-\frac{\hbar^2}{2m}\right)\nabla^2\psi_0$$

The cancelation of the exponential factors shows that, under the condition of path independence, there is no effect of the potential on the charged particles. Another way to see this is by making a gauge transformation [7a]–[7c] with $\Lambda(\boldsymbol{x}) = f(\boldsymbol{x})$, which changes $\psi \to \psi_0$ and $\boldsymbol{A} \to \boldsymbol{A}' = \boldsymbol{A} - \nabla \int_{x_0}^{x} \boldsymbol{A}(\boldsymbol{x}') \cdot \mathrm{d}\boldsymbol{x}' = \boldsymbol{A} - \boldsymbol{A} = 0$.

The condition of path independence amounts, however, to the condition that no magnetic field is present since, if $\int_\gamma A$ depends on γ, then for some pair of paths γ and γ' from $(t, \boldsymbol{x})$ to $(t', \boldsymbol{x}')$, $0 \neq \int_\gamma A - \int_{\gamma'} A = \int_\gamma A + \int_{-\gamma'} A = \oint_{\gamma \cup (-\gamma')} A = \int_\Sigma \mathrm{d}\boldsymbol{\sigma} \cdot (\nabla \times \boldsymbol{A})$, where in the last equality we applied Stokes theorem (Σ is any surface with boundary $\gamma \cup (-\gamma')$), which shows that $\boldsymbol{B} = \nabla \times \boldsymbol{A}$ must not vanish everywhere and has a nonzero flux Φ through Σ given by

$$\Phi = \int_\Sigma \mathrm{d}\boldsymbol{\sigma} \cdot \boldsymbol{B} \qquad [20]$$

The conclusion of this section is that the ansatz [19] for solving [15] can only be applied in simply connected regions with no magnetic field strength present.

Aharonov–Bohm Proposal

In 1959, Aharonov and Bohm proposed an experiment to test, in quantum mechanics, the coupling of electric charges to electromagnetic field strengths through a local interaction with the electromagnetic potential A_μ, but not with the field strengths themselves. However, as we saw before, no physical effect exists, that is, A_μ can be gauged away, unless magnetic and/or electric fields exist somewhere, although not necessarily overlapping the wave function of the particles.

Consider the usual two-slit experiment as depicted in **Figure 1**, with the additional presence, behind the slits, of a long and narrow solenoid enclosing a nonvanishing magnetic flux Φ due to a constant and homogeneous magnetic field $\boldsymbol{B}$ normal to the plane of the figure (in direction z); outside of the solenoid, the magnetic field is zero. If the radius of the solenoid is R, a vector potential $\boldsymbol{A}$ that produces such field strength is given by

$$\boldsymbol{A}(\boldsymbol{x}) = \begin{cases} (|\boldsymbol{B}|r/2)\hat{\varphi}, & r \leq R \\ (\Phi/2\pi r)\hat{\varphi}, & r > R \end{cases} \qquad [21]$$

where $\Phi = \pi R^2 |\boldsymbol{B}|$ and $\hat{\varphi}$ is a unit vector in the azimuthal direction. In fact,

$$\boldsymbol{B} = \nabla \times \boldsymbol{A}(\boldsymbol{x}) = \begin{cases} |\boldsymbol{B}|\hat{z}, & r \leq R \\ 0, & r > R \end{cases} \qquad [22]$$

Notice that at $r = R$, $\boldsymbol{A}$ is continuous but not continuously differentiable. Also, the ideal limit of an infinitely long solenoid makes the problem two-dimensional, that is, in the x–y plane.

The probability amplitude for an electron emitted at the source S to arrive at the point P on the screen Π, is given by the sum of two probability amplitudes, namely those corresponding to passing through the slits 1 and 2. The solenoid is assumed to be impenetrable to the electrons; mathematically, this corresponds to a motion in a non-simply-connected region. In the approximation for the path integral [16], in which one considers the contribution of only two classes of paths, that is, the class $\{\gamma\}$ represented by path I, and the class $\{\gamma'\}$ represented by path II, if the wave function at the source is ψ_S, then the wave function at P is given by

$$\begin{aligned}
\psi_P &= \left(\int_{\{\gamma\}} \mathrm{e}^{(\mathrm{i}/\hbar)S_0(\gamma)} \mathrm{e}^{-(\mathrm{i}|e|/\hbar c)\int_\gamma A} \right. \\
&\quad \left. + \int_{\{\gamma'\}} \mathrm{e}^{(\mathrm{i}/\hbar)S_0(\gamma')} \mathrm{e}^{-(\mathrm{i}|e|/\hbar c)\int_{\gamma'} A} \right) \psi_S \\
&= \mathrm{e}^{-(\mathrm{i}|e|/\hbar c)\int_{\mathrm{I}} A} \int_{\{\gamma\}} \mathrm{e}^{(\mathrm{i}/\hbar)S_0(\gamma)} \psi_S \\
&\quad + \mathrm{e}^{-(\mathrm{i}|e|/\hbar c)\int_{\mathrm{II}} A} \int_{\{\gamma'\}} \mathrm{e}^{(\mathrm{i}/\hbar)S_0(\gamma')} \psi_S \\
&= \mathrm{e}^{-(\mathrm{i}|e|/\hbar c)\int_{\mathrm{I}} A} \left(\psi_P^0(\mathrm{I}) \right. \\
&\quad \left. + \mathrm{e}^{-(\mathrm{i}|e|/\hbar c)\left(\int_{\mathrm{II}\cup(-\mathrm{I})} A\right)} \psi_P^0(\mathrm{II}) \right) \\
&= \mathrm{e}^{-(\mathrm{i}|e|/\hbar c)\int_{\mathrm{I}} A} \left(\psi_P^0(\mathrm{I}) + \mathrm{e}^{-2\pi\mathrm{i}(\Phi/\Phi_0)} \psi_P^0(\mathrm{II}) \right)
\end{aligned} \qquad [23]$$

where, in the second line, we used the path independence of the integral of $\boldsymbol{A}$ within each class of paths;

$$\psi_P^0(\mathrm{I}) = \int_{\{\gamma\}} \mathrm{e}^{(\mathrm{i}/\hbar)\int_{\{\gamma\}} S_0(\gamma)} \psi_S$$

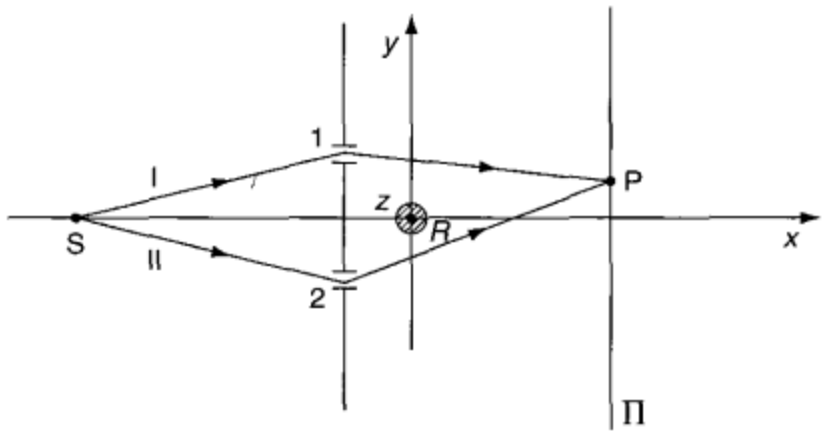


Figure 1 Magnetic Aharonov–Bohm effect.

and

$$\psi_{\mathrm{P}}^{0}(\mathrm{II}) = \int_{\{\gamma'\}} \mathrm{e}^{(\mathrm{i}/\hbar)S_0(\gamma')}\psi_{\mathrm{S}}$$

and, in the last equality, we applied the extended version of Stokes theorem (by Craven), to allow for noncontinuously differentiable vector potentials; and the quantum of magnetic flux associated with the charge $|e|$ is defined by

$$\Phi_0 = 2\pi\frac{\hbar c}{|e|} \cong 4.135 \times 10^{-7}\ \mathrm{G\ cm^2} \qquad [24]$$

$(=2\pi/|e| = \sqrt{\pi/\alpha} \cong \sqrt{137\pi}$ in the natural system of units (n.s.u.) $\hbar = c = 1$; α is the fine structure constant). Then the probability of finding the electron at P is proportional to

$$|\psi_{\mathrm{P}}|^2 = |\psi_{\mathrm{P}}^{0}(\mathrm{I})|^2 + |\psi_{\mathrm{P}}^{0}(\mathrm{II})|^2 + 2\mathrm{Re}\left(\mathrm{e}^{2\pi\mathrm{i}(\Phi/\Phi_0)}\psi_{\mathrm{P}}^{0}(\mathrm{I})\psi_{\mathrm{P}}^{0}(\mathrm{II})^*\right) \qquad [25]$$

which exhibits an interference pattern shifted with respect to that without the magnetic field: as $\boldsymbol{B}$ and therefore Φ change, dark and bright interference fringes alternate periodically at the screen Π, with period Φ_0. This is the magnetic A–B effect, which has been quantitatively verified in many experiments, the first one in 1960 by Chambers. The effect is:

1. *gauge invariant*, since $\boldsymbol{B}$ and therefore Φ are gauge invariant;
2. *nonlocal*, since it depends on the magnetic field inside the solenoid, where the electrons never enter;
3. *quantum mechanical*, since classically the charges do not feel any force and therefore no effect would be expected in this limit; and
4. *topological*, since the electrons necessarily move in a non-simply-connected space.

But perhaps the most important implication of the A–B effect is a dramatic additional confirmation of the nonlocal character of quantum mechanics: the electron has to "travel" along the two paths (I and II) simultaneously; on the contrary, no flux would be surrounded and then no shift of the (then nonexistent) interference fringes would be observed at the screen Π.

Calculations in the path-integral approach including the whole set of homotopy classes of paths around the solenoid, indexed by an integer m, have been performed by several authors, leading to a formula of the type

$$\psi_{\mathrm{P}} = \sum_{m=-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}m\delta}\psi_{\mathrm{P}}^{0}(m) \qquad [26]$$

with

$$\delta = 2\pi\frac{\Phi}{\Phi_0} \qquad [27]$$

(Schulman 1971, Kobe 1979). As in [23],

$$\psi_{\mathrm{P}}(\Phi + k\Phi_0) = \psi_{\mathrm{P}}(\Phi), \quad k \in \mathbb{Z} \qquad [28]$$

There is a close relation between the A–B effect and the Dirac quantization condition (DQC) in the presence of electric and magnetic charges: according to [25] (or [26]) the A–B effect disappears when the flux Φ equals $n\Phi_0 = 2\pi n(\hbar c/|e|)$, $n \in \mathbb{Z}$, that is, when the condition

$$|e|\Phi = nhc \qquad [29]$$

holds. But this is the DQC (Dirac 1931) when Φ is the flux associated with a magnetic charge g: $\Phi(g) = (g/4\pi r^2) \times 4\pi r^2 = g$, leading to $|e|g = nhc$ ($2\pi n$ in the n.s.u.). This is precisely the condition for the Dirac string to be unobservable in quantum mechanics: to give no A–B effect.

Geometry of the A–B Effect

In this section we study the space of gauge classes of flat potentials outside the solenoid, which determine the A–B effect; the topological structure of the A–B bundle; and the holonomy groups of the connections, which precisely give the phase shifts of the wave functions. We use the n.s.u. system; in particular, if [L] is the unit of length, then $[A_\mu] = [\mathrm{L}]^{-1}$, $[|e|] = [\mathrm{L}]^0$, and $\Phi_0 = 2\pi/|e| = \sqrt{\pi/\alpha} \cong \sqrt{137\pi}$, where α is the fine structure constant.

To synthesize, one can say that *the abelian A–B effect is a nonlocal gauge-invariant quantum effect due to the coupling of the wave function (section of an associated bundle) to a nontrivial (non-exact) flat (closed) connection in a trivial principal bundle with a non-simply-connected base space*. In the following subsections, we will give a detailed explanation of these statements.

The A–B Bundle

The gauge group of electromagnetism is the abelian Lie group U(1) with Lie algebra (the tangent space at the identity) $u(1) = \mathrm{i}\mathbb{R}$. In the limit of an infinitely long and infinitesimally thin solenoid carrying the magnetic flux Φ, the space available to the electrons is the plane minus a point, that is, $\mathbb{R}^{2*}$, which is of the same homotopy type as the circle S^1. Then the set of isomorphism classes of U(1) bundles over $\mathbb{R}^{2*}$ is in one-to-one correspondence with the set of homotopy classes of maps from S^0 to S^1 (Steenrod 1951), which consists of only one point: if f, g:

$S^0 \to S^1$ are given by $f(1) = e^{i\varphi_1}$, $f(-1) = e^{i\varphi_2}$, $g(1) = e^{i\theta_1}$, and $g(-1) = e^{i\theta_2}$, then $H : S^0 \times [0,1] \to S^1$ given by $H(1,t) = e^{i((1-t)\varphi_1 + t\theta_1)}$ and $H(-1,t) = e^{i((1-t)\varphi_2 + t\theta_2)}$ is a homotopy between f and g. Then, up to equivalence, the relevant bundle for the A–B effect is the product bundle

$$\xi_{A-B} : U(1) \to \mathbb{R}^{2*} \times U(1) \to \mathbb{R}^{2*} \qquad [30a]$$

Since $\mathbb{R}^{2*}$ is homeomorphic to an open disk minus a point $(D_0^2)^*$, then the total space of the bundle is homeomorphic to an open solid 2-torus minus a circle, since $(T_0^2)^* = (D_0^2)^* \times S^1$. Then the A–B bundle has the topological structure

$$\xi_{A-B} : S^1 \to (T_0^2)^* \to (D_0^2)^* \qquad [30b]$$

The Gauge Group and the Moduli Space of Flat Connections

The gauge group of the bundle ξ_{A-B} is the set of smooth functions from the base space to the structure group, that is, $\mathcal{G} = C^\infty(\mathbb{R}^{2*}, U(1))$. Since $\mathcal{G} \subset C^0(\mathbb{R}^{2*}, U(1)) = \{\text{continuous functions } \mathbb{R}^{2*} \to U(1)\}$ and $[\mathbb{R}^{2*}, U(1)] = \{\text{homotopy classes of continuous functions } \mathbb{R}^{2*} \to U(1)\} \cong [S^1, S^1] \cong \pi_1(S^1) \cong \mathbb{Z}$, given $f \in \mathcal{G}$ there exists a unique $n \in \mathbb{Z}$ such that f is homotopic to $f_n (f \sim f_n)$, where $f_n : \mathbb{R}^{2*} \to U(1)$ is given by $f_n(re^{i\varphi}) = e^{in\varphi}$, $\varphi \in [0, 2\pi)$.

$\mathcal{G}$ acts on the space of flat connections on ξ_{A-B} given by the closed $u(1)$-valued differential 1-forms on $\mathbb{R}^{2*}$:

$$\mathcal{C}_0 = \{\mathcal{A} \in \Omega^1(\mathbb{R}^{2*}; u(1)), d\mathcal{A} = 0\} \qquad [31]$$

through

$$\mathcal{C}_0 \times \mathcal{G} \to \mathcal{C}_0, \quad (\mathcal{A}, f) \to \mathcal{A} + f^{-1}df \qquad [32]$$

where $f^{-1}(x,y) = (f(x,y))^{-1}$. The moduli space

$$\mathcal{M}_0 = \frac{\mathcal{C}_0}{\mathcal{G}} = \{\text{gauge equivalence classes of flat connections on } \xi_{A-B}\}$$
$$= \{[\mathcal{A}] = \{\mathcal{A} + f^{-1}df,\ f \in \mathcal{G}\}, \mathcal{A} \in \mathcal{C}_0\} \qquad [33]$$

is isomorphic to the circle S^1 with length 1. This can be seen as follows: the de Rham cohomology of $\mathbb{R}^{2*}$ with coefficients in $i\mathbb{R}$ in dimension 1 is

$$H^1_{DR}(\mathbb{R}^{2*}; i\mathbb{R}) = \{\lambda[\mathcal{A}_0]_{DR}, \lambda \in \mathbb{R}\} \cong H^1_{DR}(S^1; i\mathbb{R}) \cong \mathbb{R} \qquad [34]$$

where

$$\mathcal{A}_0 = i\frac{x\,dy - y\,dx}{x^2 + y^2} \in \mathcal{C}_0 \qquad [35]$$

is the connection that, once multiplied by $-|e|^{-1}$ (see below) generates the flux $-\Phi_0$ and therefore no A–B effect: $\mathcal{A}_0$ is closed ($d\mathcal{A}_0 = 0$) but not exact ($(x\,dy - y\,dx)/(x^2 + y^2) = d\varphi$ only for $\varphi \in (0, 2\pi)$, $\varphi = 0$ is excluded); $[\mathcal{A}_0]_{DR} = \mathcal{A}_0 + d\beta$ with $\beta \in \Omega^0(R^{2*}; i\mathbb{R})$. β gives an element of $\mathcal{G}$ through the composite $\exp \circ \beta : \mathbb{R}^{2*} \to U(1), (x,y) \mapsto e^{i\beta(x,y)}$. The A–B effect with flux $\Phi = -\lambda\Phi_0$ is produced by the connection $\mathcal{A} = \lambda\mathcal{A}_0$. To determine $\mathcal{M}_0$, one finds the smallest $\sigma \in \mathbb{R}$ such that $(\lambda + \sigma)\mathcal{A}_0 \sim \lambda\mathcal{A}_0$, that is, $(\lambda + \sigma)\mathcal{A}_0 \in [\lambda\mathcal{A}_0]$, which means, from [33], that $(\lambda + \sigma)\mathcal{A}_0 = \lambda\mathcal{A}_0 + f^{-1}df$ or $\sigma\mathcal{A}_0 = f^{-1}df$. For $\varphi \neq 0$, $\mathcal{A}_0 = id\varphi$ and $f_1^{-1}df_1 = id\varphi$, then $\sigma = 1$, and therefore $(\lambda + 1)\mathcal{A}_0 \sim \lambda\mathcal{A}_0$, in particular $\mathcal{A}_0 \sim 0$.

A remark concerning the gauge group $\mathcal{G}$ is the following. In classical electrodynamics, according to [7a] and [7b], the symmetry group could be taken to be the additive group $(\mathbb{R}, +)$ instead of the multiplicative group $U(1)$. Since $\mathbb{R}$ is contractible, then the gauge group would be $\mathcal{G}_{cl} = C^\infty(\mathbb{R}^{2*}, \mathbb{R})$ with $[\mathbb{R}^{2*}, \mathbb{R}] \cong 0$, so that the homomorphism $\Psi : \mathcal{G}_{cl} \to \mathcal{G}$, $\Psi(f)(x) = e^{if(x)}$ would not exhaust $\mathcal{G}$ since $\Psi(f) \in [1]$ for any $f \in \mathcal{G}_{cl}$: in fact, $H : \mathbb{R}^{2*} \times [0,1] \to U(1)$ given by $H(x,t) = e^{i(1-t)f(x)}$ is a homotopy between $\Psi(f)$ and 1. However, the quantization of electric charges implies that in fact the gauge group is $U(1)$ and not $\mathbb{R}$. This is equivalent mathematically to the possible existence of magnetic monopoles which require nontrivial bundles for their description.

Covariant Derivative, Parallel Transport, and Holonomy

Let G be a matrix Lie group with Lie algebra $\mathfrak{g}$, B a differentiable manifold, $\xi : G \to P \xrightarrow{\pi} B$ a principal bundle, V a vector space, $G \times V \to V$ an action, and $\xi_V : V \to P \times_G V \xrightarrow{\pi_V} B$ the corresponding associated vector bundle (ξ_V is trivial if ξ is trivial). Call $\Gamma(\xi_V)$ the sections of ξ_V, $\Gamma(TB)(\Gamma(TP))$ the sections of the tangent bundle of $B(P)$, and $\Gamma_{eq}(P, V)$ the set of functions $\gamma : P \to V$ satisfying $\gamma(pg) = g^{-1}\gamma(p)$ (equivariant functions from P to V). $s \in \Gamma(\xi_V)$ induces $\gamma_s \in \Gamma_{eq}(P, V)$ with $\gamma_s(p) = v$, where $s(\pi(p)) = [p, v]$ and $\gamma \in \Gamma_{eq}(P, V)$ induces $s_\gamma \in \Gamma(\xi_V)$ with $s_\gamma(b) = [p, \gamma(p)]$, where $p \in \pi^{-1}(\{b\})$. If H is a connection on ξ, that is, a smooth assignment of a (horizontal) vector subspace H_p of T_pP at each p of P, algebraically determined by a smooth $\mathfrak{g}$-valued 1-form ω on P through $H_p = \ker(\omega_p)$, $s \in \Gamma(\xi_V)$, $X \in \Gamma(TB)$, and $X^\uparrow \in \Gamma(TP)$ the horizontal lifting of X by ω, then $X^\uparrow(\gamma_s) \in \Gamma_{eq}(P, V)$, and covariant

derivative of s with respect to ω in the direction of X is defined by

$$\nabla^{\omega}_{X} s := s_{X^{\uparrow}(\gamma_s)} \qquad [36a]$$

If $\phi: \pi^{-1}(U) \to U \times G$ is a local trivialization of ξ, x^{μ}, $\mu = 1, \ldots, \dim B$ are local coordinates on U, and e_i, $i = 1, \ldots, \dim V$ is a basis of the local sections in $\pi_V^{-1}(U)$, then the local expression of [36a] is

$$\nabla^{\omega_U}_{X^{\mu}\partial/\partial x^{\mu}}(s^i e_i) = X^{\mu}\left(\delta^j_i \frac{\partial}{\partial x^{\mu}} + \mathcal{A}^j_{\mu i}\right) s^i e_j \qquad [36b]$$

where

$$\mathcal{A}_U{}^j_i = \mathcal{A}^j_{\mu i} \mathrm{d}x^{\mu} = (\sigma^* \omega_U)^j_i \qquad [36c]$$

is the geometrical gauge potential in U, given by the pullback of ω_U, the restriction of ω to $\pi^{-1}(U)$, by the local section $\sigma: U \to \pi^{-1}(U)$, $\sigma(b) = \phi^{-1}(b, 1)$. ($\mathcal{A}^j_{\mu i}$ is defined through $\nabla^{\omega_U}_{\partial/\partial x^{\mu}} e_i = \mathcal{A}^j_{\mu i} e_j$.) The operator

$$D^j_{\mu i} = \delta^j_i \frac{\partial}{\partial x^{\mu}} + \mathcal{A}^j_{\mu i} \qquad [36d]$$

is the usual local covariant derivative. In an overlapping trivialization, [36b] is replaced by

$$\nabla^{\omega_{U'}}_{X^{\mu}\partial/\partial x^{\mu}}(s'^i e'_i) = X^{\mu}\left(\delta^j_i \frac{\partial}{\partial x^{\mu}} + \mathcal{A}'^j_{\mu i}\right) s'^i e'_j$$

with $e'_j = g^k_j e_k$ and $s'^i = g^{-1}{}^i_l s^l$ on $U \cap U'$, then the local potential transforms as

$$\mathcal{A}^j_{\mu l} = g^j_k \mathcal{A}'^k_{\mu i} g^{-1}{}^i_l + (\partial_{\mu} g^j_k) g^{-1}{}^k_l \qquad [36e]$$

which for G abelian has the form [32].

For each smooth path $c: [0, 1] \to B$ joining the points b and b', and each $p \in P_b = \pi^{-1}(\{b\})$, there exists a unique path $c^{\uparrow}$ in P through p with $\dot{c}^{\uparrow}(t) \in H_{c(t)}$ for all $t \in [0, 1]$. $c^{\uparrow}$ is the horizontal lifting of c by ω through p. Thus, for each connection and path there exists a diffeomorphism $P^{\omega}_c: P_b \to P_{b'}$ called parallel transport. If c is a loop at b, then $P^{\omega}_c \in \mathrm{Diff}(P_b)$ is called the holonomy of ω at b along c. To the loop space of B at b, $\Omega(B; b)$, corresponds a subgroup Hol^{ω}_b of $\mathrm{Diff}(P_b)$ called the holonomy of ω at b. If $c \in \Omega(B; b)$ and β is a lifting of c through $q \in P_b$, then there exists a unique path $g: [0, 1] \to G$ such that $c^{\uparrow}(t) = \beta(t) g(t)$ with $c^{\uparrow}(0) = qg(0) = p$; g satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} g(t) + \omega_{\beta(t)}(\dot{\beta}(t)) = 0 \qquad [37]$$

whose solution is the time-ordered exponential

$$\begin{aligned} g(t)g(0)^{-1} &= T\exp\left(\int_0^t \mathrm{d}\tau\, \omega_{\beta(\tau)}(\dot{\beta}(\tau))\right) \\ &= 1 + \sum_{m=1}^{\infty} (-1)^m \int_0^t \mathrm{d}\tau_1 \omega_{\beta(\tau_1)}(\dot{\beta}(\tau_1)) \\ &\quad \times \int_0^{\tau_1} \mathrm{d}\tau_2 \omega_{\beta(\tau_2)}(\dot{\beta}(\tau_2)) \cdots \\ &\quad \times \int_0^{\tau_{m-1}} \mathrm{d}\tau_m \omega_{\beta(\tau_m)}(\dot{\beta}(\tau_m)) \end{aligned} \qquad [38]$$

If $q = p$ then $g(0) = 1$. For each $p \in P$, the set of elements $g' \in G$ such that $c^{\uparrow}(1) = pg'$ for $c \in \Omega(B; \pi(p))$ is a subgroup of G, Hol^{ω}_p, called the holonomy *of* ω at p. (For each p, there exists a group isomorphism $\mathrm{Hol}^{\omega}_p \to \mathrm{Hol}^{\omega}_{\pi(p)}$, and if p and p' are connected by a horizontal curve, then $\mathrm{Hol}^{\omega}_p = \mathrm{Hol}^{\omega}_{p'}$; if all p's *in* P are horizontally connected, then $\mathrm{Hol}^{\omega}_p = G$ for all $p \in P$.) If (U, ϕ) is a local trivialization of ξ, $c \subset U$, and $\beta(t) = \sigma(c(t))$, then one has the local formula

$$c^{\uparrow}(t) = \phi^{-1}(c(t), 1)\left(T\exp\left(-\int_{c(0)}^{c(t)} \mathcal{A}_U\right)\right) g(0) \qquad [39]$$

In particular, if ξ is a product bundle, then ϕ is the identity, and choosing $g(0) = 1$ gives

$$c^{\uparrow}(t) = \left(c(t), T\exp\left(-\int_{c(0)}^{c(t)} \mathcal{A}_U\right)\right) \qquad [40]$$

In our case, $V = \mathbb{C}$, ξ is a product bundle, $s = \psi$, the wave function, is a global section of the associated bundle

$$\xi_{\mathbb{C}}: \mathbb{C} \to \mathbb{R}^{2*} \times \mathbb{C} \xrightarrow{\pi_{\mathbb{C}}} \mathbb{R}^{2*} \qquad [41]$$

$G = \mathrm{U}(1)$ with $\mathfrak{g} = \mathrm{i}\mathbb{R}$ and an action $\mathrm{U}(1) \times \mathbb{C} \to \mathbb{C}$, $(e^{\mathrm{i}\varphi}, z) \mapsto e^{\mathrm{i}\varphi} z$; therefore, $\mathcal{A}_{\mu} = \mathcal{A}_{0\mu} = \mathrm{i}a_{\mu}$ with a_{μ} real valued, and the covariant derivative is

$$D_{\mu}\psi = \left(\frac{\partial}{\partial x_{\mu}} + \mathrm{i}a_{\mu}\right)\psi \qquad [36f]$$

If ψ carries the electric charge q, we define the physical gauge potential A_{μ} through

$$a_{\mu} = qA_{\mu} \qquad [42]$$

and, for the covariant derivative, after multiplying by i, we obtain the operator appearing in eqn [15], $\mathrm{i}D_{\mu}\psi = (\mathrm{i}(\partial/\partial x^{\mu}) - qA_{\mu})\psi$: in fact, for the spatial part the coupling is $(\mathrm{i}\nabla + qA)\psi$, and for the temporal part one has $(\mathrm{i}\partial/\partial t - q\varphi)\psi$. For the electron, $q = -|e|$ and $a_{\mu} = -|e|A_{\mu} = -(2\pi/\Phi_0)A_{\mu}$.

For $c \in \Omega(\mathbb{R}^{2*};(x_0,y_0))$, which turns n times around the solenoid at $(0,0)$, eqn [40] gives

$$c^{\uparrow} = ((x_0,y_0), e^{-n\oint_c A}) = ((x_0,y_0), e^{-in\oint_c a})$$
$$= ((x_0,y_0), e^{-i|e|n\oint_c A\cdot dx}) = ((x_0,y_0), e^{-2\pi i n\Phi/\Phi_0})$$

and therefore, for $\Phi/\Phi_0 = \lambda \in [0,1)$ we have the holonomy groups

$$\mathrm{Hol}^{\omega(\Phi)}_{((x_0,y_0),1)} = \{e^{-2\pi i n(\Phi/\Phi_0)}\}_{n\in Z}$$
$$= \begin{cases} \mathbb{Z}_q, & \lambda = p/q,\ p,q \in \mathbb{Z},\ (p,q) = 1 \\ \mathbb{Z}, & \lambda \notin \mathbb{Q} \end{cases} \qquad [43]$$

In the second case, $\mathrm{Hol}^{\omega(\Phi)}_{((x_0,y_0),1)}$ is dense in U(1): in fact, suppose that for $n_1, n_2 \in \mathbb{Z}$, $n_1 \neq n_2$, $e^{2\pi i n_1\lambda} = e^{2\pi i n_2\lambda}$, then $e^{2\pi i(n_1-n_2)\lambda} = 1$ and so $(n_1 - n_2)\lambda = m$ for some $m \in \mathbb{Z}$; therefore, $\lambda \in \mathbb{Q}$, which is a contradiction.

Finally, we should mention that the A–B effect can be understood as a geometric phase *à la* Berry, though not necessarily through an adiabatic change of the parameters on which the Hamiltonian depends. The Berry potential a_B turns out to be proportional to the real magnetic vector potential A: in the n.s.u., and for electrons,

$$a_B = -|e|A \qquad [44]$$

Nonabelian and Gravitational A–B Effects

Since the fundamental group $\Pi_1(\mathbb{R}^{2*},(x_0,y_0)) \cong \mathbb{Z}$, eqn [43] shows that there is a homomorphism $\varphi(\omega)$: $\Pi_1(\mathbb{R}^{2*},(x_0,y_0)) \to \mathrm{U}(1)$, $\varphi(\omega)(n) = e^{-2\pi i n\lambda}$, with $\varphi(\omega)\ (\Pi_1(\mathbb{R}^{2*})) = \mathrm{Hol}^{\omega(\Phi)}_{((x_0,y_0),1)}$, which characterizes the A–B effect in that case. In general, an A–B effect in a G-bundle with a connection ω is characterized by a group homomorphism from the fundamental group of the base space B onto the holonomy group of the connection, which is a subgroup of the structure group. The A–B effect is nonabelian if the holonomy group is nonabelian, which requires both G and $\Pi_1(B,x)$ to be nonabelian. Examples with Yang–Mills and gravitational fields are considered in the literature.

Acknowledgment

The author thanks the University of Valencia, Spain, where part of this work was done.

See also: Deformation Quantization and Representation Theory; Fractional Quantum Hall Effect; Geometric Phases; Moduli Spaces: An Introduction; Quantum Chromodynamics; Variational Techniques for Ginzburg–Landau Energies.

Further Reading

Aguilar MA and Socolovsky M (2002) Aharonov–Bohm effect, flat connections, and Green's theorem. *International Journal of Theoretical Physics* 41: 839–860.

Aharonov Y and Bohm D (1959) Significance of electromagnetic potentials in the quantum theory. *Physical Review* 15: 485–491.

Berry MV (1984) Quantal phase factors accompanying adiabatic changes. *Proceedings of the Royal Society of London A* 392: 45–57.

Chambers RG (1960) Shift of an electron interference pattern by enclosed magnetic flux. *Physical Review Letters* 5: 3–5.

Corichi A and Pierri M (1995) Gravity and geometric phases. *Physical Review D* 51: 5870–5875.

Dirac PMA (1931) Quantised singularities in the electromagnetic field. *Proceedings of the Royal Society of London A* 133: 60–72.

Kobe DH (1979) Aharonov–Bohm effect revisited. *Annals of Physics* 123: 381–410.

Peshkin M and Tonomura A (1989) *The Aharonov–Bohm Effect.* Berlin: Springer.

Schulman LS (1971) Approximate topologies. *Journal of Mathematical Physics* 12: 304–308.

Steenrod N (1951) *The Topology of Fibre Bundles.* Princeton, NJ: Princeton University Press.

Sundrum R and Tassie LJ (1986) Non-abelian Aharonov–Bohm effects, Feynman paths, and topology. *Journal of Mathematical Physics* 27: 1566–1570.

Wu TT and Yang CN (1975) Concept of nonintegrable phase factors and global formulation of gauge fields. *Physical Review D* 12: 3845–3857.

Anomalies

S L Adler, Institute for Advanced Study, Princeton, NJ, USA

© 2006 Elsevier Ltd. All rights reserved.

Synopsis

Anomalies are the breaking of classical symmetries by quantum mechanical radiative corrections, which arise when the regularizations needed to evaluate small fermion loop Feynman diagrams conflict with a classical symmetry of the theory. They have important implications for a wide range of issues in quantum field theory, mathematical physics, and string theory.

Chiral Anomalies, Abelian and Nonabelian

Consider quantum electrodynamics, with the fermionic Lagrangian density

$$\mathcal{L} = \bar{\psi}(\mathrm{i}\gamma^\mu \partial_\mu - e_0 \gamma^\mu B_\mu - m_0)\psi \qquad [1a]$$

where $\bar{\psi} = \psi^\dagger \gamma^0$, e_0 and m_0 are the bare charge and mass, and B_μ is the electromagnetic gauge potential. (We reserve the notation A for axial-vector quantities.) Under a chiral transformation

$$\psi \to \mathrm{e}^{\mathrm{i}\lambda\gamma_5}\psi \qquad [1b]$$

with constant λ, the kinetic term in eqn [1a] is invariant (because γ_5 commutes with $\gamma^0\gamma^\mu$), whereas the mass term is not invariant. Therefore, naive application of Noether's theorem would lead one to expect that the axial-vector current

$$j_\mu^5 = \bar{\psi}\gamma_\mu\gamma_5\psi \qquad [1c]$$

obtained from the Lagrangian density by applying a chiral transformation with spatially varying λ, should have a divergence given by the change under chiral transformation of the mass term in eqn [1a]. Up to tree approximation, this is indeed true, but when one computes the AVV Feynman diagram with one axial-vector and two vector vertices (see **Figure 1**), and insists on conservation of the vector current $j_\mu = \bar{\psi}\gamma_\mu\psi$, one finds that to order e_0^2, the classical Noether theorem is modified to read

$$\partial^\mu j_\mu^5(x) = 2\mathrm{i}m_0 j^5(x) + \frac{e_0^2}{16\pi^2} F^{\xi\sigma}(x)F^{\tau\rho}(x)\epsilon_{\xi\sigma\tau\rho} \qquad [2]$$

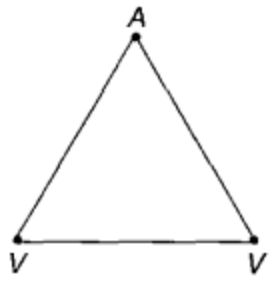


Figure 1 The AVV triangle diagram responsible for the abelian chiral anomaly.

with $F^{\xi\sigma}(x) = \partial^\sigma B^\xi(x) - \partial^\xi B^\sigma(x)$ the electromagnetic field strength tensor. The second term in eqn [2], which would be unexpected from the application of the classical Noether theorem, is the abelian axial-vector anomaly (often called the Adler–Bell–Jackiw (or ABJ) anomaly after the seminal papers on the subject). Since vector current conservation, together with the axial-vector current anomaly, implies that the left- and right-handed chiral currents $j_\mu \pm j_\mu^5$ are also anomalous, the axial-vector anomaly is frequently called the "chiral anomaly," and we shall use the terms interchangeably in this article.

There are a number of different ways to understand why the extra term in eqn [2] appears. (1) Working through the formal Feynman diagrammatic Ward identity proof of the Noether theorem, one finds that there is a step where the closed fermion loop contributions are eliminated by a shift of the loop-integration variable. For Feynman diagrams that are convergent, this is not a problem, but the AVV diagram is linearly divergent. The linear divergence vanishes under symmetric integration, but the shift then produces a finite residue, which gives the anomaly. (2) If one defines the AVV diagram by Pauli–Villars regularization with regulator mass M_0 that is allowed to approach infinity at the end of the calculation, one finds a classical Noether theorem in the regulated theory,

$$\partial^\mu j_\mu^5|_{m_0} - \partial^\mu j_\mu^5|_{M_0} = 2\mathrm{i}m_0 j^5|_{m_0} - 2\mathrm{i}M_0 j^5|_{M_0} \qquad [3a]$$

with the subscripts m_0 and M_0 indicating that fermion loops are to be calculated with fermion mass m_0 and M_0, respectively. Taking the vacuum to two-photon matrix element of eqn [3a], one finds that the matrix element $\langle 0|j^5|_{M_0}|\gamma\gamma\rangle$, which is unambiguously computable after imposing vector-current conservation, falls off only as M_0^{-1} as the regulator mass approaches infinity. Thus, the product of $2\mathrm{i}M_0$ with this matrix element has a finite limit, which gives the anomaly. (3) If the

Notes:

gauge-invariant axial-vector current is defined by point-splitting

$$j^5_\mu(x) = \bar{\psi}(x + \epsilon/2)\gamma_\mu\gamma_5\psi(x - \epsilon/2)e^{-ie_0\epsilon^\sigma B_\sigma(x)} \quad [3b]$$

with $\epsilon \to 0$ at the end of the calculation, one observes that the divergence of eqn [3b] contains an extra term with a factor of ϵ. On careful evaluation, one finds that the coefficient of this factor is an expression that behaves as ϵ^{-1}, which gives the anomaly in the limit of vanishing ϵ. (4) Finally, if the field theory is defined by a functional integral over the classical action, the standard Noether analysis shows that the classical action is invariant under the chiral transformation of eqn [1b], apart from the contribution of the mass term, which gives the naive axial-vector divergence. However, as pointed out by Fujikawa, the chiral transformation must also be applied to the functional integration measure, and since the measure is an infinite product, it must be regularized to be well defined. Careful calculation shows that the regularized measure is not chiral invariant, but contributes an extra term to the axial-vector Ward identity that is precisely the chiral anomaly.

A key feature of the anomaly is that it is irreducible: a local polynomial counter term cannot be added to the AVV diagram that preserves vector-current conservation and eliminates the anomaly. More generally, one can show that there is no way of modifying quantum electrodynamics so as to eliminate the chiral anomaly, without spoiling either vector-current conservation (i.e., electromagnetic gauge invariance), renormalizability, or unitarity. Thus, the chiral anomaly is a new physical effect in renormalizable quantum field theory, which is not present in the prequantization classical theory.

The abelian chiral anomaly is the simplest case of the anomaly phenomenon. It was extended to nonabelian gauge theories by Bardeen using a point-splitting method to compute the divergence, followed by adding polynomial counter terms to remove as many of the residual terms as possible. The resulting irreducible divergence is the nonabelian chiral anomaly, which in terms of Yang–Mills field strengths for vector and axial-vector gauge potentials V^μ and A^μ,

$$\begin{aligned} F^{\mu\nu}_V(x) &= \partial^\mu V^\nu(x) - \partial^\nu V^\mu(x) - i[V^\mu(x), V^\nu(x)] \\ &\quad - i[A^\mu(x), A^\nu(x)] \\ F^{\mu\nu}_A(x) &= \partial^\mu A^\nu(x) - \partial^\nu A^\mu(x) - i[V^\mu(x), A^\nu(x)] \\ &\quad - i[A^\mu(x), V^\nu(x)] \end{aligned} \quad [4a]$$

is given by

$$\begin{aligned} \partial^\mu j^a_{5\mu}(x) = {} & \text{normal divergence term} \\ & + (1/4\pi^2)\epsilon_{\mu\nu\sigma\tau}\mathrm{tr}\lambda^a_A[(1/4)F^{\mu\nu}_V(x)F^{\sigma\tau}_V(x) \\ & + (1/12)F^{\mu\nu}_A(x)F^{\sigma\tau}_A(x) \\ & + (2/3)iA^\mu(x)A^\nu(x)F^{\sigma\tau}_V(x) \\ & + (2/3)iF^{\mu\nu}_V(x)A^\sigma(x)A^\tau(x) \\ & + (2/3)iA^\mu(x)F^{\nu\sigma}_V(x)A^\tau(x) \\ & - (8/3)A^\mu(x)A^\nu(x)A^\sigma(x)A^\tau(x)] \end{aligned} \quad [4b]$$

In eqn [4b], "tr" denotes a trace over internal degrees of freedom, and λ^a_A is the internal symmetry matrix associated with the axial-vector external field. In the abelian case, where there is no internal symmetry structure, the terms involving two or four factors of $A^\mu, A^\nu, \ldots$ vanish by antisymmetry of $\epsilon_{\mu\nu\sigma\tau}$, and one recovers the AVV triangle anomaly, as well as a kinematically related anomaly in the AAA triangle diagram. In the nonabelian case, with nontrivial internal symmetry structure, there are also box- and pentagon-diagram anomalies.

In addition to coupling to spin-1 gauge fields, fermions can also couple to spin-2 gauge fields, associated with the graviton. When the coupling of fermions to gravitation is taken into account, the axial-vector current $\bar{\psi}T\gamma_\mu\gamma_5\psi$, with T an internal symmetry matrix, has an additional anomalous contribution to its divergence proportional to

$$\mathrm{tr}\, T\epsilon_{\xi\sigma\tau\rho}R^{\xi\sigma\alpha\beta}R^{\tau\rho}{}_{\alpha\beta} \quad [4c]$$

where $R_{\xi\sigma\tau\rho}$ is the Riemann curvature tensor of the gravitational field.

Chiral Anomaly Nonrenormalization

A salient feature of the chiral anomaly is the fact that it is not renormalized by higher-order radiative corrections. In other words, the one-loop expressions of eqns [2] and [4b] give the exact anomaly coefficient without modification in higher orders of perturbation theory. In gauge theories such as quantum electrodynamics and quantum chromodynamics, this result (the Adler–Bardeen theorem) can be understood heuristically as follows. Write down a modified Lagrangian, in which regulators are included for all gauge-boson fields. Since the gauge-boson regulators do not influence the chiral-symmetry properties of the theory, the divergences of the chiral currents are not affected by their inclusion, and so the only sources of anomalies in the regularized theory are small single-fermion loops, giving the anomaly expressions of eqns [2] and [4b]. Since the renormalized theory is obtained as the limit of

the regularized theory as the regulator masses approach infinity, this result applies to the renormalized theory as well.

The above argument can be made precise, and extends to nongauge theories such as the σ-model as well. For both gauge theories and the σ-model, cancellation of radiative corrections to the anomaly coefficient has been explicitly demonstrated in fourth-order calculations. Nonperturbative demonstrations of anomaly renormalization have also been given using the Callan–Symanzik equations. For example, in quantum electrodynamics, Zee, and Lowenstein and Schroer, showed that a factor f that gives the ratio of the true anomaly to its one-loop value obeys the differential equation

$$\left(m\frac{\partial}{\partial m}+\alpha\beta(\alpha)\frac{\partial}{\partial\alpha}\right)f=0 \qquad [5]$$

Since f is dimensionless, it can have no dependence on the mass m, and since $\beta(\alpha)$ is nonzero this implies $\partial f/\partial\alpha=0$. Thus, f has no dependence on α, and so $f=1$.

Applications of Chiral Anomalies

Chiral anomalies have numerous applications in the standard model of particle physics and its extensions, and we describe here a few of the most important ones.

Neutral Pion Decay $\pi^0\to\gamma\gamma$

As a result of the abelian chiral anomaly, the partially conserved axial-vector current (PCAC) equation relevant to neutral pion decay is modified to read

$$\partial^\mu\mathcal{F}^5_{3\mu}(x) = \left(f_\pi\mu_\pi^2/\sqrt{2}\right)\phi_\pi(x)+S\frac{\alpha_0}{4\pi}F^{\xi\sigma}(x)F^{\tau\rho}(x)\epsilon_{\xi\sigma\tau\rho} \qquad [6a]$$

with μ_π the pion mass, $f_\pi\simeq 131$ MeV the charged-pion decay constant, and S a constant determined by the constituent fermion charges and axial-vector couplings. Taking the matrix element of eqn [6a] between the vacuum state and a two-photon state, and using the fact that the left-hand side has a kinematic zero (the Sutherland–Veltman theorem), one sees that the $\pi^0\to\gamma\gamma$ amplitude F is completely determined by the anomaly term, giving the formula

$$F=-(\alpha/\pi)2S\sqrt{2}/f_\pi \qquad [6b]$$

For a single set of fractionally charged quarks, the amplitude F is a factor of three too small to agree with experiment; for three fractionally charged quarks (or an equivalent Han–Nambu triplet), eqn [6b] gives the correct neutral pion decay rate. This calculation was one of the first pieces of evidence for the color degree of freedom of quarks.

Anomaly Cancellation in Gauge Theories

In quantum electrodynamics, the gauge particle (the photon) couples to the vector current, and so the anomalous conservation properties of the axial-vector current have no effect. The same statement holds for the gauge gluons in quantum chromodynamics, when treated in isolation from the other interactions. However, in the electroweak theory that embeds quantum electrodynamics in a theory of the weak force, the gauge particles (the $W^\pm$ and Z intermediate bosons) couple to chiral currents, which are left- or right-handed linear combinations of the vector and axial-vector currents. In this case, the chiral anomaly leads to problems with the renormalizability of the theory, unless the anomalies cancel between different fermion species. Writing all fermions as left-handed, the condition for anomaly cancellation is

$$\mathrm{tr}\{T_\alpha,T_\beta\}T_\gamma=\mathrm{tr}(T_\alpha T_\beta+T_\beta T_\alpha)T_\gamma=0 \quad \text{for all } \alpha,\beta,\gamma \qquad [7]$$

with T_α the coupling matrices of gauge bosons to left-handed fermions. These conditions are obeyed in the standard model, by virtue of three nontrivial sum rules on the fermion gauge couplings being satisfied (four sum rules, if one includes the gravitational contribution to the chiral anomaly given in eqn [4c], which also cancels in the standard model). Note that anomaly cancellation in the locally gauged currents of the standard model does not imply anomaly cancellation in global-flavor currents. Thus, the flavor axial-vector current anomaly that gives the $\pi^0\to\gamma\gamma$ matrix element remains anomalous in the full electroweak theory. Anomaly cancellation imposes important constraints on the construction of grand unified models that combine the electroweak theory with quantum chromodynamics. For instance, in SU(5) the fermions are put into a $\bar{5}$ and 10 representation, which together, but not individually, are anomaly free. The larger unification groups SO(10) and E_6 satisfy eqn [7] for all representations, and so are automatically anomaly free.

Instanton Physics and the Theta Vacuum

The theory of anomalies is intimately tied to the physics associated with instanton classical Yang–Mills theory solutions. Since the instanton field

strength is self-dual, the nonvanishing instanton Euclidean action

$$S_E = \int d^4x \frac{1}{4} F_{\mu\nu} F^{\mu\nu} = 8\pi^2 \qquad [8a]$$

implies that the integral of the pseudoscalar density $F_{\mu\nu}F_{\lambda\sigma}\epsilon^{\mu\nu\lambda\sigma}$ over the instanton is also nonzero,

$$\int d^4x F_{\mu\nu} F_{\lambda\sigma} \epsilon^{\mu\nu\lambda\sigma} = 64\pi^2 \qquad [8b]$$

Referring back to eqn [4b], this means that the integral of the nonabelian chiral anomaly for fermions in the background field of an instanton is an integer, which in the Minkowski space continuation has the interpretation of a topological winding number change produced by the instanton tunneling solution. This fact has a number of profound consequences. Since a vacuum with a definite winding number $|\nu\rangle$ is unstable under instanton tunneling, careful analysis shows that the nonabelian vacuum that has correct clustering properties is a Fourier superposition

$$|\theta\rangle = \sum_\nu e^{i\theta\nu} |\nu\rangle \qquad [8c]$$

giving rise to the θ-vacuum of quantum chromodynamics, and a host of issues associated with (the lack of) strong CP violation, the Peccei–Quinn mechanism, and axion physics. Also, the fact that the integral of eqn [8b] is nonzero means that the U(1) chiral symmetry of quantum chromodynamics is broken by instantons, which as shown by 't Hooft resolves the longstanding "U(1) problem" of strong interactions, that of explaining why the flavor singlet pseudoscalar meson η' is not light, unlike its flavor octet partners.

Anomaly Matching Conditions

The anomaly structure of a theory, as shown by 't Hooft, leads to important constraints on the formation of massless composite bound states. Consider a theory with a set of left-handed fermions ψ^{if}, with i a "color" index acted on by a nonabelian gauge force, and f an ungauged family or "flavor" index. Suppose that the family multiplet structure is such that the global chiral symmetries associated with the flavor index have nonvanishing anomalies $\mathrm{tr}\{T_\alpha, T_\beta\}T_\gamma$. Then the 't Hooft condition asserts that if the color forces result in the formation of composite massless bound states of the original completely confined fermions, and if there is no spontaneous breaking of the original global flavor symmetries, then these bound states must contain left-handed spin-1/2 composites with a representation structure S that has the same anomaly coefficient as that in the underlying theory. In other words, we must have

$$\mathrm{tr}\{S_\alpha, S_\beta\}S_\gamma = \mathrm{tr}\{T_\alpha, T_\beta\}T_\gamma \qquad [9]$$

To prove this, one adjoins to the theory a set of right-handed spectator fermions ψ^f with the same flavor structure as the original set, but which are not acted on by the color force. These right-handed fermions cancel the original anomaly, making the underlying theory anomaly free at zero color coupling; since dynamics cannot spontaneously generate anomalies, the theory, when the color dynamics is turned on, must also have no global chiral anomalies. This implies that the bound-state spectrum must conspire to cancel the anomalies associated with the right-handed spectators; in other words, the bound-state anomaly structure must match that of the original fermions. This anomaly matching condition has found applications in the study of the possible compositeness of quarks and leptons. It has also been applied to the derivation of nonperturbative dynamical results in whole classes of supersymmetric theories, where the combined tools of holomorphicity, instanton physics, and anomaly matching have given incisive results.

Global Structure of Anomalies

We noted earlier that chiral anomalies are irreducible, in that they cannot be eliminated by adding a local polynomial counter-term to the action. However, anomalies can be described by a nonlocal effective action, obtained by integrating out the fermion field dynamics, and this point of view proves very useful in the nonabelian case. Starting with the abelian case for orientation, we note that if A^μ is an external axial-vector field, and we write an effective action $\Gamma[A]$, then the axial-vector current j^5_μ associated with A^μ is given (up to an overall constant) by the variational derivative expression

$$j^5_\mu(x) = \frac{\delta\Gamma[A]}{\delta A^\mu(x)} \qquad [10a]$$

and the abelian anomaly appears as the fact that the expression

$$\partial^\mu j^5_\mu = X\Gamma[A] = G \neq 0, \quad X = \partial^\mu \frac{\delta}{\delta A^\mu(x)} \qquad [10b]$$

is nonvanishing even when the theory is classically chiral invariant. Turning now to the nonabelian case, the variational derivative appearing in eqns [10a] and [10b] must be replaced by an appropriate

covariant derivative. In terms of the internal-symmetry component fields A^a_μ and V^a_μ of the Yang–Mills potentials of eqn [4a], one introduces operators

$$\begin{aligned} -X^a(x) &= \partial^\mu \frac{\delta}{\delta A^a_\mu(x)} + f_{abc} V^b_\mu \frac{\delta}{\delta A^c_\mu(x)} \\ &\quad + f_{abc} A^b_\mu \frac{\delta}{\delta V^c_\mu(x)} \\ -Y^a(x) &= \partial^\mu \frac{\delta}{\delta V^a_\mu(x)} + f_{abc} V^b_\mu \frac{\delta}{\delta V^c_\mu(x)} \\ &\quad + f_{abc} A^b_\mu \frac{\delta}{\delta A^c_\mu(x)} \end{aligned} \qquad [11a]$$

with f_{abc} the antisymmetric nonabelian group structure constants. The operators X^a and Y^a are easily seen to obey the commutation relations

$$\begin{aligned} [X^a(x), X^b(y)] &= f_{abc}\delta(x-y)Y_c(x) \\ [X^a(x), Y^b(y)] &= f_{abc}\delta(x-y)X_c(x) \\ [Y^a(x), Y^b(y)] &= f_{abc}\delta(x-y)Y_c(x) \end{aligned} \qquad [11b]$$

Let $\Gamma[V, A]$ be the effective action as a functional of the fields V^μ, A^μ, constructed so that the vector currents are covariantly conserved, as expressed formally by

$$Y^a\Gamma[V, A] = 0 \qquad [12a]$$

Then the nonabelian axial-vector current anomaly is given by

$$X^a\Gamma[V, A] = G^a \qquad [12b]$$

From eqns [12a] and [12b] and the first line of eqn [11b], we have

$$\begin{aligned} X^b G^a - X^a G^b &= (X^b X^a - X^a X^b)\Gamma[V, A] \\ &\propto f_{abc} Y^c \Gamma[V, A] = 0 \end{aligned} \qquad [12c]$$

which is the Wess–Zumino consistency condition on the structure of the anomaly G^a. It can be shown that this condition uniquely fixes the form of the nonabelian anomaly to be that of eqn [4b], up to an overall constant, which can be determined by comparison with the simplest anomalous AVV triangle graph. A physical consequence of the consistency condition is that the $\pi^0 \to \gamma\gamma$ decay amplitude determines uniquely certain other anomalous amplitudes, such as $2\gamma \to 3\pi, \gamma \to 3\pi$, and a five pseudoscalar vertex.

Although the action $\Gamma[V, A]$ is necessarily nonlocal, Wess and Zumino were able to write down a local action, involving an auxiliary pseudoscalar field, that obeys the anomalous Ward identities and the consistency conditions. Subsequently, Witten gave a new construction of this local action, in terms of the integral of a fifth-rank antisymmetric tensor over a five-dimensional disk which has a four-dimensional space as its boundary. He also showed that requiring $e^{i\Gamma}$ to be independent of the choice of the spanning disk requires, in analogy with Dirac's quantization condition for monopole charge, the condition that the overall coefficient in the nonabelian anomaly be quantized in integer multiples. Comparison with the lowest-order triangle diagram shows that in the case of $SU(N_c)$ gauge theory, this integer is just the number of colors N_c. Thus, global considerations tightly constrain the nonabelian chiral anomaly structure, and dictate that up to an integer-proportionality constant, it must have the form given in eqns [4a] and [4b].

Trace Anomalies

The discovery of chiral anomalies inspired the search for other examples of anomalous behavior. First indications of a perturbative trace anomaly obtained in a study of broken scale invariance by Coleman and Jackiw were shown by Crewther, and by Chanowitz and Ellis, to correspond to an anomaly in the three-point function $\theta^\sigma_\sigma V_\mu V_\nu$, where θ^μ_μ is the energy–momentum tensor. Letting $\Delta_{\mu\nu}(p)$ be the momentum space expression for this three-point function, and $\Pi_{\mu\nu}$ the corresponding $V_\mu V_\nu$ two-point function, the trace anomaly equation in quantum electrodynamics reads

$$\begin{aligned} \Delta_{\mu\nu}(p) &= \left(2 - p_\sigma \frac{\partial}{\partial p_\sigma}\right)\Pi_{\mu\nu}(p) \\ &\quad - \frac{R}{6\pi^2}(p_\mu p_\nu - \eta_{\mu\nu}p^2) \end{aligned} \qquad [13a]$$

with the first term on the right-hand side the naive divergence, and the second term the trace anomaly, with anomaly coefficient R given by

$$R = \sum_{i,\text{spin}\frac{1}{2}} Q_i^2 + \frac{1}{4}\sum_{i,\text{spin}\,0} Q_i^2 \qquad [13b]$$

The fact that there should be a trace anomaly can readily be inferred from a trace analog of the Pauli–Villars regulator argument for the chiral anomaly given in eqn [3a]. Letting $j = \bar{\psi}\psi$ be the scalar current in abelian electrodynamics, one has

$$\theta^\mu_\mu|_{m_0} - \theta^\mu_\mu|_{M_0} = m_0 j|_{m_0} - M_0 j|_{M_0} \qquad [13c]$$

Taking the vacuum to two-photon matrix element of this equation, and imposing vector-current conservation, one finds that the matrix element $\langle 0|j|_{M_0}|\gamma\gamma\rangle$ is proportional to $M_0^{-1}\langle 0|F_{\lambda\sigma}F^{\lambda\sigma}|\gamma\gamma\rangle_{M_0}$ for a large regulator mass, and so makes a

nonvanishing contribution to the right-hand side of eqn [13c], giving the lowest-order trace anomaly.

Unlike the chiral anomaly, the trace anomaly is renormalized in higher orders of perturbation theory; heuristically, the reason is that whereas boson field regulators do not affect the chiral symmetry properties of a gauge theory (which are determined just by the fermionic terms in the Lagrangian), they do alter the energy–momentum tensor, since gravitation couples to all fields, including regulator fields. An analysis using the Callan–Symanzik equations shows, however, that the trace anomaly is computable to all orders in terms of various renormalization group functions of the coupling. For example, in abelian electrodynamics, defining $\beta(\alpha)$ and $\delta(\alpha)$ by $\beta(\alpha) = (m/\alpha)\partial\alpha/\partial m$ and $1 + \delta(\alpha) = (m/m_0)\partial m_0/\partial m$, the trace of the energy–momentum tensor is given to all orders by

$$\theta^{\mu}_{\mu} = [1 + \delta(\alpha)]m_0\bar{\psi}\psi + \tfrac{1}{4}\beta(\alpha)N[F_{\lambda\sigma}F^{\lambda\sigma}] + \cdots \quad [14]$$

with $N[\ ]$ specifying conditions that make the division into two terms in eqn [14] unique, and with the ellipsis $\cdots$ indicating terms that vanish by the equations of motion. A similar relation holds in the nonabelian case, again with the β function appearing as the coefficient of the anomalous $\mathrm{tr}\, N[F_{\lambda\sigma}F^{\lambda\sigma}]$ term.

Just as in the chiral anomaly case, when spin-0, spin-1/2, or spin-1 fields propagate on a background spacetime, there are curvature-dependent contributions to the trace anomaly, in other words, gravitational anomalies. These typically take the form of complicated linear combinations of terms of the form $R^2, R_{\mu\nu}R^{\mu\nu}, R_{\mu\nu\lambda\sigma}R^{\mu\nu\lambda\sigma}, R_{,\mu}{}^{,\mu}$, with coefficients depending on the matter fields involved.

In supersymmetric theories, the axial-vector current and the energy–momentum tensor are both components of the supercurrent, and so their anomalies imply the existence of corresponding supercurrent anomalies. The issue of how the nonrenormalization of chiral anomalies (which have a supercurrent generalization given by the Konishi anomaly), and the renormalization of trace anomalies, can coexist in supersymmetric theories originally engendered considerable confusion. This apparent puzzle is now understood in the context of a perturbatively exact expression for the β function in supersymmetric field theories (the so-called NSVZ, for Novikov, Shifman, Vainshtein, and Zakharov, β function). Supersymmetry anomalies can be used to infer the structure of effective actions in supersymmetric theories, and these in turn have important implications for possibilities for dynamical supersymmetry breaking. Anomalies may also play a role, through anomaly mediation, in communicating supersymmetry breaking in "hidden sectors" of a theory, which do not contain the physical fields that we directly observe, to the "physical sector" containing the observed fields.

Further Anomaly Topics

The above discussion has focused on some of the principal features and applications of anomalies. There are further topics of interest in the physics and mathematics of anomalies that are discussed in detail in the references cited in the "Further reading" section. We briefly describe a few of them here.

Anomalies in Other Spacetime Dimensions and in String Theory

The focus above has been on anomalies in four-dimensional spacetime, but anomalies of various types occur both in lower-dimensional quantum field theories (such as theories in two- and three-dimensional spacetimes) and in quantum field theories in higher-dimensional spacetimes (such as $N = 1$ supergravity in ten-dimensional spacetime). Anomalies also play an important role in the formulation and consistency of string theory. The bosonic string is consistent only in 26-dimensional spacetime, and the analogous supersymmetric string only in ten-dimensional spacetime, because in other dimensions both these theories violate Lorentz invariance after quantization. In the Polyakov path-integral formulation of these string theories, these special dimensions are associated with the cancellation of the Weyl anomaly, which is the relevant form of the trace anomaly discussed above. Yang–Mills, gravitational, and mixed Yang–Mills gravitational anomalies make an appearance both in $N = 1$ ten-dimensional supergravity and in superstring theory, and again special dimensions play a role. In these theories, only when the associated internal symmetry groups are either SO(32) or $E_8 \times E_8$ is elimination of all anomalies possible, by cancellation of hexagon-diagram anomalies with anomalous tree diagrams involving exchange of a massless antisymmetric two-form field. This mechanism, due to Green and Schwarz, requires the factorization of a sixth-order trace invariant that appears in the hexagon anomaly in terms of lower-order invariants, as well as two numerical conditions on the adjoint representation generator structure, restricting the allowed gauge groups to the two noted above.

Covariant versus Consistent Anomalies; Descent Equations

The nonabelian anomaly of eqns [4a] and [4b] is called the "consistent anomaly," because it obeys the

Wess–Zumino consistency conditions of eqn [12c]. This anomaly, however, is not gauge covariant, as can be seen from the fact that it involves not only the Yang–Mills field strengths $F^{\mu\nu}_{V,A}$, but the potentials V^μ, A^μ as well. It turns out to be possible, by adding appropriate polynomials to the currents, to transform the consistent anomaly to a form, called the "covariant anomaly," which is gauge covariant under gauge transformations of the potentials V^μ, A^μ. This anomaly, however, does not obey the Wess–Zumino consistency conditions, and cannot be obtained from variation of an effective action functional.

The consistent anomalies (but not the covariant anomalies) obey a remarkable set of relations, called the Stora–Zumino descent equations, which relate the abelian anomaly in $2n+2$ spacetime dimensions to the nonabelian anomaly in $2n$ spacetime dimensions. This set of equations has been interpreted physically by Callan and Harvey as reflecting the fact that the Dirac equation has chiral zero modes in the presence of strings in $2n+2$ dimensions and of domain walls in $2n+1$ dimensions.

Anomalies and Fermion Doubling in Lattice Gauge Theories

A longstanding problem in lattice formulations of gauge field theories is that when fermions are introduced on the lattice, the process of discretization introduces an undesirable doubling of the fermion particle modes. In particular, when an attempt is made to put chiral gauge theories, such as the electroweak theory, on the lattice, one finds that the doublers eliminate the chiral anomalies, by cancellation between modes with positive and negative axial-vector charge. Thus, for a long time, it appeared doubtful whether chiral gauge theories could be simulated on the lattice. However, recent work has led to formulations of lattice fermions that use a mathematical analog of a domain wall to successfully incorporate chiral fermions and the chiral anomaly into lattice gauge theory calculations.

Relation of Anomalies to the Atiyah–Singer Index Theorem

The singlet ($\lambda^a_A = 1$) anomaly of eqn [4b] is closely related to the Atiyah–Singer index theorem. Specifically, the Euclidean spacetime integral of the singlet anomaly constructed from a gauge field can be shown to give the index of the related Dirac operator for a fermion moving in that background gauge field, where the index is defined as the difference between the numbers of right- and left-handed zero-eigenvalue normalizable solutions of the Dirac equation. Since the index is a topological invariant, this again implies that the Euclidean spacetime integral of the anomaly is a topological invariant, as noted above in our discussion of instanton-related applications of anomalies.

Retrospect

The wide range of implications of anomalies has surprised – even astonished – the founders of the subject. New anomaly applications have appeared within the last few years, and very likely the future will see continued growth of the area of quantum field theory concerned with the physics and mathematics of anomalies.

Acknowledgment

This work is supported, in part, by the Department of Energy under grant #DE-FG02-90ER40542.

See also: Bosons and Fermions in External Fields; BRST Quantization; Effective Field Theories; Gauge Theories from Strings; Gerbes in Quantum Field Theory; Index Theorems; Lagrangian Dispersion (Passive Scalar); Lattice Gauge Theory; Nonperturbative and Topological Aspects of Gauge Theory; Quantum Electrodynamics and Its Precision Tests; Quillen Determinant; Renormalization: General Theory; Seiberg–Witten Theory.

Further Reading

Adler SL (1969) Axial-vector vertex in spinor electrodynamics. *Physical Review* 177: 2426–2438.

Adler SL (1970) Perturbation theory anomalies. In: Deser S, Grisaru M, and Pendleton H (eds.) *Lectures on Elementary Particles and Quantum Field Theory*, vol. 1, pp. 3–164. Cambridge, MA: MIT Press.

Adler SL (2005) Anomalies to all orders. In: 't Hooft G (ed.) *Fifty Years of Yang–Mills Theory*, pp. 187–228. Singapore: World Scientific.

Adler SL and Bardeen WA (1969) Absence of higher order corrections in the anomalous axial-vector divergence equation. *Physical Review* 182: 1517–1536.

Bardeen W (1969) Anomalous ward identities in spinor field theories. *Physical Review* 184: 1848–1859.

Bell JS and Jackiw R (1969) A PCAC puzzle: $\pi^0 \to \gamma\gamma$ in the σ-model. *Nuovo Cimento A* 60: 47–61.

Bertlmann RA (1996) *Anomalies in Quantum Field Theory*. Oxford: Clarendon.

De Azcárraga JA and Izquierdo JM (1995) *Lie Groups, Lie Algebras, Cohomology and Some Applications in Physics*, ch. 10. Cambridge: Cambridge University Press.

Fujikawa K and Suzuki H (2004) *Path Integrals and Quantum Anomalies*. Oxford: Oxford University Press.

Golterman M (2001) Lattice chiral gauge theories. *Nuclear Physics Proceeding Supplements* 94: 189–203.

Green MB, Schwarz JH, and Witten E (1987) *Superstring Theory*. vol. 2, sects. 13.3–13.5. Cambridge: Cambridge University Press.

Hasenfratz P (2005) Chiral symmetry on the lattice. In: 't Hooft G (ed.) *Fifty Years of Yang–Mills Theory*, pp. 377–398. Singapore: World Scientific.

Jackiw R (1985) Field theoretic investigations in current algebra and topological investigations in quantum gauge theories. In: Treiman S, Jackiw R, Zumino B, and Witten E (eds.) *Current Algebra and Anomalies*. Singapore: World Scientific and Princeton: Princeton University Press.

Jackiw R (2005) Fifty years of Yang–Mills theory and our moments of triumph. In: 't Hooft G (ed.) *Fifty Years of Yang–Mills Theory*, pp. 229–251. Singapore: World Scientific.

Makeenko Y (2002) *Methods of Contemporary Gauge Theory*, ch. 3. Cambridge: Cambridge University Press.

Neuberger H (2000) Chiral fermions on the lattice. *Nuclear Physics Proceeding Supplements* 83: 67–76.

Polchinski J (1999) *String Theory*, vol. 1, sect. 3.4; vol. 2, sect. 12.2. Cambridge: Cambridge University Press.

Shifman M (1997) Non-perturbative dynamics in supersymmetric gauge theories. *Progress in Particle and Nuclear Physics* 39: 1–116.

van Nieuwenhuizen P (1988) *Anomalies in Quantum Field Theory: Cancellation of Anomalies in $d=10$ Supergravity*. Leuven: Leuven University Press.

Volovik GE (2003) *The Universe in a Helium Droplet*, ch. 18. Oxford: Clarendon.

Weinberg S (1996) *The Quantum Theory of Fields, Vol. II Modern Applications*, ch. 22. Cambridge: Cambridge University Press.

Zee A (2003) *Quantum Field Theory in a Nutshell*, sect. IV.7. Princeton: Princeton University Press.

BRST Quantization

M Henneaux, *Université Libre de Bruxelles, Bruxelles, Belgium*

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The BRST symmetry was originally introduced in the seminal papers by Becchi *et al.* (1976) and Tyutin (1975) for Yang–Mills gauge theories as a tool for controlling the renormalization of the models in a consistent (gauge-independent) way. This symmetry was discovered as a residual symmetry of the gauge-fixed action. It was realized later that, in fact, the BRST construction is quite general, in the sense that it covers arbitrary gauge theories and not just Yang–Mills gauge models. Furthermore, it is intrinsic, in that no gauge choice is actually necessary to define it.

The purpose of this review is to explain the general, intrinsic features of the BRST formalism applicable to "any" gauge theory. The proper setting for discussing these issues is that of homological algebra (Stasheff (1998), and references therein). This article first explains the necessary algebraic material underlying the construction and then illustrates it in the cases of the Hamiltonian BRST formalism and the Lagrangian BRST formalism.

A Result from Homological Algebra

The main result of homological algebra needed in the BRST construction deals with a differential complex $\mathcal{C}$ with two gradings. The first grading is an $\mathbb{N}$-degree and is called the "resolution degree," or "r-degree." The second grading is a $\mathbb{Z}$-degree and is called the total ghost number. It is denoted by gh. We assume that there are two odd derivations δ and s_0 that have the following properties:

$$\begin{aligned} r(\delta) &= -1, \qquad \mathrm{gh}(\delta) = 1 \\ r(s_0) &= 0, \qquad \mathrm{gh}(s_0) = 1 \end{aligned} \tag{1}$$

and

$$\delta^2 = 0, \quad s_0\delta + \delta s_0 = 0, \quad s_0^2 = -[\delta, s_1] \tag{2}$$

Notes:

for some derivation s_1 of r-degree 1 and ghost number 1. The bracket $[\cdot,\cdot]$ is the graded commutator – in this specific case, the anticommutator. We also assume that the homology of δ vanishes at nonzero value of the r-degree, both in the original complex $\mathcal{C}$,

$$H_k(\delta,\mathcal{C}) = 0, \quad k > 0 \qquad [3]$$

(which is equivalent to $\delta a = 0$, $r(a) > 0 \Rightarrow a = \delta b$) and in the space of derivations,

$$[\alpha,\delta] = 0, \quad r(\alpha) \neq 0 \Rightarrow \alpha = [\beta,\delta] \qquad [4]$$

where α and β are both derivations in $\mathcal{C}$. The r-degree of a homogeneous linear operator α is defined through $r(\alpha(x)) = r(\alpha) + r(x)$ for any element $x \in \mathcal{C}$ and is negative when α decreases the r-degree.

In $H_0(\delta,\mathcal{C})$, the (odd) derivation s_0 defines a differential. The cohomology of s_0 modulo δ, denoted $H^k(s_0, H_0(\delta,\mathcal{C}))$, is the cohomology of s_0 in $H_0(\delta,\mathcal{C})$. It is explicitly defined through the cocycle condition

$$s_0 a = \delta m \qquad [5]$$

with coboundaries of the form

$$s_0 b + \delta n \qquad [6]$$

The central result underlying the BRST construction is:

Theorem 1 *Given the above setting, there exists an odd derivation s in $\mathcal{C}$ with the following properties*:

$$s = \delta + s_0 + s_1 + \cdots \qquad [7]$$

$$r(s_k) = k, \qquad \mathrm{gh}(s_k) = 1 \qquad [8]$$

$$s^2 = 0 \qquad [9]$$

Furthermore, one has

$$H^k(s,\mathcal{C}) = H^k(s_0, H_0(\delta,\mathcal{C})) \qquad [10]$$

The proof is straightforward (see, e.g., Henneaux and Teitelboim (1992)). In particular, the proof of [10] is a standard spectral sequence argument with a sequence that collapses after the second step. It is interesting to note that, contrary to s_0, which is only a differential modulo δ,s is a true differential. The construction of s provides a model for $H^k(s_0, H_0(\delta,\mathcal{C}))$. The differential s is not unique, but this does not affect the subsequent discussion.

In physical applications, the total ghost number is a derived quantity. The primary gradings are the resolution degree and the "filtration degree" called the pure ghost number and denoted pgh. It is an $\mathbb{N}$-degree and one has

$$\mathrm{gh} = \mathrm{pgh} - r \qquad [11]$$

The r-degree is known as the antighost or antifield number, depending on the context (see below). When $r(x) = 0$, one has $\mathrm{gh}(x) = \mathrm{pgh}(x)$. Since the pure ghost number is non-negative, this implies that

$$H^k(s,\mathcal{C}) = 0, \quad k < 0 \qquad [12]$$

A Geometric Application

Geometric Setting

Theorem 1 is relevant to the following situation. Consider a surface Σ in a manifold M, defined by equations

$$f_a = 0 \qquad [13]$$

which may or may not be independent. (We assume for definiteness that the variables in M are bosonic, that is, that M is an ordinary manifold – as opposed to a supermanifold. The graded case can be covered without difficulty by including appropriate sign factors at the relevant places.) Assume that Σ is partitioned by orbits generated by vector fields $\boldsymbol{X}_\alpha$ defined everywhere in M, tangent to Σ and closing on Σ in the Lie bracket,

$$[\boldsymbol{X}_\alpha, \boldsymbol{X}_\beta] = C^\gamma{}_{\alpha\beta}\boldsymbol{X}_\gamma + \text{“more”} \qquad [14]$$

where "more" denotes terms that vanish on Σ. We assume, for simplicity, that the vector fields $\boldsymbol{X}_\alpha$ are linearly independent of Σ, although this is not necessary. The formalism can be developed in the nonindependent case, but it then requires more variables. We are interested in the quotient space $\Sigma/\mathcal{O}$ of the surface Σ by the orbits. To guide the geometrical intuition, we shall assume that this quotient space is a smooth manifold (the fiber of the orbits, etc.), and we shall suggestively adopt notations adapted to this best possible case. The approach, being purely algebraic, is in fact more general. (Accordingly, the notations should be understood with a liberal mind.)

The aim here is to describe the algebra of "observables," that is, the algebra $C^\infty(\Sigma/\mathcal{O})$ of functions on the quotient space $\Sigma/\mathcal{O}$. The terminology "observables" anticipates the physical situation discussed below, where the orbits are the "gauge orbits." In order to describe algebraically the algebra of observables, one observes that this algebra is obtained

through a two-step procedure. First, one restricts the functions from M to Σ. Second, one imposes the invariance condition along the orbits. To each of these steps corresponds a separate differential.

Longitudinal Complex

The longitudinal complex is associated with the second step. One can consider on Σ an "exterior derivative operator D along the gauge orbits." This operator is defined on functions on Σ as

$$Df = X_\alpha(f)C^\alpha \qquad [15]$$

where the 1-forms C^α dual to the X's are called ghosts. In the physical context, the form-degree is the pgh described earlier, and so $\mathrm{pgh}(C^\alpha)=1$. The action of D on the ghosts is given by

$$DC^\alpha = -\tfrac{1}{2}C^\gamma{}_{\alpha\beta}C^\alpha C^\beta \qquad [16]$$

The longitudinal complex $\mathcal{L}_\Sigma$ is the complex of exterior forms along the gauge orbits. In our representation used here, it is given by the space of polynomials in the ghosts C^α with coefficients that are functions on Σ. The exterior derivative D is defined on this space by extending the formulas [15] and [16] so that it is an odd derivation. One clearly has (on Σ)

$$D^2 = 0 \qquad [17]$$

The functions on the quotient space $\Sigma/\mathcal{O}$ are just the elements of the zeroth cohomological group $H^0(D,\mathcal{L}_\Sigma)$,

$$H^0(D,\mathcal{L}_\Sigma) = C^\infty(\Sigma/\mathcal{O}) \qquad [18]$$

In general, $H^k(D,\mathcal{L}_\Sigma)\neq 0$.

Koszul–Tate Differential δ

The Koszul–Tate differential δ implements the first step in the reduction procedure. More precisely, it provides an algebraic resolution of the algebra $C^\infty(\Sigma)$ of the smooth functions on the surface Σ.

That algebra can be identified with the quotient algebra

$$C^\infty(\Sigma) = C^\infty(M)/\mathcal{N} \qquad [19]$$

where $\mathcal{N}$ is the ideal of functions that vanish on Σ. The Koszul–Tate complex $\mathcal{K}$ is defined by adding one new generator for each equation $f_a=0$ defining Σ, denoted t^*_a and assigned r-degree 1. In the algebra $C^\infty(M)\otimes\wedge(t^*_a)$ (where $\wedge(t^*_a)$ is the exterior algebra on t^*), one defines δ through

$$\delta f = 0 \quad \forall f\in C^\infty(M), \qquad \delta t^*_a = f_a \qquad [20]$$

and extends it as an odd derivation. It is clear that $r(\delta)=-1$ and that $\delta^2=0$. Because the functions on M are annihilated by δ, they are clearly cycles at r-degree zero. Because the left-hand side f_a of the equations $f_a=0$ are exact (equal to δt^*_a), the ideal $\mathcal{N}$ coincides with the set of boundaries in degree zero.

Thus,

$$H_0(\delta,\mathcal{K}) = C^\infty(\Sigma) \qquad [21]$$

We see accordingly that δ successfully enforces the restriction to the surface Σ through its homology in degree zero.

However, if the equations $f_a=0$ are not independent, this is not the end of the story. Indeed, any identity $Z^a_A f_a=0$ on the functions f_a leads to a nontrivial cycle $Z^a_A t^*_a$ in r-degree 1, $\delta(Z^a_A t^*_a)=0$. This is undesirable. To cure this drawback, one introduces further generators t^*_A in r-degree 2, one for each identity $Z^a_A f_a=0$, and defines

$$\delta t^*_A = Z^a_A t^*_a, \quad r(t^*_A)=2 \qquad [22]$$

in order to "kill" the unwanted cycles $Z^a_A t^*_a$. The Koszul complex $\mathcal{K}$ is thus enlarged to contain these new (even) variables and redefined as

$$\mathcal{K} = C^\infty(M)\otimes\wedge(t^*_a)\otimes S(t^*_A) \qquad [23]$$

where $S(t^*_A)$ is the symmetric algebra in t^*_A. The operator δ is extended to $\mathcal{K}$ as an odd derivation. One has $\delta^2=0$ and the property [21] is unaffected by the inclusion of the new generators. Furthermore, by construction,

$$H_1(\delta,\mathcal{K}) = 0 \qquad [24]$$

If there is no "identity on the identities," we shall assume that the process stops. Otherwise, one needs to introduce further generators in r-degree 3 and possibly higher. When all the appropriate variables are included, there is no homology at higher r-degree. Thus,

$$H_k(\delta,\mathcal{K}) = 0, \quad k>0 \qquad [25]$$

Combining δ with D

We now turn to the problem of combining the Koszul–Tate complex with the longitudinal complex, so as to implement the full reduction. To that end, we define $\mathcal{C}$ by adding the ghosts to $\mathcal{K}$,

$$\mathcal{C} = \mathcal{K}\otimes\wedge(C_\alpha) = 0 \qquad [26]$$

We then extend the action of the Koszul–Tate differential in the simplest way which preserves all gradings, namely

$$\delta C_\alpha = 0 \qquad [27]$$

It is clear that the homology of δ in $\mathcal{C}$ is given by

$$H_0(\delta,\mathcal{C}) = \mathcal{L}_\Sigma, \qquad H_k(\delta,\mathcal{C}) = 0 \quad (k>0) \qquad [28]$$

One can also extend the longitudinal derivative D to the whole complex $\mathcal{C}$ because the vector fields X_α are defined throughout M and so, the definitions [15] and [16] make sense in $\mathcal{C}$. One defines the action of D on the generators t^* by requiring that

$$D\delta + \delta D = 0 \qquad [29]$$

This is easily verified to be possible. However, the (odd) derivation so obtained fails to be a differential in $\mathcal{C}$ when the vector fields X_α do not close off the surface Σ. In that case, the gauge transformations are not integrable off Σ; one says that they form an "open algebra." One has then $D^2 = 0$ only on Σ, or, more precisely,

$$D^2 = -\delta s_1 - s_1\delta \qquad [30]$$

for some (odd) derivation s_1 (that vanishes in the "closed algebra" case). But this situation is precisely the one discussed earlier, with the Koszul–Tate differential being indeed δ, as anticipated by the notation, and the longitudinal differential D playing the role of s_0 (the degrees also match). Applying the theorem discussed there, we can conclude:

Theorem 2 *There exists a differential s in $\mathcal{C}$,*

$$s = \delta + D + s_1 + \cdots, \qquad s^2 = 0 \qquad [31]$$

such that

$$H^0(s,\mathcal{C}) = C^\infty(\Sigma/\mathcal{O}) \qquad [32]$$

This is an immediate consequence of Theorem 1 and eqns [18] and [28]. The differential s is known in the physical applications described below as the BRST differential.

Hamiltonian BRST Construction

As a first application of the above setting, we consider the Hamiltonian description of gauge systems. As already known, gauge systems are characterized in the Hamiltonian description by constraints and, for this reason, are called "constrained Hamiltonian systems." Furthermore, the gauge transformations generate gauge orbits on the constraint surface and the physical observables are the functions on the quotient space of the constraint surface by the gauge orbits.

A further important feature arises in the Hamiltonian formalism: the gauge transformations are canonical transformations that are generated by the first-class constraints. Assuming that all the second-class constraints have been eliminated and that the bracket being used is the Dirac bracket, one sees that there is a vector field X_α for each constraint function $f_a, \alpha \equiv a$. (The functions f_a are thus assumed to be independent since the vector fields X_α are assumed to be so. If not, further variables are needed, but the analysis proceeds along the same ideas.)

This implies, in turn, that there is a pairing between the ghosts C^a associated with the longitudinal exterior derivative and the generators t^*_a of the Koszul–Tate complex. This pairing enables one to extend the bracket structure defined on the phase space to the pairs (C^a, t^*_a) by declaring that these are canonically conjugate. The variables t^*_a are the momenta conjugate to the ghosts, $[t^*_a, C^b] = \delta^b_a$. Accordingly, the complex $\mathcal{C}$ relevant to the Hamiltonian situation,

$$\mathcal{C} = C^\infty(P) \otimes \wedge(C^a) \wedge (t^*_a) \qquad [33]$$

has a phase-space structure (here, $P \equiv M$ is the manifold obtained after eliminating the second-class constraints, equipped with the Dirac bracket). The space $\mathcal{C}$ is known as the "extended phase space." The r-degree is called "antighost number" in the Hamiltonian context.

By the general theorem described in the previous section, one knows that the cohomology at $\mathrm{gh} = 0$ of the BRST differential is isomorphic to the algebra of the observables. Thus, there are two alternative ways to describe this physical algebra, either through reduction, by eliminating the redundant (gauge) variables, or cohomologically in an extended space containing additional variables, the ghosts, and their momenta.

There is an additional interesting feature of the BRST construction in the Hamiltonian case: the BRST transformation is a canonical transformation in the extended phase space, in the sense that

$$sF = [\Omega, F] \qquad [34]$$

for some "BRST generator" Ω of ghost number 1 ($F, \Omega \in \mathcal{C}$). The nilpotency s^2 of the BRST differential is equivalent to

$$[\Omega, \Omega] = 0 \qquad [35]$$

That s is canonically generated implies that the cohomological BRST groups come with a natural bracket structure: the Poisson bracket of the extended phase space passes on to the BRST cohomological groups. In particular, $H^0(s,\mathcal{C})$, equipped with this bracket structure, is isomorphic (as Poisson algebra) to the algebra of physical observables.

Lagrangian BRST Construction

The analysis of the Lagrangian BRST construction, due to Batalin and Vilkovisky (1981) ("antifield formalism"), proceeds in the same way because the covariant description of the space of observables involves also the same geometric ingredients. The surface Σ is now the "stationary surface," that is, the space of solutions to the equations of motion. The space M in which it is embedded is the space of all field histories. The gauge symmetry acts on this space. Furthermore, the gauge vector fields are tangent to Σ since a solution is mapped on a solution by a gauge transformation. The integral submanifolds are the gauge orbits. The observables are the functions on the quotient space.

Since the equations of motion follow from an action principle, there are as many equations as there are fields φ^i. The corresponding generators t_a^* in the Koszul–Tate complex (at degree 1) are called "antifields conjugate to the fields" and are denoted φ_i^*. The r-degree is known as "antifield" (or also "antighost") number. The gauge symmetry of the action implies Noether identities on the equations of motion. These are, therefore, not independent. According to the above general discussion, there are further generators in the Koszul–Tate complex, at degree 2. More precisely, there are as many new generators in degree 2 as there are Noether identities or independent gauge symmetries. These are called antifields conjugate to the ghosts and denoted C_α^*.

In the longitudinal complex, one has the ghosts C^α, with as many ghosts as there are gauge symmetries. Thus, the BRST complex is the space

$$\mathcal{C} = C^\infty(M) \otimes \wedge(C^\alpha) \otimes \wedge(\varphi_i^*) \otimes S(C_\alpha^*) \qquad [36]$$

where M is the space of all field histories. There is now a natural pairing between the original field variables φ^i and the antifields φ_i^*, as well as between the ghosts C^α and the antifields C_α^*. One thus defines a bracket in which the fields φ^i and the ghosts C^α on the one hand, and the antifields φ_i^* and C_α^* on the other, are declared to be conjugate. This bracket is denoted by parentheses,

$$(\varphi^i, \varphi_j^*) = \delta_j^i, \qquad (C^\alpha, C_\beta^*) = \delta_\beta^\alpha \qquad [37]$$

However, since the bracket pairs variables with degrees that add up to -1, it is in fact an "odd bracket," called the "antibracket."

The BRST differential is again canonically generated, but this time in the antibracket,

$$sF = (S, F), \quad F \in \mathcal{C} \qquad [38]$$

where the generator S is an even function of the fields, the ghosts and the antifields, with $\mathrm{gh} = 0$ (the ghost number is carried by the odd antibracket). The nilpotency $s^2 = 0$ of the BRST differential is equivalent to the crucial "master equation,"

$$(S, S) = 0 \qquad [39]$$

Because the BRST differential is canonically generated, there is a natural bracket in cohomology. This bracket is not the Poisson bracket of observables (at $\mathrm{gh} = 0$) because it changes the ghost number by one unit. One can, however, relate it to the Poisson bracket of observables (Barnich and Henneaux 1996); furthermore, it plays an important role in the study of the consistent deformations of the action.

Spacetime Locality

In the context of local field theory, one is often interested in a particular class of functions of the field histories, namely the so-called space of local functionals. A local functional is, by definition, the integral of a local n-form (where n is the spacetime dimension). A local n-form reads, in local coordinates,

$$\omega = f(x)\, \mathrm{d}^n x \qquad [40]$$

where $f(x)$ depends on the fields at x as well as on a finite number of their derivatives. When the ghosts and the antifields are included, the local functions depend on them in the same way.

The previous general cohomological result was derived in the space of all function(al)s, without locality restriction. When changing the space of cochains, one may change the cohomology. For instance, a local functional which is BRST-trivial in the space of all functionals may become nontrivial in the space of local functionals. This indeed happens here because the homology of the Koszul–Tate differentials usually no longer vanishes at strictly positive r-degree in the space of local functionals, where it is related to local conservation laws. As a result, the analysis of the BRST cohomology in the space of local functionals is an interesting and nontrivial problem. In particular, the cohomological groups $H^k(s)$ in the space of local functionals may not vanish at negative ghost numbers.

BRST Quantization

The quantization of a dynamical system can proceed along different lines. For gauge models, the path-integral approach is most efficiently pursued in the context of the antifield formalism. We shall briefly outline here the general principles underlying the

operator approach, which is based on the Hamiltonian formalism.

In the operator approach, all the variables, including the ghosts and the conjugate momenta, are realized as operators in a space endowed with a nonpositive-definite inner product (because of the ghosts and the gauge modes). Real dynamical variables become formally Hermitian operators. Ignoring anomalies, the BRST generator Ω becomes an operator that fulfills the conditions

$$\Omega^* = \Omega, \qquad \Omega^2 = 0 \qquad [41]$$

(which allows for nontrivial solutions $\Omega \neq 0$ because the inner product is not positive definite). The second relation is a consequence of the classical Poisson bracket relation $[\Omega,\Omega] = 0$ and the fact that the graded Poisson bracket of two odd objects becomes the anticommutator.

To remove the ghost and gauge redundancy, which has no physical content, one must impose a condition that selects physical states. The appropriate condition is motivated by the general cohomological result connecting the BRST cohomology with the algebra of physical observables. One imposes the condition

$$\Omega|\psi\rangle = 0 \qquad [42]$$

Because of [41], states of the form $\Omega|\chi\rangle$ are solutions of [42], but they have a vanishing inner product with any other physical states, including themselves. They are called null states. The physical states are given by the BRST state cohomology. The physical operators are given by the BRST operator cohomology at $\mathrm{gh} = 0$ and induce a well-defined action in the state cohomology. In particular, the Hamiltonian, being gauge invariant in the original theory, is represented by a BRST cohomological class, so that the time evolution maps physical states on physical states.

The whole scheme is (formally) consistent because exact BRST operators have vanishing matrix elements between states annihilated by the BRST operator Ω, while null states $|\phi\rangle$ are such that $\langle\psi|A|\phi\rangle = 0$ whenever A is a BRST-closed operator, $[A,\Omega] = 0$, and $|\psi\rangle$ a physical state. Problems may arise, however, if the classical relations $[\Omega,\Omega] = 0$ and $[H,\Omega] = 0$ are not satisfied in presence of extra terms of order $\hbar$, that is,

$$\Omega^2 \neq 0 \quad \text{or} \quad H\Omega + \Omega H \neq 0 \qquad [43]$$

In such cases, one says that they are anomalies. These are usually fatal to the consistency of the theory.

Some Applications

The number of applications of the BRST formalism is so large that it would be out of place to try being exhaustive here. Some of its main successes are outlined here, with suggestions for "Further reading."

Renormalization of Gauge Theories

First, there is the original context of perturbative renormalization and anomalies for gauge theories of the Yang–Mills type. The relevant cohomology here is the BRST cohomology in the space of local functionals involving the fields, the ghosts, and the antifields. The antifields are also known in this context as Zinn-Justin sources for the BRST variations of the fields and ghosts, since Zinn-Justin was the first to introduce them (with that meaning). Many authors have contributed to the full computation of the local BRST cohomology. A review is given in Barnich *et al.* (2000), where extensions to other theories are also indicated.

String Theory

Modern string theory would be inconceivable without the BRST formalism. This started with the pioneering paper by Kato and Ogawa (1983), where the critical dimension of the bosonic string was derived from the condition that Ω^2 should vanish (quantum mechanically), and where it was shown that the string physical states could be identified with the state BRST cohomology. The reader is referred to excellent monographs on modern string theory (see "Further reading").

Deformations of Gauge Models

The study of consistent deformations of a given gauge theory (i.e., the problem of introducing consistent couplings) is also efficiently dealt with in the BRST context. References to applications may be found in Henneaux (1998).

See also: Anomalies; Batalin–Vilkovisky Quantization; BF Theories; Constrained Systems; Functional Integration in Quantum Physics; Graded Poisson Algebras; Indefinite Metric; Perturbative Renormalization Theory and BRST; Quantum Chromodynamics; Quantum Field Theory: A Brief Introduction; Renormalization: General Theory; String Field Theory; Supermanifolds; Topological Sigma Models.

Further Reading

Barnich G, Brandt F, and Henneaux M (2000) Local BRST cohomology in gauge theories. *Physics Reports* 338: 439.

Barnich G and Henneaux M (1996) Isomorphisms between the Batalin–Vilkovisky antibracket and the Poisson bracket. *Journal of Mathematical Physics* 37: 5273.

Batalin IA and Vilkovisky GA (1977) Relativistic *S*-matrix of dynamical systems with boson and fermion constraints. *Physics Letters* B69: 309.

Batalin IA and Vilkovisky GA (1981) Gauge algebra and quantization. *Physics Letters* B102: 27.

Becchi C, Rouet A, and Stora R (1976) Renormalization of gauge theories. *Annals of Physics, NY* 98: 287.

Fradkin ES and Vilkovisky GA (1975) Quantization of relativistic systems with constraints. *Physics Letters* B55: 224.

Green MB, Schwarz JH, and Witten E (1987) *Superstring Theory*, vols. 1 and 2. Cambridge: Cambridge University Press.

Henneaux M (1998) Consistent interactions between gauge fields: the cohomological approach. *Contemporary Mathematics* 219: 93.

Henneaux M and Teitelboim C (1992) *Quantization of Gauge Systems*. Princeton: Princeton University Press.

Kato M and Ogawa K (1983) Covariant quantization of strings based on BRS invariance. *Nuclear Physics* B212: 443.

Kugo T and Ojima I (1979) Local covariant operator formalism of nonabelian gauge theories and quark confinement problem. *Progress of Theoretical Physics (Suppl.)* 66: 1.

Polchinski J (1998) *String Theory*,vols. 1 and 2. Cambridge: Cambridge University Press.

Stasheff JD (1998) The (secret?) homological algebra of the Batalin–Vilkovisky approach. *Contemporary Mathematics* 219: 195.

Tyutin IV (1975) Gauge invariance in field theory and statistical physics in the operator formalism. Preprint Lebedev-75-39.

Chern–Simons Models: Rigorous Results

A N Sengupta, Louisiana State University, Baton Rouge, LA, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The relationship between topological invariants and functional integrals from quantum Chern–Simons theory discovered by Witten (1989) raised several challenges for mathematicians. Most of the tremendous amount of mathematical activity generated by Witten's discovery has been concerned primarily with issues that arise after one has accepted the functional integral as a formal object. This has left, as an important challenge, the task of giving rigorous meaning to the functional integrals themselves and to rigorously derive their relation to topological invariants. The present article will discuss efforts to put the functional integral itself on a rigorous basis.

Notes:

Chern–Simons Functional Integrals

We shall describe here the typical Chern–Simons functional integral. For the purposes of this article, we will confine ourselves to a simpler setting rather than the most general possible one. In fact, we shall work with fields over three-dimensional Euclidean space $\mathbb{R}^3$ (instead of a general 3-manifold).

The typical Chern–Simons functional integral is of the form

$$\int_{\mathcal{A}} e^{i(k/4\pi)S_{CS}(A)} W_{C_1,R_1}(A) \dots W_{C_n,R_n}(A) DA \qquad [1]$$

Our objective in this section will be to specify what the terms in this formal integral mean. Very briefly, the integration is with respect to a formal "Lebesgue measure" on $\mathcal{A}$, an infinite-dimensional space of geometric objects A called connections over $\mathbb{R}^3$ with values in the Lie algebra LG of a group G. In the first term in the integrand, in the exponent, k is a real number, and $S_{CS}(A)$ is the Chern–Simons action for the connection A. Each term $W_{C_i,R_i}(A)$ is a Wilson loop observable, the trace in some representation R_i of the holonomy of the connection A around the loop C_i. The entire integral, formal though it may be, provides an invariant associated with the system of loops $C_1, \dots, C_n$.

Let G be a compact Lie group; for ease of exposition, let us take G to be a closed, connected subgroup of $U(n)$. Thus, each element of G is an $n \times n$ complex matrix g with $g^*g = I$, the identity. The Lie algebra LG consists of all $n \times n$ matrices A which are skew-Hermitian, that is, satisfy $A^* = -A$, and for which $e^{tA} \in G$ for all real numbers t. On LG there is a convenient inner product given by

$$\langle A, B \rangle = \mathrm{tr}(AB^*)$$

This inner product is invariant under the conjugation action of the group G on its Lie algebra LG.

By a connection over $\mathbb{R}^3$ we shall mean a C^∞ 1-form with values in LG. The set of all connections is an affine (in our case, actually a linear) space $\mathcal{A}$. If $A \in \mathcal{A}$, then define

$$S_{CS}(A) = \int_{\mathbb{R}^3} \mathrm{tr}(A \wedge dA + \tfrac{2}{3} A \wedge A \wedge A) \qquad [2]$$

This is, up to constant multiple, the Chern–Simons action functional.

Let A be a connection and consider a piecewise smooth path

$$C : [0, 1] \to \mathbb{R}^3$$

With this one can associate a G-valued path $[0,1] \to G : t \mapsto g(t) \in G$ satisfying the differential equation

$$g'(t)g(t)^{-1} = -A(C'(t))$$

subject to the initial condition $g(0) = I$, the identity. The path $t \mapsto g(t)$ describes parallel transport along C by the connection A. If C is a loop then the final value $g(1)$ is the holonomy of A around C. If R is a representation of G on some finite-dimensional vector space then the trace of $R(g(1))$ is the Wilson loop observable:

$$W_{C,R}(A) = \mathrm{tr}(R(g(1))) \qquad [3]$$

Thus, we have specified the meaning of the terms appearing in the formal integral [1], where $C_1, \dots, C_n$ of eqn [1] form a link (a family of nonintersecting, imbedded loops) in $\mathbb{R}^3$ and $R_1, \dots, R_n$ are finite-dimensional representations of G. Witten showed that, at least for suitable values of k, integrals of this form ought to produce topological invariants, which he identified, for the link.

The integral [1] is problematic for several reasons. First, there is no reasonable and useful analog of Lebesgue measure on an infinite-dimensional space. Even if one were to regularize this measure in some simple way, one would run into the problem that the measure would not live on the space of smooth connections, and so the integrand would become meaningless.

There are several different approaches to a mathematical interpretation of [1]. The approach that is often taken in practice is to simply ignore the analytical problem and define the value of the integral [1] to be what Witten's calculations have given. One approach, used, for instance, by Bar-Natan (1995) is to expand the integrand in a series and relate each individual integral in this expansion separately to topological invariants. Discrete approximation procedures to the continuum integral have also been explored. In the abelian case, infinite-dimensional oscillatory integral techniques have been used to understand the functional integral. Fröhlich and King (1999) showed the possibility of interpreting parallel transport using ideas from stochastic differential equations. Such an approach has been used successfully in the case of two-dimensional Yang–Mills theory, where the functional integral actually corresponds to integration with respect to a measure. In this article, we focus on a method of understanding the normalized Chern–Simons functional integral in terms of infinite-dimensional distribution theory and examining some ideas for understanding Wilson loop expectation values in this setting.

Infinite Dimensional Distributions

Let (x^0, x^1, x^2) denote the usual coordinates on $\mathbb{R}^3$. Gauge symmetry, an issue which will not be examined here, may be used to simplify the problem of the Chern–Simons integral. In particular, one

need only focus on connections which vanish in the x^2-direction, that is, connections of the form $A = A_0 dx^0 + A_1 dx^1$. For such A, the triple wedge-product term in the Chern–Simons action disappears, and we are left with the quadratic expression:

$$S_{CS}(A) = \int_{\mathbb{R}^3} \mathrm{tr}(A \wedge dA) \qquad [4]$$

This is good, since the functional integral now involves a quadratic exponent and so stands a good chance of rigorous realization, just as Gaussian measure can be given rigorous meaning in infinite dimensions. However, in the Chern–Simons situation, there is no hope of actually getting a measure, not even a complex measure.

The next best thing to a measure is a distribution or "generalized function." A distribution over a space Y is a continuous linear functional on a topological vector space of functions on Y. Thus, the objective is to realize the Chern–Simons functional integral as a continuous linear functional on some space of test functions over $\mathcal{A}$ (more precisely, on an extension of $\mathcal{A}$). Before turning to the specific case of the Chern–Simons integral, let us examine some elements of the theory of infinite-dimensional distributions, in as much as they are relevant to our needs.

Let us consider a Hilbert space $\mathcal{E}_0$, and a positive Hilbert–Schmidt operator T on $\mathcal{E}_0$. For each integer $p \geq 0$, let $\mathcal{E}_p = T^p(\mathcal{E}_0)$, which is a Hilbert space with the inner product $\langle x, y\rangle_p = \langle T^{-p}x, T^{-p}y\rangle$. Then we have the chain of inclusions

$$\mathcal{E} = \bigcap_{p\geq 1} \mathcal{E}_p \subset \cdots \mathcal{E}_2 \subset \mathcal{E}_1 \subset \mathcal{E}_0 \qquad [5]$$

with each inclusion $\mathcal{E}_{p+1} \to \mathcal{E}_p$ being Hilbert–Schmidt. Let $\mathcal{E}_{-p} = \mathcal{E}_p'$ be the topological dual of $\mathcal{E}_p$, the space of continuous linear functionals on $\mathcal{E}_p$, and let $\mathcal{E}'$ be the topological dual of $\mathcal{E}$, where the latter is given the topology generated by all the norms $\|\cdot\|_p$. Then we have the inclusions

$$\mathcal{E}_0 \simeq \mathcal{E}_0' \subset \mathcal{E}_{-1} \subset \mathcal{E}_{-2} \subset \cdots \subset \mathcal{E}' = \bigcup_{p\geq 0} \mathcal{E}_{-p} \qquad [6]$$

For each $x \in \mathcal{E}$ there is the evaluation map $\hat{x}: \mathcal{E}' \to \mathbb{R} : \phi \mapsto \phi(x)$. A very special case of a general theorem of Minlos guarantees that on the dual $\mathcal{E}'$ there is a measure μ on the sigma algeba generated by all the functions $\hat{x}$ such that each $\hat{x}$ is a Gaussian random variable of mean zero and variance $|x|_0^2$, that is,

$$\int_{\mathcal{E}'} e^{it\hat{x}}\, d\mu = e^{-t^2|x|_0^2/2}$$

for all $x \in \mathcal{E}$ and $t \in \mathbb{R}$. This measure μ is the standard Gaussian measure on $\mathcal{E}'$ for the infinite-dimensional nuclear space $\mathcal{E}$.

The inner products $\langle \cdot, \cdot \rangle_p$ give rise to a nuclear space structure on function spaces over $\mathcal{E}$. Let $\mathcal{U}$ be the algebra of functions on $\mathcal{E}'$ generated by the exponentials $e^{\lambda\hat{x}}$, with x running over $\mathcal{E}$ and λ over $\mathbb{C}$. For each $p \geq 0$, there is an inner product $\langle\langle \cdot, \cdot \rangle\rangle_p$ on $\mathcal{U}$ such that

$$\left\langle\left\langle e^{\lambda\hat{x} - \lambda^2|x|_p^2/2}, e^{\mu\hat{y} - \mu^2|y|_p^2/2}\right\rangle\right\rangle_p = e^{\lambda\bar{\mu}\langle x, y\rangle_p} \qquad [7]$$

For $p = 0$ the left-hand side coincides with the $L^2(\mu)$ inner product. Let $[\mathcal{E}]_p$ be the Hilbert space completion of $\mathcal{U}$ in the $\langle\langle \cdot, \cdot \rangle\rangle_p$ inner product. Then

$$\cdots [\mathcal{E}]_3 \subset [\mathcal{E}]_2 \subset [\mathcal{E}]_1 \subset [\mathcal{E}]_0 = L^2(\mathcal{E}', \mu) \qquad [8]$$

Let $[\mathcal{E}] = \cap_{p\geq 0} [\mathcal{E}]_p$, equipped with topology from all the norms $\|\cdot\|_p$, and $[\mathcal{E}]'$ its topological dual. Elements of $[\mathcal{E}]'$, being continuous linear functionals on the "test function space" $[\mathcal{E}]$, are called distributions over $\mathcal{E}$, in the language of white-noise analysis.

A fundamental tool in the study of infinite-dimensional distributions is the S-transform. This generalizes the traditional Segal–Bargmann transform from the L^2-setting to the context of distributions. Let $\mathcal{E}_c$ be the complexification of $\mathcal{E}$. The inner product $\langle \cdot, \cdot \rangle_0$ on $\mathcal{E}$ extends to a complex-bilinear pairing $\mathcal{E}_c \times \mathcal{E}_c \to \mathbb{C} : (z, w) \mapsto z \cdot w$. The evaluation pairing $\mathcal{E}' \times \mathcal{E} \to \mathbb{R}$ also extends naturally to the complexifications. For Φ a distribution belonging to $[\mathcal{E}]'$, define a function $S\Phi$ on $\mathcal{E}$ by

$$S\Phi(z) = \Phi(c_z)$$

for all $z \in \mathcal{E}_c$. Here c_z is the coherent state function on $\mathcal{E}'$ given by $c_z(\phi) = e^{\phi(z) - (1/2)z\cdot z}$. A fundamental and useful result in white-noise analysis, due originally to Potthoff and Streit, specifies the range of the transform S and allows reconstruction of a distribution Φ from the function $S\Phi$. Briefly, the range of S consists of functions which are holomorphic, in an appropriate sense, and have at most quadratic exponential growth. In particular, this theorem implies that a function of the form $z \mapsto e^{az\cdot z}$, for any constant a, is in the range of Φ.

Rigorous Realization of Chern–Simons Integrals

We return to the Chern–Simons context. As mentioned earlier, gauge symmetry may be invoked to reduce the space of connections to the smaller space:

$$\mathcal{E} = X \oplus X \qquad [9]$$

where $X = \mathcal{S}(\mathbb{R}^3) \otimes LG$ is the space of rapidly decreasing functions with values in the Lie algebra LG. Let

$$T_1 = \left(-\frac{d^2}{dx^2} + \frac{x^2}{4}\right)^{-1}$$

as a linear operator on $L^2(\mathbb{R}^3)$, $T_2 = T_1^{\otimes 3} \otimes I$ the induced operator on $L^2(\mathbb{R}^3) \otimes LG$, and $T = T_2 \oplus T_2$. Then, as described in the preceding section, we have the space $\mathcal{E}$ and its dual $\mathcal{E}'$. There is then the standard Gaussian measure μ on $\mathcal{E}'$, and the space $[\mathcal{E}]'$ of distributions over $\mathcal{E}'$.

The normalized Chern–Simons integral may be viewed as a linear functional

$$\Phi_{CS} : F \mapsto \frac{1}{N} \int_{\mathcal{E}} e^{i(k/4\pi)S_{CS}(A)} F(A) DA \qquad [10]$$

where N is a "normalizing" factor. Rigorous meaning can be given to this by first formally working out what the S-transform of Φ_{CS} ought to be. Calculation shows that $S\Phi$ is indeed a holomorphic function on $\mathcal{E}_c$ of quadratic growth. The Potthoff–Streit theorem then implies that Φ_{CS} does exist as a distribution in the space $[\mathcal{E}]'$. Let us examine this in some more detail.

As before, we take A to be of the form $A = A_0 dx^0 + A_1 dx^1$, with the component A_2 equal to 0. Integration by parts shows that

$$\frac{k}{4\pi} S_{CS}(A) = -\frac{k}{2\pi} \int_{\mathbb{R}^3} \mathrm{tr}(A_0 \partial_2 A_1) \, \mathrm{dvol} \qquad [11]$$

A formal computation reveals that $S(\Phi_{CS})(j)$ should be given by

$$\exp\left(\frac{2\pi i}{k} \mathrm{tr}(j_0 \partial_2^{-1} j_1)\right) \qquad [12]$$

where $j = (j_0, j_1)$, and

$$\partial_2^{-1} f(x) = \frac{1}{2} \int ds [1_{(-\infty, x_2]}(s) - 1_{[x_2, \infty)}(s)] \, f(x^0, x^1, s)$$

The Potthoff–Streit criterion implies the existence of a distribution Φ_{CS}, whose S-transform is given by the above expression.

The distribution Φ_{CS} is, however, not a sufficiently powerful object to allow determination of the Wilson loop expectations that one would really like to have. For instance, Φ_{CS} does not live on the space of smooth connections and so the meaning of parallel transport needs to be defined. The state of knowledge, at the rigorous level, at this point is still evolving, with progress reported by A. Hahn. We describe some ideas for the Wilson loop expectations in the following.

The strategy for defining parallel transport along a path is to smear out the path by means of bump functions and essentially replace the path by a path of test functions in $\mathcal{E}$. The description given here is mainly for the case of abelian G. Choose first a C^∞ non-negative bump function ψ on $\mathbb{R}^3$, vanishing outside the unit ball and having L^1 norm equal to 1. For $\epsilon > 0$, let ψ^ϵ be the scaled bump function given by $\psi^\epsilon(x) = \epsilon^{-3} \psi(x/\epsilon)$. Next, for a smooth loop $[0, 1] \to l(t) = (l_0(t), l_1(t), l_2(t))$, let $l^\epsilon(t) = \psi^\epsilon(\cdot - l(t))$, the scaled bump function centered now at the path point $l(t)$. Now consider a generalized connection $A = (A_0, A_1) \in \mathcal{E}'$. Set

$$B_A^{l^\epsilon}(t) = A_0(l^\epsilon(t)) l'(t)_0 + A_1(l^\epsilon(t)) l'(t)_1 \qquad [13]$$

The equation of parallel transport can be reformulated as a differential equation for a matrix-valued path $t \mapsto P_A^{l^\epsilon}(t)$ satisfying

$$\frac{d}{dt} P_A^{l^\epsilon}(t) + B_A^{l^\epsilon}(t) P_A^{l^\epsilon}(t) = 0 \qquad [14]$$

and the initial condition $P_A^{l^\epsilon}(t) = I$. With this smearing, one can consider functions of the form

$$W_\epsilon(L; A) = \prod_{i=1}^{n} \mathrm{tr}(P^{l_i^\epsilon}(A)) \qquad [15]$$

for a link L consisting of loops $l_1, \ldots, l_n$, instead of the classical Wilson loop variable.

At this stage, it would be natural to consider taking $\epsilon \downarrow 0$ in $\Phi(W_\epsilon(L))$. However, this is still problematic. A further regularization is needed, roughly corresponding to the geometric notion of framing. In the definition of Φ_{CS}, alteration is made to the quadratic form $Q(j, j)$ in the exponent which appears in the expression for $S(\Phi_{CS})$, replacing it with $Q(j, \phi_s^* j)$, where $\{\phi_s\}_{s>0}$ is a family of suitable diffeomorphisms of $\mathbb{R}^3$, with ϕ_0 being the identity. In a sense, this splits a single loop l into l and a neighboring loop $\phi_s \circ l$. At the end, one has to take $s \downarrow 0$. The resulting limiting value is the expected link-invariant. We shall not go into the case of nonabelian G, which is more complex, for which work continues to be in progress.

Infinite-dimensional distributions can be used to formulate a rigorous theory for normalized Chern–Simons functional integrals. The more specific questions raised by the Wilson-loop integrals in this setting opens up new problems for further developments in the distribution theory, connecting geometry, topology, and infinite-dimensional analysis.

Acknowledgment

This research is supported by US NSF grant DMS-0201683.

See also: BF Theories; Feynman Path Integrals; Fractional Quantum Hall Effect; Knot Theory and Physics; Large-*N* and Topological Strings; Large-*N* Dualities; Quantum 3-Manifold Invariants; Quantum Hall Effect; Spin Foams; String Field Theory; Topological Quantum Field Theory: Overview; Twistor Theory: Some Applications.

Further Reading

Albeverio S, Hahn A, and Sengupta AN (2003) Chern–Simons theory, Hida distributions, and state models. *Infinite Dimensional Analysis Quantum Probability and Related Topics* 6: 65–81.

Albeverio S and Schäfer J (1994) Abelian Chern–Simons theory and linking numbers via oscillatory integrals. *Journal of Mathematical Physics (N.Y.)* 36 (suppl. 5): 2135–2169.

Albeverio S and Sengupta A (1997) A mathematical construction of the non-Abelian Chern–Simons functional integral. *Communications in Mathematical Physics* 186: 563–579.

Altschuler D and Freidel L (1997) Vassiliev Knot invariants and Chern–Simons perturbation theory to all orders. *Communications in Mathematical Physics* 187: 261–287.

Atiyah M (1990) *The Geometry and Physics of Knot Polynomials.* Cambridge: Cambridge University Press.

Bar-Natan D (1995) Perturbative Chern–Simons theory. *Journal of Knot Theory and its Ramifications* 4: 503.

Chern S-S and Simons J (1974) Characteristic forms and geometric invariants. *Annals of Mathematics* 99: 48–69.

Fröhlich J and King C (1989) The Chern–Simons theory and Knot polynomials. *Communications in Mathematical Physics* 126: 167–199.

Kondratiev Yu, Leukert P, Potthoff J, Streit L, and Westerkamp W (1996) Generalized functionals in Gaussian spaces – the characterization theorem revisited. *Journal of Functional Analysis* 141 (suppl. 2): 301–318.

Kuo H-H (1996) *White Noise Distribution Theory.* Boca Raton, FL: CRC Press.

Landsman NP, Pflaum M, and Schlichenmaier M (2001) *Quantization of Singular Symplectic Quotients.* Basel–Boston–Berlin: Birkhäuser.

Leukert P and Schäfer J (1996) A rigorous construction of Abelian Chern–Simons path integrals using White Noise analysis. *Rev. Math. Phys.* 8 (suppl. 3): 445–456.

Sen Samik, Sen Siddhartha, Sexton JC, and Adams DH (2000) Geometric discretization scheme applied to the Abelian Chern–Simons theory. *Physical Review E* 61: 3174–5185.

Simon B (1971) Distributions and their Hermite expansions. *Journal of Mathematical Physics (N.Y.)* 12: 140–148.

Witten E (1989) Quantum field theory and the Jones polynomial. *Communications in Mathematical Physics* 121: 351–399.

Dirac Fields in Gravitation and Nonabelian Gauge Theory

J A Smoller, University of Michigan, Ann Arbor, MI, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

In this article we describe some recent results (Finster *et al.* 1999a,b, 2000 a–c, 2002a) concerning the existence of both particle-like, and black hole solutions of the coupled Einstein–Dirac–Yang–Mills (EDYM) equations. We show that there are stable globally defined static, spherically symmetric solutions. We also show that for static black hole solutions, the Dirac wave function must vanish identically outside the event horizon. The latter result indicates that the Dirac particle (fermion) must either enter the black hole or tend to infinity.

The plan of the article is as follows. The next section describes the background material. It is followed by a discussion of the coupled EDYM equations for static, spherically symmetric particle-like and black hole solutions. The final section of the article is devoted to a discussion of these results.

Background Material

Einstein's Equations

We begin by describing the Einstein equation for the gravitational field (for more details, see, e.g., Adler *et al.* (1975)). We first note Einstein's hypotheses of general relativity (GR):

(E1) The gravitational field is the metric g_{ij} in $3+1$ spacetime dimensions. The metric is assumed to be symmetric.
(E2) At each point in spacetime, the metric can be diagonalized as diag(−1,1,1,1).
(E3) The equations which describe the gravitational field should be covariant; that is, independent of the choice of coordinate system.

The hypothesis (E1) is Einstein's brilliant insight, whereby he "geometrizes" the gravitational field. (E2) means that there are inertial frames at each point (but not globally), and guarantees that special relativity (SR) is included in GR, while (E3) implies that the gravitational field equations must be tensor equations; that is, coordinates are an artifact, and physics should not depend on the choice of coordinates.

Einstein's Equations of GR

The metric $g_{ij}=g_{ij}(x), i,j=0,1,2,3, x=(x^0,x^1,x^2,x^3)$, $x^0=ct$ (c=speed of light, t=time), is the metric tensor defined on four-dimensional spacetime. Einstein's equations are ten (tensor) equations for the unknown metric g_{ij} (gravitational field), and take the form

$$R_{ij} - \tfrac{1}{2}Rg_{ij} = \sigma T_{ij} \qquad [1]$$

where the left-hand side $G_{ij} = R_{ij} - \frac{1}{2}Rg_{ij}$ is the Einstein tensor and depends only on the geometry, $\sigma = 8\pi G/c^4$, where G is Newton's gravitational constant, while T_{ij}, the energy–momentum tensor, represents the source of the gravitational field, and encodes the distribution of matter. (The word "matter" in GR refers to everything which can produce a gravitational field, including elementary particles, electromagnetic or Yang–Mills (YM) fields. From the Bianchi identities in geometry (cf. Adler *et al.* (1975)), the (covariant) divergence of the Einstein tensor, G_{ij}, vanishes identically, namely

$$G^j_{i,j} = 0$$

so, on solutions of Einstein's equations,

$$T^j_{i,j} = 0$$

and this in turn expresses the conservation of energy and momentum. The quantities which comprise the Einstein tensor are given as follows: first, from the metric tensor g_{ij}, we form the Levi-Civita connection Γ^k_{ij} defined by:

$$\Gamma^k_{ij} = \frac{1}{2}g^{k\ell}\left(\frac{\partial g_{\ell j}}{\partial x^i} - \frac{\partial g_{i\ell}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^\ell}\right)$$

where (4×4 matrix) $[g^{k\ell}] = [g_{k\ell}]^{-1}$, and summation convention is employed; namely, an index which appears as both a subscript and a superscript is to be summed from 0 to 3. With the aid of Γ^k_{ij}, we can

Notes:

construct the celebrated Riemann curvature tensor $R^i_{qk\ell}$:

$$R^i_{qk\ell} = \frac{\partial \Gamma^i_{q\ell}}{\partial x^k} - \frac{\partial \Gamma^i_{qk}}{\partial x^\ell} + \Gamma^i_{pk}\Gamma^p_{q\ell} - \Gamma^i_{p\ell}\Gamma^p_{qk}$$

Finally, the terms R_{ij} and R which appear in the Einstein tensor G_{ij} are given by

$$R_{ij} = R^s_{isj}$$

(the Ricci tensor), and

$$R = g^{ij}R_{ij}$$

is the scalar curvature.

From the above definitions, one sees at once the enormous complexity of the Einstein equations. For this reason, one usually seeks solutions which have a high degree of symmetry, and in what follows, in this section, we shall only consider static, spherically symmetric solutions; that is, solutions which depend only on $r = |x| = \sqrt{(x^1)^2 + (x^2)^2 + (x^3)^2}$. In this case, the metric g_{ij} takes the form

$$ds^2 = -T(r)^2 dt^2 + A(r)^{-1} dr^2 + r^2 d\Omega^2 \qquad [2]$$

where $d\Omega^2 = d\theta^2 + \sin^2\theta\, d\varphi^2$ is the standard metric on the unit 2-sphere, r,θ,φ are the usual spherical coordinates, and t denotes time.

Black Hole Solutions

Consider the problem of finding the gravitational field outside a ball of mass M in $\mathbb{R}^3$; that is, there is no matter exterior to the ball. Solving Einstein's equations $G_{ij}' = 0$ gives the famous Schwarzschild solution (1916):

$$ds^2 = -\left(1 - \frac{2m}{r}\right)c^2 dt^2 + \left(1 - \frac{2m}{r}\right)^{-1} dr^2 + r^2 d\Omega^2 \qquad [3]$$

where $m = GM/c^2$. Since $2m$ has the dimensions of length, it is called the Schwarzschild radius. Observe that when $r = 2m$, the metric is singular; namely, $g_{tt} = 0$ and $g_{rr} = \infty$. By transforming the metric [2] to the so-called Kruskal coordinates (cf. Adler *et al.* (1975)), one observes that the Schwarzschild sphere $r = 2m$ has the physical characteristics of a black hole: light and nearby particles can enter the region $r < 2m$, nothing can exit this region, and there is an intrinsic (nonremovable) singularity at the center $r = 0$.

For the general metric [2], we define a black hole solution of Einstein's equations to be a solution which satisfies, for some $\rho > 0$,

$$A(\rho) = 0, \quad A(r) > 0 \text{ if } r > \rho$$

ρ is called the radius of the black hole, or the event horizon.

Yang–Mills Equations

The YM equations generalize Maxwell's equations. To see how this comes about, we first write Maxwell's equations in an invariant way. Thus, let A denote a scalar-valued 1-form:

$$A = A_i dx^i, \quad A_i \in \mathbb{R}$$

which is called the electromagnetic potential (by physicists), or a connection (by geometers). The electromagnetic field (curvature) is the 2-form

$$F = dA$$

In local coordinates,

$$F = F_{ij} dx^i \wedge dx^j, \quad F_{ij} = \frac{\partial A_j}{\partial x^i} - \frac{\partial A_i}{\partial x^j}$$

In this framework, Maxwell's equations are given by

$$d^*F = 0, \qquad dF = 0 \qquad [4]$$

where $\star$ is the Hodge star operator, mapping 2-forms to 2-forms (in $\mathbb{R}^4$), and is defined by

$$(^*F)_{k\ell} = \tfrac{1}{2}\sqrt{|g|}\,\varepsilon_{ijk\ell}F^{ij}$$

where $g = \det(g_{ij})$ and $\varepsilon_{ijk\ell}$ is the completely antisymmetric symbol defined by $\varepsilon_{ijk\ell} = \operatorname{sgn}(ijk\ell)$. As usual, indices are raised (or lowered) via the metric, so that, for example,

$$F^{ij} = g^{\ell i} g^{mj} F_{jm}$$

It is important to notice that $^*F$ depends on the metric. Note also that Maxwell's equations are linear equations for the A_i's.

The YM equations generalize Maxwell's equations and can be described as follows. With each YM field (described below) is associated a compact Lie group G called the gauge group. For such G, we denote its Lie algebra by $\mathfrak{g}$, defined to be the tangent space at the identity of G. Now let A be a $\mathfrak{g}$-valued 1-form

$$A = A_i dx^i$$

where each A_i is in $\mathfrak{g}$. In this case, the curvature 2-form is defined by

$$F = dA + A \wedge A$$

or, in local coordinates,

$$F_{ij} = \frac{\partial A_j}{\partial x^i} - \frac{\partial A_i}{\partial x^j} + [A_i, A_j]$$

The commutator $[A_i, A_j] = 0$ if G is an abelian group, but is generally nonzero if G is a matrix group. In this framework, the YM equations can be written in the form $\mathrm{d}^* F = 0$, where now d is an appropriately defined covariant exterior derivative. For Maxwell's equations, the gauge group $G = \mathrm{U}(1)$ (the circle group $\{e^{i\theta}: \theta \in \mathbb{R}\}$) so $\mathfrak{g}$ is abelian and we recover Maxwell's equations from the YM equations. Observe that if G is nonabelian, then the YM equations $\mathrm{d}^* F = 0$ are nonlinear equations for the connection coefficients A_i.

The Dirac Equation in Curved Spacetime

The Dirac equation is a generalization of Schrödinger's equation, in a relativistic setting (Bjorken and Drell 1964). It thus combines quantum mechanics with the theory of relativity. In addition, the Dirac equation also describes the intrinsic "spin" of fermions and, for this reason, solutions of the Dirac equation are often called spinors.

The Dirac equation can be written as

$$(G - m)\Psi = 0 \qquad [5]$$

where G is the Dirac operator, m is the mass of the Dirac particle (fermion), and Ψ is a complex-valued 4-vector called the wave function, or spinor. The Dirac operator G is of the form

$$G = \mathrm{i} G^j(x) \frac{\partial}{\partial x^j} + B(x) \qquad [6]$$

where G^j as well as B are 4×4 matrices, m is the (rest) mass of the fermion, and $\mathrm{i} = \sqrt{-1}$. The Dirac equation is thus a linear equation for the spinors. The G^j (called Dirac matrices) and the Lorentzian metric g_{ij} are related by

$$g^{jk} I = \tfrac{1}{2}\{G^j, G^k\} \qquad [7]$$

where $\{G^j, G^k\}$ is the anticommutator

$$\{G^j, G^k\} = G^j G^k + G^k G^j$$

Thus, the Dirac matrices depend on the underlying metric in four-dimensional spacetime.

Suppose that H is a spacelike hypersurface in $\mathbb{R}^4$, with future-directed normal vector $\nu = \nu(x)$, and let $\mathrm{d}\mu$ be the invariant measure on H induced by the metric g_{ij}. We define a scalar product on solutions Ψ, Φ of the Dirac equation by

$$\langle \Psi | \Phi \rangle = \int_H \bar{\Psi} G^j \Phi \nu_j \, \mathrm{d}\mu \qquad [8]$$

This scalar product is positive definite, and because of current conservation (cf. Finster (1988))

$$\nabla_j \bar{\Psi} G^j \Phi = 0$$

it is also independent of H. By generalizing the expression (due to Dirac), $\bar{\Psi}\gamma^0\Psi = |\Psi|^2$, in Minkowski space, where γ^0 and $\bar{\Psi}$, the adjoint spinors, are defined by

$$\gamma^0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \bar{\Psi} = \Psi^* \gamma^0$$

where $*$ denotes complex conjugation, and 1 is the 2×2 identity matrix, the quantity $\bar{\Psi} G^j \Psi \nu_j$ is interpreted as the probability density of the Dirac particle. We normalize solutions of the Dirac equation by requiring

$$\langle \Psi | \Psi \rangle = 1 \qquad [9]$$

Spherically Symmetric EDYM Equations

In the remainder of this article we assume that all fields are spherically symmetric, so they depend only on the variable $r = |x|$. In this case, the Lorentzian metric in polar coordinates (t, r, θ, φ) takes the form [2]. The Dirac wave function can be (Finster *et al.* 2000b) described by two real functions, $(\alpha(r), \beta(r))$, and the potential $W(r)$ corresponds to the magnetic component of an SU(2) YM field. As shown in Finster *et al.* (2000b), the EDYM equations are

$$\sqrt{A}\alpha' = \frac{w}{r}\alpha - (m + \omega T)\beta \qquad [10]$$

$$\sqrt{A}\beta' = (-m + \omega T)\alpha - \frac{w}{r}\beta \qquad [11]$$

$$rA' = 1 - A - \frac{1}{e^2}\frac{(1 - w^2)^2}{r^2} - 2\omega T^2(\alpha^2 + \beta^2) - \frac{2}{e^2} A w'^2 \qquad [12]$$

$$2rA'\frac{T'}{T} = -1 + A + \frac{1}{e^2}\frac{(1 - w^2)^2}{r^2} + 2mT(\alpha^2 - \beta^2) - 2\omega T^2(\alpha^2 + \beta^2) + 4\frac{T}{r} w\alpha\beta - \frac{2}{e^2} A w'^2 \qquad [13]$$

$$rAw'' = -(1 - w^2)w + e^2 rT\alpha\beta - r^2 \frac{A'T - 2AT'}{2T} w' \qquad [14]$$

Equations [10] and [11] are the Dirac equations, [12] and [13] are the Einstein equations, and [14] is

the YM equation. The constants m, ω, and e denote, respectively, the rest mass of the Dirac particle, its energy, and the YM coupling constant.

Nonexistence of Black Hole Solutions

Let the surface $r = \rho > 0$ represent a black hole event horizon:

$$A(\rho) = 0, \quad A(r) > 0 \text{ if } r > \rho \qquad [15]$$

In this case, the normalization condition [9] is replaced by

$$\int_{r_0}^{\infty} (\alpha^2 + \beta^2) \frac{\sqrt{T}}{A} \, dr < \infty, \quad \text{for every } r_0 > \rho \qquad [16]$$

In addition, we assume that the following global conditions hold:

$$\lim_{r \to \infty} r(1 - A(r)) = M < \infty \qquad [17]$$

(finite mass),

$$\lim_{r \to \infty} T(r) = 1 \qquad [18]$$

(gravitational field is asymptotically flat Minkowskian), and

$$\lim_{r \to \infty} (w(r)^2, w'(r)) = (1, 0) \qquad [19]$$

(the YM field is well behaved).

Concerning the event horizon $r = \rho$, we make the following regularity assumptions:

1. The volume element $\sqrt{|\det g_{ij}|} = |\sin\theta| r^2 A^{-1} T^{-2}$ is smooth and nonzero on the horizon; that is,

$$T^{-2} A^{-1}, \; T^2 A \in C^1([\rho, \infty))$$

2. The strength of the YM field F_{ij} is given by

$$\mathrm{tr}(F_{ij} F^{ij}) = \frac{2Aw'^2}{r^4} + \frac{(1 - w^2)^2}{r^4}$$

(cf. Bartnik and McKinnon 1988). We assume that this scalar is bounded near the horizon; that is, outside the event horizon and near $r = \rho$, assume that

$$w \text{ and } Aw'^2 \text{ are bounded} \qquad [20]$$

3. The function $A(r)$ is monotone increasing outside of and near the event horizon.

As discussed in Finster *et al.* (1999a), if assumption 1 or 2 were violated, then an observer freely falling into the black hole would feel strong forces when crossing the horizon. Assumption 3 is considerably weaker than the corresponding assumption in Finster *et al.* (1999b), where, indeed, it was assumed that the function $A(r)$ obeyed a power law $A(r) = c(r - \rho)^s + \mathcal{O}((r - \rho)^{s+1})$, with positive constants c and s, for $r > \rho$.

The main result in this subsection is the following theorem:

Theorem 1 *Every black hole solution of the EDYM equations* [10]–[14] *satisfying the regularity conditions 1–3 cannot be normalized and coincides with a Bartnik–McKinnon (BM) black hole of the corresponding Einstein–Yang–Mills (EYM) equations; that is, the spinors α and β must vanish identically outside the event horizon.*

Remark Smoller and Wasserman (1998) proved that any black hole solution of the EYM equations that has finite mass (i.e., that satisfies [17]) must be one of the BM black hole solutions (Bartnik and McKinnon 1988) whose existence was first demonstrated in Smoller *et al.* (1993). Thus, amending the EYM equations by taking quantum-mechanical effects into account – in the sense that both the gravitational and YM fields can interact with Dirac particles – does not yield any new types of black hole solutions.

The present strategy in proving this theorem is to assume that we have a black hole solution of the EDYM equations [10]–[18] satisfying assumptions 1–3, where the spinors do not vanish identically outside of the black hole. We shall show that this leads to a contradiction. The proof is broken up into two cases: either $A^{-1/2}$ is integrable or nonintegrable near the event horizon. We shall only discuss the proof for the case when $A^{-1/2}$ is integrable near the event horizon, leaving the alternate case for the reader to view in Finster *et al.* (2000a).

If $A^{-1/2}$ is integrable, then one shows that there are positive constants c, ε such that

$$c \leq \alpha^2(r) + \beta^2(r) \leq \frac{1}{c}, \quad \text{if} \quad \rho < r < \rho + \varepsilon \qquad [21]$$

Indeed, multiplying [10] by α, and [11] by β and adding gives an estimate of the form

$$\sqrt{A}(\alpha^2 + \beta^2)' \leq \kappa(\alpha^2 + \beta^2)$$

Upon dividing by $\sqrt{A}(\alpha^2 + \beta^2)$ and integrating from $r > \rho$ to $\rho + \varepsilon$ gives

$$|\log(\alpha^2 + \beta^2)(\rho + \varepsilon) - \log(\alpha^2 + \beta^2)(r)| \leq \text{const.}$$

from which the desired result follows. Next, from [12] and [13],

$$r(AT^2)' = 4 - \omega T^4(\alpha^2 + \beta^2) + T^3\left[2m(\alpha^2 - \beta^2) + \frac{4w}{r}\alpha\beta\right] - \frac{4}{e^2}(Aw')^2T^2 \qquad [22]$$

Using assumption 2 together with the last theorem, we see that the coefficients of T^4, T^3, and T^2 on the right-hand side of [21] are bounded near ρ, and from assumption 1 the left-hand side of [21] is bounded near ρ. Since assumption 1 implies $T(r) \to \infty$ as $r \searrow \rho$, we see that $\omega = 0$. Since $\omega = 0$, the Dirac equations simplify and we can show that $\alpha\beta$ is a positive decreasing function which tends to 0 as $r \to \infty$. Then the YM equation can be written in the form

$$r^2(Aw')' = -w(1-w)^2 + e^2\frac{r(T\sqrt{A})\alpha\beta}{\sqrt{A}} + r^2\frac{(AT^2)'}{2AT^2}(Aw') \qquad [23]$$

From assumption 2, Aw'^2 is bounded so $A^2w'^2 \to 0$ as $r \searrow \rho$. Thus, from [22] we can write, for r near ρ,

$$(Aw')'(r) \geq c_1 + \frac{c_2}{\sqrt{A(r)}}$$

where c_1 and c_2 are positive constants. Using this inequality, we can show that for r near ρ,

$$A(r) = (r - \rho)B(r)$$

where $0 < \lim_{r\searrow\rho} B(r) < \infty$. It follows that $A(\rho) = 0$ and $A'(\rho) > 0$. Thus, the Einstein metric has the same qualitative features as the Schwarzschild metric near the event horizon. Hence, the metric singularity can be removed via a Kruskal transformation (Adler *et al.* 1975). In these Kruskal coordinates, the YM potential is continuous and bounded (as is easily verified). As a consequence, the arguments in Finster *et al.* (2000c) go through and show that the spinors must vanish identically outside the horizon. For this, one must note that continuous zero-order terms in the Dirac operator are irrelevant for the derivation of the matching conditions in Finster *et al.* 2000c, section 2.4). Thus, the matching conditions (equations (2.31), (2.34) of Finster *et al.* (2000c)) are valid without changes in the presence of our YM field. Using conservation of the (electromagnetic) Dirac current and its positivity in timelike directions, the arguments in Finster *et al.* (2000c, section 4) all carry over. This completes the proof.

We have thus proved that the only black hole solutions of our EDYM equations are the BM black holes; that is, the spinors must vanish identically. In other words, the EDYM equations do not admit normalizable black hole solutions. Thus, in the presence of quantum-mechanical Dirac particles, static and spherically symmetric black hole solutions do not exist. Another interpretation of these our result is that Dirac particles can only either disappear into the black hole or escape to infinity. These results were proved under very weak regularity assumptions on the form of the event horizon (see assumptions 1–3).

Particle-Like Solutions

By a particle-like (bound state) solution of the (SU(2)) EDYM equations, we mean a smooth solution of eqns [10]–[14], which is defined for all $r \geq 0$, and satisfies condition [9], which explicitly becomes

$$\int_0^r (\alpha^2 + \beta^2)\frac{\sqrt{T}}{A}\,dr = 1 \qquad [24]$$

In addition, we demand that [17]–[19] also hold. It is easily shown that, near $r = 0$, we must have

$$w(r) = 1 - \frac{\lambda}{2}r^2 + \mathcal{O}(r^2) \qquad [25]$$

where λ is a real parameter. From this, via a Taylor expansion, one finds that

$$\alpha(r) = \alpha_1 r + \mathcal{O}(r^3)$$
$$\beta(r) = \tfrac{1}{2}(\omega T_0 - m)\alpha_1 r^2 + \mathcal{O}(r^3) \qquad [26]$$

$$A(r) = 1 + \mathcal{O}(r^2), \quad T(r) = T_0 + \mathcal{O}(r^2) \qquad [27]$$

with two parameters α_1 and $T_0 > 0$. Using linearity of the Dirac equation, we can always assume that $\alpha_1 > 0$.

Under all realistic conditions, the coupling of Dirac particles to the YM field (describing the weak or strong interactions) is much stronger than the coupling to the gravitational field. Thus, we are particularly intrested in the case of weak gravitational coupling. As shown in Finster *et al.* (2000b), the gravitational field is essential for the formation of bound states. However, for arbitrarily weak gravitational coupling, we can hope to find bound states. It is even conceivable that these bound-state solutions might have a well-defined limit when the gravitational coupling tends to zero, if we let the YM coupling go to infinity at the same time. Our idea is that this limiting case might yield a system of equations which is simpler than the full EDYM system, and can thus serve as a physically interesting starting point for the analysis of the coupled

interactions described by the EDYM equations. Expressed in dimensionless quantities, we shall thus consider the limits

$$m^2\kappa \to 0 \quad \text{and} \quad e^2 \to \infty \tag{28}$$

That is, we ask whether weak gravitational coupling can give rise to bound states. Using numerical methods, we find particle-like solutions which are stable, even for arbitrarily weak gravitational coupling.

Now assuming that [27] holds (weak gravitational coupling), so that $(A, T) \approx (1, 1)$, then we find that the Dirac equations have a meaningful limit only under the assumptions that α converges and that

$$\begin{aligned} m\beta(r) &\to \hat{\beta}(r), \quad m^2(T(r)-1) \to \varphi \\ m(\omega - m) &\to E \end{aligned} \tag{29}$$

with two real functions $\hat{\beta}, \varphi$ and a real parameter E. Multiplying [29] with m and taking the limits [28] as well as $A, T \to 1$, the Dirac equations become

$$\alpha' = \frac{w}{r}\alpha - 2\hat{\beta} \tag{30}$$

$$\hat{\beta}' = (E + \varphi)\alpha - \frac{w}{r}\hat{\beta} \tag{31}$$

We next consider the YM equation [14]. The last term in [14] drops out in the limit of weak gravitational coupling [27]. The second summand converges only under the assumption that

$$\frac{e^2}{m} \to q \tag{32}$$

with q a real parameter, playing the role of an "effective" coupling constant. Together with [27], this implies that $m \to \infty$. The YM equations thus have the limit

$$r^2 w'' = -(1-w)^2 w + qr\alpha\hat{\beta} \tag{33}$$

In order to get a well-defined and nontrivial limit of the Einstein equations [13] and [14], we need to assume that the parameter $m^3\kappa$ has a finite, nonzero limit. Since this parameter has the dimension of inverse length, we can arrange by a scaling of our coordinates that

$$m^3\kappa \to 1 \tag{34}$$

We differentiate the T-equation [13] with respect to r and substitute [12]. Taking the limits [28] and [33], a straightforward calculation yields the equation

$$r^2\Delta\varphi = -\alpha^2 \tag{35}$$

where $\Delta = r^{-2}\partial_r(r^2\partial_r)$ is the radial Laplacian in Euclidean $\mathbb{R}^3$. Indeed, this equation can be regarded as Newton's equation with the Newtonian potential φ. Thus, the limiting case [34] for the gravitational field corresponds to taking the Newtonian limit. Finally, the normalization condition [16] reduces to

$$\int_0^\infty \alpha(r)^2\, dr = 1 \tag{36}$$

The boundary conditions [17]–[19], [24]–[26] are transformed into

$$w(r) = 1 - \frac{\lambda}{2}r^2 + \mathcal{O}(r^3), \quad \lim_{r\to\infty} w(r) = \pm 1 \tag{37}$$

$$\alpha(r) = \alpha_1 r + \mathcal{O}(r^3), \quad \hat{\beta}(r) = \mathcal{O}(r^3) \tag{38}$$

$$\varphi(r) = \varphi_0 + \mathcal{O}(r^3), \quad \lim_{r\to\infty} \varphi(r) < \infty \tag{39}$$

with the three parameters λ, α_1, and φ_0. We point out that the limiting system contains only one coupling constant q. According to [31] and [33], q is in dimensionless form given by

$$e^2 m^2 \kappa \to q \tag{40}$$

Hence, in dimensionless quantities, the limit [17] describes the situation where the gravitational coupling goes to zero, while the YM coupling constant goes to infinity like $e^2 \sim (m^2\kappa)^{-1}$. Therefore, this limiting case is called the reciprocal coupling limit (RCL). The reciprocal coupling system is given by eqns [29], [30], [32], and [34] together with the normalization conditions [35] and the boundary conditions [36]–[38]. According to [28], the parameter E coincides up to a scaling factor with $\omega - m$, and thus has the interpretation as the (properly scaled) energy of the Dirac particle. As in Newtonian mechanics, the potential φ is determined only up to a constant $\mu \in \mathbb{R}$; namely, the reciprocal limit equations are invariant under the transformation

$$\varphi \to \varphi + \mu, \qquad E \to E - \mu \tag{41}$$

To simplify the connection between the EDYM equations, and the RCL equations, we introduce a parameter ε in such a way that as $\varepsilon \to 0$, EDYM $\to$ RCL; namely,

$$\varepsilon = \frac{m^2\kappa}{e^2}$$

Notice that ε describes the relative strength of gravity versus the YM interaction. For realistic physical situations, the gravitational coupling is weak; namely, $m^2\kappa \ll 1$, but the YM coupling constant is of order $1 : e^2 \sim 1$. So we investigate the parameter range $\varepsilon \ll 1, q \sim 0$. These form the starting points for the numeric below.

We seek stable bound states for weak gravitational coupling. For this purpose, we consider the total binding energy

$$B = M - m \qquad [42]$$

where M is the ADM mass defined by [17] and m is the rest mass of the Dirac particle. B is thus the amount of energy set free when the binding is broken. If $B < 0$, then energy is needed to break up the binding. According to Lee (1987), a solution is stable if $B < 0$. In order to find solutions of the RCL equations with $B < 0$, Lee's treatment and a new two-parameter shooting method (Finster *et al.* 2000b) can be used. Stable solutions of these RCL equations then follow (see Finster *et al.* (2000b) for details).

We now turn to the full EDYM equations. Here are the key steps of our method:

1. Find solutions which are small perturbations of the limiting (RCL) solutions.
2. Trace these solutions by gradually changing the coupling constants.
3. This should yield a one-parameter family of solutions which are "far" from the known limiting solutions.

The point is that we use the RCL solutions as a starting point for numerics, and we "continue" these solutions to solutions of the full EDYM equations.

To be somewhat more specific, we see that if we fix ε and q, we have two parameters:

$$\alpha_1 = \alpha'(0) \quad \text{and} \quad E = \omega - m$$

and two conditions at ∞:

$$\alpha^2 + \beta^2 \to 0, \qquad w^2 \to 1$$

We consider the EDYM equations with weaker side conditions

$$0 < \lambda^2 \equiv \int_0^\infty (\alpha^2 + \beta^2) \frac{\sqrt{T}}{A} \mathrm{d}r < \infty$$
$$0 < \tau = \lim_{r\to\infty} T(r) < \infty$$
$$\lim_{r\to\infty} w^2(r) = 1$$
$$\rho = \lim_{r\to\infty} r(1 - A(r)) < \infty$$

Then we rescale these solutions to obtain the true side conditions via the transformations

$$\tilde{\alpha}(r) = \sqrt{\tau}\lambda^{-2}\alpha(\lambda^{-2}r)$$
$$\tilde{\beta}(r) = \sqrt{\tau}\lambda^{-2}\beta(\lambda^{-2}r)$$
$$\tilde{A}(r) = A(\lambda^{-2}r), \quad \tilde{T}(r) = \tau^{-1}T(\lambda^{-2}r)$$
$$\tilde{m} = \lambda^{-2}m, \quad \tilde{\omega} = \tau\lambda^{-2}\omega$$
$$\tilde{\kappa} = \lambda^6\kappa, \quad \tilde{e}^2 = \lambda^2 e^2$$

Discussion

In this article we have considered the SU(2) EDYM equations. Our first result shows that the only black hole solutions of these equations are the BM black holes; that is, the spinors must vanish identically outside of the black hole. In other words, the EDYM equations do not admit normalizable black hole solutions. Thus, as mentioned earlier, this result indicates that the Dirac particle either enters the black hole or escapes to infinity. Two recent publications (Finster *et al.* 2002a,b) we consider the Cauchy problem for a massive Dirac equation in a charged, rotating-black-hole geometry (the non-extreme Kerr–Newman black hole), with compactly supported initial data outside the black hole. We prove that, in this case, the probability that the Dirac particle lies in any compact set tends to zero as $t \to \infty$. This means that the Dirac particle indeed either enters the black hole or tends to infinity. We also show that the wave function decays at a rate $t^{-5/6}$ on any compact set outside of the event horizon.

For particle-like solutions of the SU(2) EDYM equations, we find stable bound states for arbitrarily weak gravitational coupling. This shows that as weak as the gravitational interaction is, it has a regularizing effect on the equations. The stability of particle-like solutions of the EDYM equations is in sharp contrast to the EYM equations, where the particle-like solutions are all unstable (Straumann and Zhou 1990).

Acknowledgment

This work was partially supported by the NSF, Contract Number DMS-010-3998.

See also: Abelian and Nonabelian Gauge Theories using Differential Forms; Black Hole Mechanics; Bosons and Fermions in External Fields; Dirac Operator and Dirac Field; Einstein Equations: Exact Solutions; Einstein Equations: Initial Value Formulation; Noncommutative Geometry and the Standard Model; Relativistic Wave Equations Including Higher Spin Fields; Symmetry Classes in Random Matrix Theory.

Further Reading

Adler R, Bazin M, and Schiffer M (1975) *Introduction to General Relativity*, 2nd edn. New York: McGraw-Hill.

Bartnik R and McKinnon J (1988) Particle-like solutions of the Einstein–Yang–Mills equations. *Physical Review Letters* 61: 141–144.

Bjorken J and Drell S (1964) *Relativistic Quantum Mechanics*. New York: McGraw-Hill.

Finster F (1988) Local U(2, 2) symmetry in relativistic quantum mechanics. *Journal of Mathematical Physics* 39: 6276–6290.

Finster F, Smoller J, and Yau S-T (1999a) Particle-like solutions of the Einstein–Dirac equations. *Physical Review D* 59: 104020/1–104020/19.

Finster F, Smoller J, and Yau S-T (2000a) The interaction of Dirac particles with non-abelian gauge fields and gravitation-black holes. *Michigan Mathematical Journal* 47: 199–208.

Finster F, Smoller J, and Yau S-T (2000b) The interaction of Dirac particles with non-abelian gauge fields and gravitation-bound states. *Nuclear Physics B* 584: 387–414.

Finster F, Smoller J, and Yau S-T (1999b) Nonexistence of black hole solutions for a spherically symmetric, static Einstein–Dirac–Maxwell system. *Communications in Mathematical Physics* 205: 249–262.

Finster F, Smoller J, and Yau S-T (2000c) Nonexistence of time-periodic solutions of the Dirac equation in a Reissner–Nordström black hole background. *Journal of Mathematical Physics* 41: 2173–2194.

Finster F, Kamran N, Smoller J, and Yau S-T (2002a) The long time dynamics of Dirac particles in the Kerr–Newman black hole geometry. *Advances in Theoretical and Mathematical Physics* 7: 25–52.

Finster F, Kamran N, Smoller J, and Yau S-T (2002b) Decay rates and probability estimates Dirac particles in the Kerr–Newman black hole geometry. *Communications in Mathematical Physics* 230: 201–244.

Lee TD (1987) Mini-soliton stars. *Physical Review D* 25: 3640–3657.

Smoller J, Wasserman A, and Yau S-T (1993) Existence of infinitely-many smooth global solutions of the Einstein–Yang/Mills equations. *Communications in Mathematical Physics* 151: 303–325.

Smoller J and Wasserman A (1998) Extendability of solutions of the Einstein–Yang/Mills equations. *Communications in Mathematical Physics* 194: 707–733.

Straumann N and Zhou Z (1990) Instability of the Bartnik–McKinnon solution of the Einstein–Yang–Mills equations. *Physics Letters B* 237: 353–356.

Donaldson Invariants *see* Gauge Theoretic Invariants of 4-Manifolds

Donaldson–Witten Theory

M Mariño, CERN, Geneva, Switzerland

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Since they were introduced by Witten in 1988, topological quantum field theories (TQFTs) have had a tremendous impact in mathematical physics (see Birmingham *et al.* (1991) and Cordes *et al.* for a review). These quantum field theories are constructed in such a way that the correlation functions of certain operators provide topological invariants of the spacetime manifold where the theory is defined. This means that one can use the methods and insights of quantum field theory in order to obtain information about topological invariants of low-dimensional manifolds.

Historically, the first TQFT was Donaldson–Witten theory, also called topological Yang–Mills theory. This theory was constructed by Witten (1998) starting from $\mathcal{N}=2$ super Yang–Mills by a procedure called

Notes:

"topological twisting." The resulting model is topological and the famous Donaldson invariants of 4-manifolds are then recovered as certain correlation functions in the topological theory. The analysis of Witten (1998) did not indicate any new method to compute the invariants, but in 1994 the progress in understanding the nonperturbative dynamics of $\mathcal{N}=2$ theories (Seiberg and Witten 1994 a, b) led to an alternative way of computing correlation functions in Donaldson–Witten theory. As Witten (1994) showed, Donaldson–Witten theory can be reduced to another, simpler topological theory consisting of a twisted abelian gauge theory coupled to spinor fields. This theory leads to a different set of 4-manifold invariants, the so-called "Seiberg–Witten invariants," and Donaldson invariants can be expressed in terms of these invariants through Witten's "magic formula." The connection between Seiberg–Witten and Donaldson invariants was streamlined and extended by Moore and Witten by using the so-called u-plane integral (Moore and Witten 1998). This has led to a rather complete understanding of Donaldson–Witten theory from a physical point of view.

In this article we provide a brief review of Donaldson–Witten theory. First, we describe the construction of the model, from both a mathematical and a physical point of view, and state the main results for the Donaldson–Witten generating functional. In the next section, we present the basic results of the u-plane integral of Moore and Witten and sketch how it can be used to solve Donaldson–Witten theory. In the final section, we mention some generalizations of the basic framework. For a complete exposition of Donaldson–Witten theory, the reader is referred to the book by Labastida and Mariño (2005). A short review of the u-plane integral can be found in Mariño and Moore (1998a).

Donaldson–Witten Theory: Basic Construction and Results

Donaldson–Witten Theory According to Donaldson

Donaldson theory as formulated in Donaldson (1990), Donaldson and Kronheimer (1990), and Friedman and Morgan (1991) starts with a principal $G=\mathrm{SO}(3)$ bundle $V \to X$ over a compact, oriented, Riemannian 4-manifold X, with fixed instanton number k and Stiefel–Whitney class $w_2(V)$ (SO(3) bundles on a 4-manifold are classified up to isomorphism by these topological data). The moduli space of anti-self-dual (ASD) connections is then defined as

$$\mathcal{M}_{\mathrm{ASD}} = \{A : F^+(A) = 0\}/\mathcal{G} \qquad [1]$$

where $F^+(A)$ is the self-dual part of the curvature, and $\mathcal{G}$ is the group of gauge transformations. To construct the Donaldson polynomials, one considers the universal bundle

$$P = (V \times \mathcal{A}^*)/(\mathcal{G} \times G) \qquad [2]$$

where $\mathcal{A}^*$ is the space of irreducible G-connections on V. This is a G-bundle over $\mathcal{B}^* \times X$, where $\mathcal{B}^* = \mathcal{A}^*/\mathcal{G}$ is the space of irreducible connections modulo gauge transformations, and as such has a Pontrjagin class

$$p_1(P) \in H^*(\mathcal{B}^*) \otimes H^*(X) \qquad [3]$$

One can then obtain differential forms on $\mathcal{B}^*$ by taking the slant product of $p_1(P)$ with homology classes in X. In this way we obtain the Donaldson map:

$$\mu : H_i(X) \longrightarrow H^{4-i}(\mathcal{B}^*) \qquad [4]$$

After restriction to $\mathcal{M}_{\mathrm{ASD}}$, we obtain the following differential forms on the moduli space of ASD connections:

$$\begin{aligned} x \in H_0(X) &\to \mathcal{O}(x) \in H^4(\mathcal{M}_{\mathrm{ASD}}) \\ S \in H_2(X) &\to I_2(S) \in H^2(\mathcal{M}_{\mathrm{ASD}}) \end{aligned} \qquad [5]$$

If the manifold X has $b_1(X) \neq 0$, there are also cohomology classes associated to 1-cycles and 3-cycles, but we will not consider them here.

We can now formally define the Donaldson invariants as follows. Consider the space

$$A(X) = \mathrm{Sym}(H_0(X) \oplus H_2(X)) \qquad [6]$$

with a typical element written as $x^\ell S_{i_1} \cdots S_{i_p}$. The Donaldson invariant corresponding to this element of $A(X)$ is the following intersection number:

$$\begin{aligned} &\mathcal{D}_X^{w_2(V),k}(x^\ell S_{i_1} \cdots S_{i_p}) \\ &\quad = \int_{\mathcal{M}_{\mathrm{ASD}}} \mathcal{O}^\ell \wedge I_2(S_{i_1}) \wedge \cdots \wedge I_2(S_{i_p}) \end{aligned} \qquad [7]$$

where $\mathcal{M}_{\mathrm{ASD}}$ is the moduli space of ASD connections with second Stiefel–Whitney class $w_2(V)$ and instanton number k. The integral in [7] will be different from zero only if the degrees of the forms add up to $\dim(\mathcal{M}_{\mathrm{ASD}})$.

It is very convenient to pack all Donaldson invariants in a generating functional. Let $\{S_i\}_{i=1,\ldots,b_2}$ be a basis of 2-cycles. We introduce the formal sum

$$S = \sum_{i=1}^{b_2} v_i S_i \qquad [8]$$

where v_i are complex numbers. We then define the Donaldson–Witten generating functional as

$$Z_{\mathrm{DW}}^{w_2(V)}(p, v_i) = \sum_{k=0}^{\infty} \mathcal{D}_X^{w_2(V),k}(e^{px+S}) \qquad [9]$$

where on the right-hand side we are summing over all instanton numbers, that is, we are summing over all topological configurations of the SO(3) gauge field with a fixed $w_2(V)$. This gives a formal power series in p and v_i.

For $b_2^+(X) > 1$, the generating functional [9] is a diffeomorphism invariant of X; therefore, it is potentially a powerful tool in four-dimensional topology. When $b_2^+(X) = 1$, Donaldson invariants are metric dependent. The metric dependence can be described in more detail as follows. Define the period point as the harmonic 2-form satisfying

$$*\omega = \omega, \qquad \omega^2 = 1 \qquad [10]$$

which depends on the conformal class of the metric. As the conformal class of the metric varies, ω describes a curve in the cone

$$V_+ = \{\omega \in H^2(X, \boldsymbol{R}) : \omega^2 > 0\} \qquad [11]$$

Let $\zeta \in H^2(X)$ satisfy

$$\zeta \equiv w_2(V) \bmod 2, \quad \zeta^2 < 0, \ (\zeta, \omega) = 0 \qquad [12]$$

Such an element ζ defines a "wall" in V_+:

$$W_\zeta = \{\omega : (\zeta, \omega) = 0\} \qquad [13]$$

The complements of these walls are called "chambers," and the cone V_+ is then divided in chambers separated by walls. A class ζ satisfying [12] is the first Chern class associated to a reducible solution of the ASD equations, and it causes a singularity in moduli space: the Donaldson invariants jump when we pass through such a wall. Therefore, when $b_2^+(X) = 1$, Donaldson invariants are metric independent in each chamber. A basic problem in Donaldson–Witten theory is to determine the jump in the generating function as we cross a wall,

$$Z_+^\zeta(p, S) - Z_-^\zeta(p, S) = WC_\zeta(p, S) \qquad [14]$$

The jump term $WC_\zeta(p, S, \delta)$ is usually called the "wall-crossing" term.

The basic goal of Donaldson theory is to study the properties of the generating functional [9] and to compute it for different 4-manifolds X. On the mathematical side, many results have been obtained on Z_{DW}, and some of them can be found in Donaldson and Kronheimer (1990), Friedman and Morgan (1991), Stern (1998), and Göttsche (1996). On the other hand, Donaldson theory can be formulated as a topological field theory, and many of these results can be obtained by using quantum field theory techniques. This will be our main focus for the rest of the article.

Donaldson–Witten Theory According to Witten

Witten (1988) constructed a twisted version of $\mathcal{N} = 2$ super Yang–Mills theory which has a nilpotent Becchi–Rouet–Stora–Tyutin (BRST) charge (modulo gauge transformations)

$$\overline{Q} = \epsilon^{\dot{\alpha}\dot{A}} Q_{\dot{\alpha}\dot{A}} \qquad [15]$$

where $Q_{\dot{\alpha}\dot{A}}$ are the supersymmetric (SUSY) charges. Here $\dot{\alpha}$ is a chiral spinor index and A has its origin in the SU(2) $\mathcal{R}$-symmetry. The field content of the theory is the standard twisted $\mathcal{N} = 2$ vector multiplet:

$$A, \psi_\mu = \psi_{\alpha\dot{\alpha}}, \phi, \quad D_{\mu\nu}^+, \chi_{\mu\nu}^+ = \overline{\psi}_{\dot{\alpha}\dot{\beta}}, \quad \overline{\phi}, \eta = \overline{\psi}^{\dot{\alpha}}{}_{\dot{\alpha}} \qquad [16]$$

where $(1/2)D_{\mu\nu}^+ \, dx^\mu \, dx^\nu$ is a self-dual 2-form derived from the auxiliary fields, etc. All fields are valued in the adjoint representation of the gauge group. After twisting, the theory is well defined on any Riemannian 4-manifold, since the fields are naturally interpreted as differential forms and the $\overline{Q}$ charge is a scalar (Witten 1988).

The observables of the theory are $\overline{Q}$ cohomology classes of operators, and they can be constructed from 0-form observables $\mathcal{O}^{(0)}$ using the descent procedure. This amounts to solving the equations

$$d\mathcal{O}^{(i)} = \{\overline{Q}, \mathcal{O}^{(i+1)}\}, \quad i = 0, \ldots, 3 \qquad [17]$$

The integration over i-cycles $\gamma^{(i)}$ in X of the operators $\mathcal{O}^{(i)}$ is then an observable. These descent equations have a canonical solution: the 1-form-valued operator $K_{\alpha\dot{\alpha}} = -i\delta_\alpha^A Q_{\dot{\alpha}A}/4$ verifies

$$d = \{\overline{Q}, K\} \qquad [18]$$

as a consequence of the supersymmetry algebra. The operators $\mathcal{O}^{(i)} = K^i \mathcal{O}^{(0)}$ solve the descent equations [17] and are canonical representatives. When the gauge group is SU(2), the observables are obtained by the descent procedure from the operator

$$\mathcal{O} = \mathrm{tr}(\phi^2) \qquad [19]$$

The topological descendant $\mathcal{O}^{(2)}$ is given by

$$\mathcal{O}^{(2)} = -\tfrac{1}{2}\mathrm{tr}\left(\tfrac{1}{\sqrt{2}}\phi(F_{\mu\nu}^- + D_{\mu\nu}^+) - \tfrac{1}{4}\psi_\mu\psi_\nu\right) dx^\mu \wedge dx^\nu \qquad [20]$$

and the resulting observable is

$$I_2(S) = \int_S \mathcal{O}^{(2)} \qquad [21]$$

$\mathcal{O}$ and $I_2(S)$ correspond to the cohomology classes in [5]. One of the main results of Witten (1988) is that the semiclassical approximation in the twisted $\mathcal{N}=2$ Yang–Mills theory is exact. The semiclassical evaluation of correlation functions of the observables above leads directly to the definition of Donaldson invariants, and the generating functional [9] can be written as a correlation function of the twisted theory. One then has

$$Z_{\mathrm{DW}}^{w_2(V)}(p,S) = \left\langle \exp\left(p\mathcal{O} + I_2(S)\right)\right\rangle \qquad [22]$$

Results for the Donaldson–Witten Generating Function

The basic results that have emerged from the physical approach to Donaldson–Witten theory are the following.

1. The Donaldson–Witten generating functional is in general the sum of the two terms,

$$Z_{\mathrm{DW}} = Z_u + Z_{\mathrm{SW}} \qquad [23]$$

(We have omitted the Stiefel–Whitney class for convenience.) The first term, Z_u, is called the "u-plane integral." It is given by a complicated integral over C which can be written, in turn, as an integral over a fundamental domain of the congruence subgroup $\Gamma^0(4)$ of $\mathrm{SL}(2,\mathbf{Z})$. Z_u depends only on the cohomology ring of X, and therefore does not contain any information beyond the one provided by classical topology. Finally, Z_u vanishes if $b_2^+(X) > 1$, and it is responsible for the wall-crossing behavior of Z_{DW} when $b_2^+(X) = 1$.

2. The second term of [23], Z_{SW}, is called the Seiberg–Witten contribution. This contribution involves the Seiberg–Witten invariants of X, which are obtained by considering the moduli problem defined by the Seiberg–Witten monopole equations (Witten 1994b):

$$\begin{aligned} F^+_{\dot\alpha\dot\beta} + 4\mathrm{i}\bar M_{(\dot\alpha} M_{\dot\beta)} &= 0 \\ D_L^{\alpha\dot\alpha} M_{\dot\alpha} &= 0 \end{aligned} \qquad [24]$$

In these equations, $M_{\dot\alpha}$ is a section of the spinor bundle $S^+ \otimes L^{1/2}$, L is the determinant line bundle of a Spin_c structure on X, $F^+_{\dot\alpha\dot\beta} = \bar\sigma^{\mu\nu}_{\dot\alpha\dot\beta} F^+_{\mu\nu}$ is the self-dual part of the curvature of a U(1) connection on L, and D_L is the Dirac operator for the bundle $S^+ \otimes L^{1/2}$. The solutions of these equations modulo gauge equivalence form the moduli space $\mathcal{M}_S W$, and the Seiberg–Witten invariants are defined by integrating suitable differential forms on this moduli space. We will label Spin_c structures by the class $\lambda = c_1(L^{1/2}) \in H^2(X,\mathbf{Z}) + w_2(X)/2$. We say that λ is a Seiberg–Witten basic class if the corresponding Seiberg–Witten invariants are not all zero. If $\mathcal{M}_S W$ is zero dimensional, the Seiberg–Witten invariant depends only on the Spin_c structure associated to $\lambda = c_1(L^{1/2})$, and is denoted by $\mathrm{SW}(\lambda)$.

3. A manifold X is said to be of Seiberg–Witten simple type if all the Seiberg–Witten basic classes have a zero-dimensional moduli space. For simply connected 4-manifolds of Seiberg–Witten simple type and with $b_2^+(X) > 1$, Witten determined the Seiberg–Witten contribution and proposed the following "magic formula" for Z_{DW} (Witten 1994b):

$$\begin{aligned} Z_{\mathrm{DW}} = 2^{1+7\chi/4+11\sigma/4} \sum_\lambda \mathrm{e}^{2\mathrm{i}\pi(\lambda_0\cdot\lambda+\lambda_0^2)} \Big[& \mathrm{e}^{2p+S^2/2}\mathrm{e}^{2(S,\lambda)} \\ & + \mathrm{i}^{\chi_h - w_2(V)^2}\mathrm{e}^{-2p-S^2/2}\mathrm{e}^{-2\mathrm{i}(S,\lambda)}\Big]\mathrm{SW}(\lambda) \end{aligned} \qquad [25]$$

In this equation, χ, σ are the Euler characteristic and signature of X, respectively, $\chi_h = (\chi+\sigma)/4$ is the holomorphic Euler characteristic of X, and λ_0 is an integer lifting of $w_2(V)$. This formula generalizes previous results by Witten (1994a) for Kähler manifolds. It also follows from this formula that the Donaldson–Witten generating function of simply connected 4-manifolds of Seiberg–Witten simple type and with $b_2^+(X) > 1$ satisfies

$$\left(\frac{\partial^2}{\partial p^2} - 4\right) Z_{\mathrm{DW}} = 0$$

which is the Donaldson simple type condition introduced by Kronheimer and Mrowka (1994).

4. Using the u-plane integral, one can find explicit expressions for Z_{DW} in more general situations (like non-simply-connected manifolds or manifolds which are not of Seiberg–Witten simple type).

In the next section we explain the formalism of the u-plane integral introduced by Moore and Witten (1998), which makes possible a detailed derivation of the above results.

The u-Plane Integral

Definition of the u-Plane Integral

The evaluation of the Donaldson–Witten generating function can be made by using the results of Seiberg and Witten (1994 a, b) on the low-energy dynamics of SU(2), $\mathcal{N}=2$ Yang–Mills theory. In their work, Seiberg and Witten determined the exact low-energy effective action of the model up to two derivatives.

From a physical point of view, there are certainly corrections to this effective action which are difficult to evaluate. Fortunately, the computation in the twisted version of the theory can be done by just considering the Seiberg–Witten effective action. This is because the correlation functions in the twisted theory are invariant under rescalings of the metric, so we can evaluate them in the limit of large distances or equivalently of very low energies. The effective action up to two derivatives is sufficient for that purpose.

One way of describing the main result of the work of Seiberg and Witten is that the moduli space of $\bar{Q}$-fixed points of the twisted SO(3) $\mathcal{N}=2$ theory on a compact 4-manifold has two branches, which we refer to as the Coulomb and Seiberg–Witten branches. On the Coulomb branch the expectation value

$$u = \frac{\langle \mathrm{tr}\, \phi^2 \rangle}{16\pi^2}$$

breaks $SO(3) \to U(1)$ via the standard Higgs mechanism. The Coulomb branch is simply a copy of the complex u-plane. The low-energy effective theory on this branch is simply the abelian $\mathcal{N}=2$ gauge theory. However, at two points, $u=\pm 1$, there is a singularity where the moduli space meets a second branch, the Seiberg–Witten branch. At these points, the effective action is given by the magnetic dual of the U(1), $\mathcal{N}=2$ gauge theory coupled to a monopole matter hypermultiplet. Therefore, this branch consists of solutions to the Seiberg–Witten equations [24].

Since the manifold X is compact, the partition function of the twisted theory is a sum over "all" vacuum states. Equation [23] then follows. In this equation, Z_u comes from "integrating over the u-plane," while Z_{SW} corresponds to the points $u=\pm 1$. As we stated before, Z_u vanishes for manifolds of $b_2^+(X) > 1$, but once this piece has been determined an argument originally presented at Moore and Witten (1998) allows one to derive the form of Z_{SW} as well for arbitrary $b_2^+(X) \geq 1$.

The computation of Z_u is presented in detail in Moore and Witten (1998). The starting point of the computation is the untwisted low-energy theory, which has been described in detail in Seiberg and Witten (1994 a, b) and Witten (1995). It is an $\mathcal{N}=2$ theory characterized by a prepotential $\mathcal{F}$ which depends on an $\mathcal{N}=2$ vector multiplet. The effective gauge coupling is given by $\tau(a)=\mathcal{F}''(a)$, where a is the scalar component of the vector multiplet. The Euclidean Lagrange density for the u-plane theory can be obtained simply by twisting the physical theory. It can be written as

$$\begin{aligned} &\frac{\mathrm{i}}{6\pi} K^4 \mathcal{F}(a) + \frac{1}{16\pi}\left\{\bar{Q}, \overline{\mathcal{F}}'' \chi (D + F_+)\right\} \\ &\quad - \frac{\mathrm{i}\sqrt{2}}{32\pi}\left\{\bar{Q}, \overline{\mathcal{F}}' d * \psi\right\} - \frac{\sqrt{2}\mathrm{i}}{3 \times 2^5 \pi} \\ &\quad \times \left\{Q, \overline{\mathcal{F}}''' \chi_{\mu\nu}\chi^{\nu\lambda}\chi_\lambda{}^\mu\right\}\sqrt{g}d^4x \\ &\quad + \mathcal{A}(u)\mathrm{tr}R \wedge R + \mathcal{B}(u)\mathrm{tr}R \wedge \tilde{R} \end{aligned} \qquad [26]$$

where $\mathcal{A}(u), \mathcal{B}(u)$ describe the coupling to gravity, and after integration of the corresponding differential forms we obtain terms proportional to the signature σ and Euler characteristic χ of X. The data of the low-energy effective action can be encoded in an elliptic curve of the form

$$y^2 = x^3 - ux^2 + \tfrac{1}{4}x \qquad [27]$$

and τ is the modulus of the curve. The monodromy group of this curve is $\Gamma^0(4)$. All the quantities involved in the action can be obtained by integrating a certain meromorphic differential on the curve, and they can be expressed in terms of modular forms.

As for the operators, we have $u=\mathcal{O}(P)$ by definition. We may then obtain the 2-observables from the descent procedure. The result is that $I(S) \to \tilde{I}(S) = \int_S K^2 u = \int_S (\mathrm{d}u/\mathrm{d}a)(D_+ + F_-) + \cdots$. Here D_+ is the auxiliary field. Although one has $I(S) \to \tilde{I}(S)$ in going from the microscopic theory to the effective theory, it does not necessarily follow that $I(S_1)I(S_2) \to \tilde{I}(S_1)\tilde{I}(S_2)$ because there can be contact terms. If S_1 and S_2 intersect, then in passing to the low-energy theory we integrate out massive modes. This can induce delta function corrections to the operator product expansion modifying the mapping to the low-energy theory as follows:

$$\exp(I(S)) \to \exp\left(\tilde{I}(S) + S^2 T(u)\right) \qquad [28]$$

where $T(u)$ is the contact term. Such contact terms were observed in Witten (1994a) and studied in detail in Losev *et al.* (1998). It can be shown that

$$T(u) = -\frac{1}{24} E_2(\tau)\left(\frac{\mathrm{d}u}{\mathrm{d}a}\right)^2 + \frac{1}{3}u \qquad [29]$$

where $E_2(\tau)$ is Eisenstein's series and $\mathrm{d}a/\mathrm{d}u$ is one of the periods of the elliptic curve [27].

The final result of Moore and Witten is the following expression:

$$Z_u(p,S) = \int_C \frac{\mathrm{d}u\,\mathrm{d}\bar{u}}{y^{1/2}} \mu(\tau) \mathrm{e}^{2pu + S^2 \hat{T}(u)} \Psi \qquad [30]$$

Here,

$$\mu(\tau) = \frac{\mathrm{d}\bar{\tau}}{\mathrm{d}\bar{u}}\left(\frac{\mathrm{d}a}{\mathrm{d}u}\right)^{1-(1/2)\chi}\Delta^{\sigma/8}$$
$$\hat{T}(u) = T(u) + \frac{(\mathrm{d}u/\mathrm{d}a)^2}{8\pi y} \qquad [31]$$

where $y = \operatorname{Im}\tau$ and Δ is the discriminant of the curve [27]. The quantity Ψ is essentially a Narain–Siegel theta function associated to the lattice $H^2(X, Z)$. Notice that this lattice is Lorentzian and has signature $(1, (-1)^{b_2^-(X)})$ (since $b_2^+(X) = 1$). The self-dual projection of a 2-form λ can be done with the period point ω as $\lambda_+ = (\lambda, \omega)\omega$. The lattice is shifted by half the second Stiefel–Whitney class of the bundle, $w_2(V)$, that is,

$$\Gamma = H^2(X, \mathbf{Z}) + \tfrac{1}{2}w_2(V)$$

and

$$\Psi = \exp\left[-\frac{1}{8\pi y}\left(\frac{\mathrm{d}u}{\mathrm{d}a}\right)^2 S_-^2\right]\mathrm{e}^{2\pi\mathrm{i}\lambda_0^2}\sum_{\lambda\in\Gamma}(-1)^{(\lambda-\lambda_0)\cdot w_2(X)}$$
$$\times\left[(\lambda,\omega) + \frac{\mathrm{i}}{4\pi y}\frac{\mathrm{d}u}{\mathrm{d}a}(S,\omega)\right]$$
$$\times\exp\left[-\mathrm{i}\pi\bar{\tau}(\lambda_+)^2 - \mathrm{i}\pi\tau(\lambda_-)^2 - \mathrm{i}\frac{\mathrm{d}u}{\mathrm{d}a}(S,\lambda_-)\right] \qquad [32]$$

Here, $w_2(X)$ is the second Stiefel–Whitney class of X, and λ_0 is a choice of lifting of $w_2(V)$ to $H^2(X, \mathbf{Z})$. This expression can be extended to the non-simply-connected case (see Mariño and Moore (1999) and Moore and Witten (1998)). The study of the u-plane integral leads to a systematic derivation of many important results in Donaldson–Witten theory. We will discuss in detail two such applications, Göttsche's wall-crossing formula and Witten's "magic formula."

Wall-Crossing Formula

As shown by Moore and Witten, the u-plane integral is well defined and does not depend on the period point (hence on the metric on X) except for discontinuous behavior at walls. There are two kinds of walls, associated, respectively, to the singularities at $u = \infty$ (the semiclassical region of the underlying Yang–Mills theory) and at $u = \pm 1$, given by

$$u = \infty\colon\ \lambda_+ = 0, \lambda \in H^2(X, \mathbf{Z}) + \tfrac{1}{2}w_2(V)$$
$$u = \pm 1\colon\ \lambda_+ = 0, \lambda \in H^2(X, \mathbf{Z}) + \tfrac{1}{2}w_2(X) \qquad [33]$$

The first type of walls is precisely the one that appears in Donaldson theory on manifolds of $b_2^+(X) = 1$. The discontinuity of the u-plane integral at these walls can be easily computed from eqn [33]:

$$\mathrm{WC}_{\zeta=2\lambda}(p, S)$$
$$= -\frac{\mathrm{i}}{2}(-1)^{(\lambda-\lambda_0, w_2(X))}\mathrm{e}^{2\pi\mathrm{i}\lambda_0^2}\Big[q^{-\lambda^2/2}h_\infty(\tau)^{-2}\vartheta_4^\sigma f_\infty^{-1}$$
$$\times\exp\{2pu_\infty + S^2T_\infty - \mathrm{i}(\lambda, S)/h_\infty\}\Big]_{q^0} \qquad [34]$$

This expression involves the modular forms $h_\infty, f_{2\infty}, u_\infty$, and T_∞ (the subscript ∞ refers to the fact that they are computed at the "electric" frame which is appropriate for the Seiberg–Witten curve at $u \to \infty$). They can be written in terms of Jacobi theta functions $\vartheta_i(q)$, with $q = \mathrm{e}^{2\pi\mathrm{i}\tau}$, and their explicit expression is

$$h_\infty(q) = \tfrac{1}{2}\vartheta_2(q)\vartheta_3(q)$$
$$f_\infty(q) = \frac{\vartheta_2(q)\vartheta_3(q)}{2\vartheta_4^8(q)}$$
$$u_\infty(q) = \frac{\vartheta_2^4(q) + \vartheta_3^4(q)}{2(\vartheta_2(q)\vartheta_3(q))^2} \qquad [35]$$
$$T_\infty(q) = -\frac{1}{24}\frac{E_2(q)}{h_\infty^2(q)} + \frac{1}{3}u_\infty(q)$$

The subindex q^0 means that in the expansion in q of the modular forms, we pick the constant term. The formula [34] agrees with the formula of Göttsche (1996) for the wall crossing of the Donaldson–Witten generating functional.

The Seiberg–Witten Contribution and Witten's Magic Formula

At $u = \pm 1, Z_u$ jumps at the second type of walls [33], which are called Seiberg–Witten (SW) walls. In fact, these walls are labeled by classes $\lambda \in H^2(X; Z) + (1/2)w_2(X)$, which correspond to Spin_c structures on X. At these walls, the Seiberg–Witten invariants have wall-crossing behavior. Since the Donaldson polynomials do not jump at SW walls, it must happen that the change of Z_u at $u = \pm 1$ is canceled by the change of Z_{SW}. As shown by Moore and Witten, this actually allows one to obtain a precise expression for Z_{SW} for general 4-manifolds of $b_2^+(X) \geq 1$.

On general grounds, Z_{SW} is given by the sum of the generating functionals at $u = \pm 1$. These involve a magnetic U(1), $\mathcal{N} = 2$ vector multiplet coupled to a hypermultiplet (the monopole field). The twisted

Lagrangian for such a system involves the magnetic prepotential $\widetilde{\mathcal{F}}_D(a_D)$, and it can be written as

$$\{\vec{Q}, W\} + \frac{\mathrm{i}}{16\pi}\tilde{\tau}_D F\wedge F + p(u)\mathrm{tr}R\wedge *R + \ell(u)\mathrm{tr}R\wedge R - \frac{\mathrm{i}\sqrt{2}}{32\pi}\frac{\mathrm{d}\tilde{\tau}_D}{\mathrm{d}a_D}(\psi\wedge\psi)\wedge F + \frac{\mathrm{i}}{3\times 2^7\pi}\frac{\mathrm{d}^2\tilde{\tau}_D}{\mathrm{d}a_D^2}\psi\wedge\psi\wedge\psi\wedge\psi \qquad [36]$$

where $\tilde{\tau}_D = \widetilde{\mathcal{F}}_D''(a_D)$. Using the cancellation of wall crossings, one can actually compute the functions $\widetilde{\mathcal{F}}_D(a_D), p(u), \ell(u)$ and determine the precise form of the Seiberg–Witten contributions. One finds that a Spin_c structure λ at $u=1$ gives the following contribution to the Donaldson–Witten generating functional:

$$Z_{\mathrm{SW}}^{u=1,\lambda} = \frac{\mathrm{SW}(\lambda)}{16}\mathrm{e}^{2\mathrm{i}\pi(\lambda_0^2-\lambda_0\cdot\lambda)} \times\left[q_D^{-\lambda^2/2}\frac{\vartheta_2^{8+\sigma}}{a_D h_M}\left(-2\mathrm{i}\frac{a_D}{h_M^2}\right)^{\chi_h} \times\exp\left(2pu_M + \mathrm{i}(\lambda,S)/h_M + S^2T_M\right)\right]_{q_D^0} \qquad [37]$$

Here, a_D, h_M, u_M, and T_M are modular forms that can be expressed as well in terms of Jacobi theta functions $\vartheta_i(q_D)$, where $q_D = \exp(2\pi\mathrm{i}\tau_D)$. The subscript M refers to the monopole point, and they are related by an S-transformation to the quantities obtained in the "electric" frame at $u\to\infty$. Their explicit expression is

$$\begin{aligned} a_D(q_D) &= -\frac{\mathrm{i}}{6}\frac{2E_2(q_D)-\vartheta_3^4(q_D)-\vartheta_4^4(q_D)}{\vartheta_3(q_D)\vartheta_4(q_D)}\\ h_M(q_D) &= \frac{1}{2\mathrm{i}}\vartheta_3(q_D)\vartheta_4(q_D)\\ u_M(q_D) &= \frac{1}{2}\frac{\vartheta_3^4(q_D)+\vartheta_4^4(q_D)}{(\vartheta_3(q_D)\vartheta_4(q_D))^2}\\ T_M(q) &= -\frac{1}{24}\frac{E_2(q_D)}{h_M^2(q_D)}+\frac{1}{3}u_M(q_D) \end{aligned} \qquad [38]$$

The contribution at $u=-1$ is related to the contribution at $u=1$ by a $u\to -u$ *symmetry:*

$$Z_{u=-1}(p,S) = \mathrm{e}^{-2\pi\mathrm{i}\lambda_0^2}\mathrm{i}^{(\chi+\sigma)/4}Z_{u=1}(-p,-\mathrm{i}S) \qquad [39]$$

If the manifold has $b_2^+(X)>1$ and is of Seiberg–Witten simple type, [37] reduces to

$$(-1)^{\chi_h}2^{1+7\chi/4+11\sigma/4}\mathrm{e}^{2p+S^2/2}\mathrm{e}^{-2(S,\lambda)} \times\mathrm{e}^{2\mathrm{i}\pi(\lambda_0^2-\lambda_0\cdot\lambda)}\mathrm{SW}(\lambda) \qquad [40]$$

This leads to Witten's "magic formula" [25] which expresses the Donaldson invariants in terms of Seiberg–Witten invariants.

Other Applications of the u-Plane Integral

The u-plane integral makes possible to derive other results on the Donaldson–Witten generating functional.

The blow-up formula. This relates the function Z_{DW} on X to Z_{DW} on the blown-up manifold $\widehat{X}$. The u-plane integral leads directly to the general blow-up formula of Fintushel and Stern (1996).

Direct evaluations. The u-plane integral can be evaluated directly in many cases, and this leads to explicit formulas for the Donaldson–Witten generating functional of certain 4-manifolds with $b_2^+(X)=1$, on certain chambers, and in terms of modular forms. For example, there are explicit formulas for the Donaldson–Witten generating functional of product ruled surfaces of the form $S^2\times\Sigma_g$ in the limiting chambers in which S^2 or Σ_g are very small (Moore and Witten 1998, Mariño and Moore 1999). Moore and Witten (1998) have also derived an explicit formula for the Donaldson invariants of CP^2 in terms of Hurwitz class numbers.

Extensions of Donaldson–Witten Theory

Donaldson–Witten theory is a twisted version of SU(2), $\mathcal{N}=2$ Yang–Mills theory. The twisting of more general $\mathcal{N}=2$ gauge theories, involving other gauge groups and/or matter content, leads to other topological field theories that give interesting generalizations of Donaldson–Witten theory. We now briefly list some of these extensions and their most important properties.

Higher-rank theories. The extension of Donaldson–Witten to other gauge groups has been studied in detail in Mariño and Moore (1998b) and Losev *et al.* (1998). One can study the higher-rank generalization of the u-plane integral, and as shown in Mariño and Morre (1998b), this leads to a fairly explicit formula for the Donaldson–Witten generating function in the SU(N) case, for manifolds with $b_2^+>1$ and of Seiberg–Witten simple type. Mathematically, higher-rank generalizations of Donaldson theory turn out to be much more complicated, but they can be studied. In particular, higher-rank generalizations of the Donaldson invariants can be defined and computed (Kronheimer 2004), and the results so far agree with the predictions of Mariño and Moore (1998b). Unfortunately it seems that these higher-rank generalizations do not contain new topological information, besides the one encoded in the Seiberg–Witten invariants.

Theories with matter. Twisted SU(2), $\mathcal{N}=2$ theories with hypermultiplets lead to generalizations of Donaldson–Witten theory involving nonabelian

monopole equations (see Mariño (1997) and Labastida and Mariño (2005) for a review of these models and some of their properties). The u-plane integral leads to explicit formulas for the generating functionals of these theories, which for manifolds of $b_2^+ > 1$ can be written in terms of Seiberg–Witten invariants. Again, no new topological information seems to be encoded in these theories. One can however exploit new physical phenomena arising in the theories with hypermultiplets (in particular, the presence of superconformal points) to obtain new information about the Seiberg–Witten invariants (see Mariño *et al.* (1999) for these developments).

Vafa–Witten theory. The so-called Vafa–Witten theory is a close cousin of Donaldson–Witten theory, and was introduced by Vafa and Witten (1994) as a topological twist of $\mathcal{N}=4$ Yang–Mills theory. In some cases, the partition function of this theory counts the Euler characteristic of the moduli space of instantons on the 4-manifold X. For a review of some properties of this theory, see Lozano (1999).

See also: Duality in Topological Quantum Field Theory; Mathai–Quillen Formalism; Seiberg–Witten Theory; Topological Quantum Field Theory: Overview.

Further Reading

Birmingham D, Blau M, Rakowski M, and Thompson G (1991) Topological field theories. *Physics Reports* 209: 129.

Cordes S, Moore G, and Rangoolam S (1994) Lectures on 2D Yang–Mills theory, equivariant cohomology and topological field theory (arXiv:hep-th/9411210).

Donaldson SK (1990) Polynomial invariants for smooth four-manifolds. *Topology* 29: 257.

Donaldson SK and Kronheimer PB (1990) *The Geometry of Four-Manifolds.* Oxford: Clarendon.

Fintushel R and Stern RJ (1996) The blowup formula for Donaldson invariants. *Annals of Mathematics* 143: 529 (arXiv:alg-geom/9405002).

Friedman R and Morgan JW (1991) *Smooth Four-Manifolds and Complex Surfaces.* New York: Springer.

Göttsche L (1996) Modular forms and Donaldson invariants for four-manifolds with $b_+ = 1$. *Journal of American Mathematical Society* 9: 827 (arXiv:alg-geom/9506018).

Kronheimer P (2004) Four-manifold invariants from higher-rank bundles (arXiv:math.GT/0407518).

Kronheimer P and Mrowka T (1994) Recurrence relations and asymptotics for four-manifolds invariants. *Bulletin of the American Mathematical Society* 30: 215.

Kronheimer P and Mrowka T (1995) Embedded surfaces and the structure of Donaldson's polynomials invariants. *Journal of Differential Geometry* 33: 573.

Labastida J and Mariño M (2005) *Topological Quantum Field Theory and Four-Manifolds.* New York: Springer.

Losev A, Nekrasov N, and Shatashvili S (1998) Issues in topological gauge theory. *Nuclear Physics B* 534: 549 (arXiv:hep-th/9711108).

Lozano C (1999) Duality in topological quantum field theories (arXiv:hep-th/9907123).

Mariño M (1997) The geometry of supersymmetric gauge theories in four dimensions (arXiv:hep-th/9701128).

Mariño M and Moore G (1998a) Integrating over the Coulomb branch in $\mathcal{N}=2$ gauge theory. *Nuclear Physics Proceedings Supplement* 68: 336 (arXiv:hep-th/9712062).

Mariño M and Moore GW (1998b) The Donaldson–Witten function for gauge groups of rank larger than one. *Communications in Mathematical Physics* 199: 25 (arXiv:hep-th/9802185).

Mariño M and Moore GW (1999) Donaldson invariants for non-simply connected manifolds. *Communications in Mathematical Physics* 203: 249 (arXiv:hep-th/9804104).

Mariño M, Moore GW, and Peradze G (1999) Superconformal invariance and the geography of four-manifolds. *Communications in Mathematical Physics* 205: 691 (arXiv:hep-th/9812055).

Moore G and Witten E (1998) Integration over the u-plane in Donaldson theory. *Advances in Theoretical and Mathematical Society* 1: 298 (arXiv:hep-th/9709193).

Seiberg N and Witten E (1994a) Electric – magnetic duality, monopole condensation, and confinement in $\mathcal{N}=2$ supersymmetric Yang–Mills theory. *Nuclear Physics B* 426: 19.

Seiberg N and Witten E (1994b) Electric – magnetic duality, monopole condensation, and confirement in $\mathcal{N}=2$ supersymmetric Yang–Mills theory – erratum. *Nuclear Physics B* 430: 485 (arXiv:hep-th/9407087).

Stern R (1998) Computing Donaldson invariants. In: *Gauge Theory and the Topology of Four-Manifolds*, IAS/Park City Mathematical Series. Providence, RI: American Mathematical Society.

Vafa C and Witten E (1994) A strong coupling test of S duality. *Nuclear Physics B* 431: 3 (arXiv:hep-th/9408074).

Witten E (1988) Topological quantum field theory. *Communications in Mathematical Physics* 117: 353.

Witten E (1994a) Supersymmetric Yang–Mills theory on a four manifold. *Journal of Mathematical Physics* 35: 5101 (arXiv:hep-th/9403195).

Witten E (1994b) Monopoles and four manifolds. *Mathematical Research Letters* 1: 769 (arXiv:hep-th/9411102).

Witten E (1995) On S duality in Abelian gauge theory. *Selecta Mathematica* 1: 383 (arXiv:hep-th/9505186).

Effective Field Theories

G Ecker, Universität Wien, Vienna, Austria

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Effective field theories (EFTs) are the counterpart of the "theory of everything." They are the field theoretical implementation of the quantum ladder: heavy degrees of freedom need not be included among the quantum fields of an EFT for a description of low-energy phenomena. For example, we do not need quantum gravity to understand the hydrogen atom nor does chemistry depend upon the structure of the electromagnetic interaction of quarks.

EFTs are approximations by their very nature. Once the relevant degrees of freedom for the problem at hand have been established, the corresponding EFT is usually treated perturbatively. It does not make much sense to search for an exact solution of the Fermi theory of weak interactions. In the same spirit, convergence of the perturbative expansion in the mathematical sense is not an issue. The asymptotic nature of the expansion becomes apparent once the accuracy is reached where effects of the underlying "fundamental" theory cannot be neglected any longer. The range of applicability of the perturbative expansion depends on the separation of energy scales that define the EFT.

EFTs pervade much of modern physics. The effective nature of the description is evident in atomic and condensed matter physics. The present article will be restricted to particle physics, where EFTs have become important tools during the last 25 years.

Classification of EFTs

A first classification of EFTs is based on the structure of the transition from the "fundamental" (energies $>\Lambda$) to the "effective" level (energies $<\Lambda$).

1. *Complete decoupling* The fundamental theory contains heavy and light degrees of freedom. Under very general conditions (decoupling theorem, Appelquist and Carazzone 1975) the effective Lagrangian for energies $\ll\Lambda$, depending only on light fields, takes the form

$$\mathcal{L}_{\text{eff}} = \mathcal{L}_{d\leq 4} + \sum_{d>4} \frac{1}{\Lambda^{d-4}} \sum_{i_d} g_{i_d} O_{i_d} \qquad [1]$$

The heavy fields with masses $>\Lambda$ have been "integrated out" completely. $\mathcal{L}_{d\leq 4}$ contains the potentially renormalizable terms with operator dimension $d \leq 4$ (in natural mass units where Bose and Fermi fields have $d=1$ and 3/2, respectively), the g_{i_d} are coupling constants and the O_{i_d} are monomials in the light fields with operator dimension d. In a slightly misleading notation, $\mathcal{L}_{d\leq 4}$ consists of relevant and marginal operators, whereas the O_{i_d} $(d>4)$ are denoted irrelevant operators. The scale Λ can be the mass of a heavy field (e.g., M_W in the Fermi theory of weak interactions) or it reflects the short-distance structure in a more indirect way.

2. *Partial decoupling* In contrast to the previous case, the heavy fields do not disappear completely from the EFT but only their high-momentum modes are integrated out. The main area of application is the physics of heavy quarks. The procedure involves one or several field redefinitions introducing a frame dependence. Lorentz invariance is not manifest but implies relations between coupling constants of the EFT (reparametrization invariance).

3. *Spontaneous symmetry breaking* The transition from the fundamental to the effective level occurs via a phase transition due to spontaneous symmetry breaking generating (pseudo-)Goldstone bosons. A spontaneously broken symmetry relates processes with different numbers of Goldstone bosons. Therefore, the distinction between renormalizable $(d \leq 4)$ and nonrenormalizable $(d>4)$ parts in the effective Lagrangian [1] becomes meaningless. The effective Lagrangian of type 3 is generically nonrenormalizable. Nevertheless, such Lagrangians define perfectly consistent quantum field theories at sufficiently low energies. Instead of the operator dimension as in [1], the number of derivatives of the fields and the number of symmetry-breaking

Notes:

insertions distinguish successive terms in the Lagrangian. The general structure of effective Lagrangians with spontaneously broken symmetries is largely independent of the specific physical realization (universality). There are many examples in condensed matter physics, but the two main applications in particle physics are electroweak symmetry breaking and chiral perturbation theory (both discussed later) with the spontaneously broken global chiral symmetry of quantum chromodynamics QCD.

Another classification of EFTs is related to the status of their coupling constants.

A. *Coupling constants can be determined by matching the EFT with the underlying theory at short distances.* The underlying theory is known and Green functions can be calculated perturbatively at energies $\sim\Lambda$ both in the fundamental and in the effective theory. Identifying a minimal set of Green functions fixes the coupling constants g_{i_d} in eqn [1] at the scale Λ. Renormalization group equations can then be used to run the couplings down to lower scales. The nonrenormalizable terms in the Lagrangian [1] can be fully included in the perturbative analysis.

B. *Coupling constants are constrained by symmetries only.*
- The underlying theory and therefore also the EFT coupling constants are unknown. This is the case of the standard model (SM) (see the next section). A perturbative analysis beyond leading order only makes sense for the known renormalizable part $\mathcal{L}_{d\leq 4}$. The nonrenormalizable terms suppressed by powers of Λ are considered at tree level only. The associated coupling constants g_{i_d} serve as bookmarks for new physics. Usually, but not always (cf., e.g., the subsection "Noncommutative spacetime"), the symmetries of $\mathcal{L}_{d\leq 4}$ are assumed to constrain the couplings.
- The matching cannot be performed in perturbation theory even though the underlying theory is known. This is the generic situation for EFTs of type 3 involving spontaneous symmetry breaking. The prime example is chiral perturbation theory as the EFT of QCD at low energies.

The SM as an EFT

With the possible exception of the scalar sector to be discussed in the subsection "Electroweak symmetry breaking" the SM is very likely the renormalizable part of an EFT of type 1B. Except for nonzero neutrino masses, the SM Lagrangian $\mathcal{L}_{d\leq 4}$ in [1] accounts for physics up to energies of roughly the Fermi scale $G_F^{-1/2} \simeq 300\,\text{GeV}$.

Since the SM works exceedingly well up to the Fermi scale where the electroweak gauge symmetry is spontaneously broken it is natural to assume that the operators O_{i_d} with $d > 4$, made up from fields representing the known degrees of freedom and including a single Higgs doublet in the SM proper, should be gauge invariant with respect to the full SM gauge group $\mathrm{SU(3)_c \times SU(2)_L \times U(1)_Y}$. An almost obvious constraint is Lorentz invariance that will be lifted in the next subsection, however.

These requirements limit the Lagrangian with operator dimension $d=5$ to a single term (except for generation multiplicity), consisting only of a left-handed lepton doublet L_L and the Higgs doublet Φ:

$$O_{d=5} = \epsilon_{ij}\epsilon_{kl} L_{iL}^{\top} C^{-1} L_{kL} \Phi_j \Phi_l + \text{h.c.} \qquad [2]$$

This term violates lepton number and generates nonzero Majorana neutrino masses. For a neutrino mass of 1 eV, the scale Λ would have to be of the order of $10^{13}\,\text{GeV}$ if the associated coupling constant in the EFT Lagrangian [1] is of order 1.

In contrast to the simplicity for $d=5$, the list of gauge-invariant operators with $d=6$ is enormous. Among them are operators violating baryon or lepton number that must be associated with a scale much larger than 1 TeV. To explore the territory close to present energies, it therefore makes sense to impose baryon and lepton number conservation on the operators with $d=6$. Those operators have all been classified (Buchmüller and Wyler 1986) and the number of independent terms is of the order of 80. They can be grouped in three classes.

The first class consists of gauge and Higgs fields only. The corresponding EFT Lagrangian has been used to parametrize new physics in the gauge sector constrained by precision data from LEP. The second class consists of operators bilinear in fermion fields, with additional gauge and Higgs fields to generate $d=6$. Finally, there are four-fermion operators without other fields or derivatives. Some of the operators in the last two groups are also constrained by precision experiments, with a certain hierarchy of limits. For lepton and/or quark flavor conserving terms, the best limits on Λ are in the few TeV range, whereas the absence of neutral flavor changing processes yields lower bounds on Λ that are several orders of magnitude larger. If there is new physics in the TeV range flavor changing neutral transitions must be strongly suppressed, a powerful constraint on model building.

It is amazing that the most general renormalizable Lagrangian with the given particle content accounts

for almost all experimental results in such an impressive manner. Finally, we recall that many of the operators of dimension 6 are also generated in the SM via radiative corrections. A necessary condition for detecting evidence for new physics is therefore that the theoretical accuracy of radiative corrections matches or surpasses the experimental precision.

Noncommutative Spacetime

Noncommutative geometry arises in some string theories and may be expected on general grounds when incorporating gravity into a quantum field theory framework. The natural scale of noncommutative geometry would be the Planck scale in this case without observable consequences at presently accessible energies. However, as in theories with large extra dimensions the characteristic scale Λ_{NC} could be significantly smaller. In parallel to theoretical developments to define consistent noncommutative quantum field theories (short for quantum field theories on noncommutative spacetime), a number of phenomenological investigations have been performed to put lower bounds on Λ_{NC}.

Noncommutative geometry is a deformation of ordinary spacetime where the coordinates, represented by Hermitian operators $\hat{x}_\mu$, do not commute:

$$[\hat{x}_\mu, \hat{x}_\nu] = i\theta_{\mu\nu} \qquad [3]$$

The antisymmetric real tensor $\theta_{\mu\nu}$ has dimensions length^2 and it can be interpreted as parametrizing the resolution with which spacetime can be probed. In practically all applications, $\theta_{\mu\nu}$ has been assumed to be a constant tensor and we may associate an energy scale Λ_{NC} with its nonzero entries:

$$\Lambda_{NC}^{-2} \sim \theta_{\mu\nu} \qquad [4]$$

There is to date no unique form for the noncommutative extension of the SM. Nevertheless, possible observable effects of noncommutative geometry have been investigated. Not unexpected from an EFT point of view, for energies $\ll\Lambda_{NC}$, noncommutative field theories are equivalent to ordinary quantum field theories in the presence of nonstandard terms containing $\theta_{\mu\nu}$ (Seiberg–Witten map). Practically all applications have concentrated on effects linear in $\theta_{\mu\nu}$.

Kinetic terms in the Lagrangian are in general unaffected by the noncommutative structure. New effects arise therefore mainly from renormalizable $d=4$ interactions terms. For example, the Yukawa coupling $g_Y\bar{\psi}\psi\phi$ generates the following interaction linear in $\theta_{\mu\nu}$:

$$\mathcal{L}_Y^{NC} = g_Y\theta_{\mu\nu}\left(\partial^\mu\bar{\psi}\partial^\nu\psi\phi + \partial^\mu\bar{\psi}\psi\partial^\nu\phi + \bar{\psi}\partial^\mu\psi\partial^\nu\phi\right) \qquad [5]$$

These interaction terms have operator dimension 6 and they are suppressed by $\theta_{\mu\nu} \sim \Lambda_{NC}^{-2}$. The major difference to the previous discussion on physics beyond the SM is that there is an intrinsic violation of Lorentz invariance due to the constant tensor $\theta_{\mu\nu}$. In contrast to the previous analysis, the terms with dimension $d > 4$ do not respect the symmetries of the SM.

If $\theta_{\mu\nu}$ is indeed constant over macroscopic distances, many tests of Lorentz invariance can be used to put lower bounds on Λ_{NC}. Among the exotic effects investigated are modified dispersion relations for particles, decay of high-energy photons, charged particles producing Cerenkov radiation in vacuum, birefringence of radiation, a variable speed of light, etc. A generic signal of noncommutativity is the violation of angular momentum conservation that can be searched for at the Large Hadron Collider (LHC) and at the next linear collider.

Lacking a unique noncommutative extension of the SM, unambiguous lower bounds on Λ_{NC} are difficult to establish. However, the range $\Lambda_{NC} \lesssim 10\,\text{TeV}$ is almost certainly excluded. An estimate of the induced electric dipole moment of the electron (noncommutative field theories violate CP in general to first order in $\theta_{\mu\nu}$) yields $\Lambda_{NC} \gtrsim 100\,\text{TeV}$. On the other hand, if the SM were CP invariant, noncommutative geometry would be able to account for the observed CP violation in $K^0 - \overline{K^0}$ mixing for $\Lambda_{NC} \sim 2\,\text{TeV}$.

Electroweak Symmetry Breaking

In the SM, electroweak symmetry breaking is realized in the simplest possible way through renormalizable interactions of a scalar Higgs doublet with gauge bosons and fermions, a gauged version of the linear σ model.

The EFT version of electroweak symmetry breaking (EWEFT) uses only the experimentally established degrees of freedom in the SM (fermions and gauge bosons). Spontaneous gauge symmetry breaking is realized nonlinearly, without introducing additional scalar degrees of freedom. It is a low-energy expansion where energies and masses are assumed to be small compared to the symmetry-breaking scale. From both perturbative and nonperturbative arguments we know that this scale cannot be much bigger than 1 TeV. The Higgs model can be viewed as a specific example of an EWEFT as long as the Higgs boson is not too light (heavy-Higgs scenario).

The lowest-order effective Lagrangian takes the following form:

$$\mathcal{L}_{EWSB}^{(2)} = \mathcal{L}_B + \mathcal{L}_F \qquad [6]$$

where $\mathcal{L}_F$ contains the gauge-invariant kinetic terms for quarks and leptons including mass terms. In addition to the kinetic terms for the gauge bosons W_μ, B_μ, the bosonic Lagrangian $\mathcal{L}_B$ contains the characteristic lowest-order term for the would-be-Goldstone bosons:

$$\mathcal{L}_B = \mathcal{L}^{\text{kin}}_{\text{gauge}} + \frac{v^2}{4}\langle D_\mu U^\dagger D^\mu U\rangle \qquad [7]$$

with the gauge-covariant derivative

$$\begin{aligned} D_\mu U &= \partial_\mu U - igW_\mu U + ig' U\hat{B}_\mu \\ W_\mu &= \frac{\vec{\tau}}{2}W_\mu, \qquad \hat{B}_\mu = \frac{\tau_3}{2}B_\mu \end{aligned} \qquad [8]$$

where $\langle\ldots\rangle$ denotes a (two-dimensional) trace. The matrix field $U(\phi)$ carries the nonlinear representation of the spontaneously broken gauge group and takes the value $U=1$ in the unitary gauge. The Lagrangian [6] is invariant under local $SU(2)_L \times U(1)_Y$ transformations:

$$\begin{aligned} W_\mu &\to g_L W_\mu g_L^\dagger + \frac{i}{g} g_L \partial_\mu g_L^\dagger \\ \hat{B}_\mu &\to \hat{B}_\mu + \frac{i}{g'} g_R \partial_\mu g_R^\dagger \\ f_L &\to g_L f_L, \quad f_R \to g_R f_R, \quad U \to g_L U g_R^\dagger \end{aligned} \qquad [9]$$

with

$$\begin{aligned} g_L(x) &= \exp(i\vec{\alpha}_L(x)\vec{\tau}/2) \\ g_R(x) &= \exp(i\alpha_Y(x)\tau_3/2) \end{aligned}$$

and $f_{L(R)}$ are quark and lepton fields grouped in doublets.

As is manifest in the unitary gauge $U=1$, the lowest-order Lagrangian of the EWEFT just implements the tree-level masses of gauge bosons ($M_W = M_Z\cos\theta_W = vg/2$, $\tan\theta_W = g'/g$) and fermions but does not carry any further information about the underlying mechanism of spontaneous gauge symmetry breaking. This information is first encoded in the couplings a_i of the next-to-leading-order Lagrangian

$$\mathcal{L}^{(4)}_{\text{EWSB}} = \sum_{i=0}^{14} a_i O_i \qquad [10]$$

with monomials O_i of $O(p^4)$ in the low-energy expansion. The Lagrangian [10] is the most general CP and $SU(2)_L \times U(1)_Y$ invariant Lagrangian of $O(p^4)$.

Instead of listing the full Lagrangian, we display three typical examples:

$$\begin{aligned} O_0 &= \frac{v^2}{4}\langle TV_\mu\rangle^2 \\ O_3 &= -g\langle W_{\mu\nu}[V^\mu, V^\nu]\rangle \\ O_5 &= \langle V_\mu V^\mu\rangle^2 \end{aligned} \qquad [11]$$

where

$$\begin{aligned} T &= U\tau_3 U^\dagger, \qquad V_\mu = D_\mu U U^\dagger \\ W_{\mu\nu} &= \frac{i}{g}[\partial_\mu - igW_\mu, \partial_\nu - igW_\nu] \end{aligned} \qquad [12]$$

In the unitary gauge, the monomials O_i reduce to polynomials in the gauge fields. The three examples in eqn [11] start with quadratic, cubic, and quartic terms in the gauge fields, respectively. The strongest constraints exist for the coefficients of quadratic contributions from the Large Electron–Positron collider LEP1, less restrictive ones for the cubic self-couplings from LEP2, and none so far for the quartic ones.

Heavy-Quark Physics

EFTs in this section are derived from the SM and they are of type 2A in the classification introduced previously. In a first step, one integrates out W, Z, and top quark. Evolving down from M_W to m_b, large logarithms $\alpha_s(m_b)\ln(M_W^2/m_b^2)$ are resummed into the Wilson coefficients. At the scale of the b-quark, QCD is still perturbative, so that at least a part of the amplitudes is calculable in perturbation theory. To separate the calculable part from the rest, the EFTs below perform an expansion in $1/m_Q$, where m_Q is the mass of the heavy quark.

Heavy-quark EFTs offer several important advantages.

1. Approximate symmetries that are hidden in full QCD appear in the expansion in $1/m_Q$.
2. Explicit calculations simplify in general, for example, the summing of large logarithms via renormalization group equations.
3. The systematic separation of hard and soft effects for certain matrix elements (factorization) can be achieved much more easily.

Heavy-Quark Effective Theory

Heavy-quark effective theory (HQET) is reminiscent of the Foldy–Wouthuysen transformation (nonrelativistic expansion of the Dirac equation). It is a systematic expansion in $1/m_Q$, when $m_Q \gg \Lambda_{\text{QCD}}$, the scale parameter of QCD. It can be applied to processes where the heavy quark remains essentially on shell: its velocity v changes only by small amounts $\sim\Lambda_{\text{QCD}}/m_Q$. In the hadron rest frame, the heavy quark is almost at rest and acts as a quasistatic source of gluons.

More quantitatively, one writes the heavy-quark momentum as $p^\mu = m_Q v^\mu + k^\mu$, where v is the hadron 4-velocity ($v^2=1$) and k is a residual

momentum of $O(\Lambda_{\rm QCD})$. The heavy quark field $Q(x)$ is then decomposed with the help of energy projectors $P_v^{\pm}=(1\pm\not{v})/2$ and employing a field redefinition:

$$\begin{aligned} Q(x) &= \mathrm{e}^{-im_Q v\cdot x}(h_v(x)+H_v(x)) \\ h_v(x) &= \mathrm{e}^{im_Q v\cdot x}P_v^+ Q(x) \\ H_v(x) &= \mathrm{e}^{im_Q v\cdot x}P_v^- Q(x) \end{aligned} \qquad [13]$$

In the hadron rest frame, $h_v(x)$ and $H_v(x)$ correspond to the upper and lower components of $Q(x)$, respectively. With this redefinition, the heavy-quark Lagrangian is expressed in terms of a massless field h_v and a "heavy" field H_v:

$$\begin{aligned} \mathcal{L}_Q &= \bar{Q}(i\not{D}-m_Q)Q \\ &= \overline{h_v}\, iv\cdot D\, h_v - \overline{H_v}(iv\cdot D+2m_Q)H_v \\ &\quad + \text{mixed terms} \end{aligned} \qquad [14]$$

At the semiclassical level, the field H_v can be eliminated by using the QCD field equation $(i\not{D}-m_Q)Q=0$ yielding the nonlocal expression

$$\mathcal{L}_Q = \overline{h_v}\, iv\cdot D h_v + \overline{h_v}\, i\not{D}_\perp \frac{1}{iv\cdot D+2m_Q-i\epsilon} i\not{D}_\perp h_v \qquad [15]$$

with $D_\perp^\mu=(g^{\mu\nu}-v^\mu v^\nu)D_\nu$. The field redefinition in [13] ensures that, in the heavy-hadron rest frame, derivatives of h_v give rise to small momenta of $O(\Lambda_{\rm QCD})$ only. The Lagrangian [15] is the starting point for a systematic expansion in m_Q.

To leading order in $1/m_Q (Q=b,c)$, the Lagrangian

$$\mathcal{L}_{b,c} = \overline{b_v}\, iv\cdot D\, b_v + \overline{c_v} iv\cdot D c_v \qquad [16]$$

exhibits two important approximate symmetries of HQET: the flavor symmetry $SU(2)_F$ relating heavy quarks moving with the same velocity and the heavy-quark spin symmetry generating an overall SU(4) spin-flavor symmetry. The flavor symmetry is obvious and the spin symmetry is due to the absence of Dirac matrices in [16]: both spin degrees of freedom couple to gluons in the same way. The simplest spin-symmetry doublet consists of a pseudoscalar meson H and the associated vector meson H^*. Denoting the doublet by $\mathcal{H}$, the matrix elements of the heavy-to-heavy transition current are determined to leading order in $1/m_Q$ by a single form factor, up to Clebsch–Gordan coefficients:

$$\langle\mathcal{H}(v')|\overline{h_{v'}}\Gamma h_v|\mathcal{H}(v)\rangle \sim \xi(v\cdot v') \qquad [17]$$

Γ is an arbitrary combination of Dirac matrices and the form factor ξ is the so-called Isgur–Wise function. Moreover, since $\overline{h_v}\gamma^\mu h_v$ is the Noether current of heavy-flavor symmetry, the Isgur–Wise function is fixed in the no-recoil limit $v'=v$ to be $\xi(v\cdot v'=1)=1$. The semileptonic decays $B\to Dl\nu_l$ and $B\to D^* l\nu_l$ are therefore governed by a single normalized form factor to leading order in $1/m_Q$, with important consequences for the determination of the Cabibbo–Kobayashi–Maskawa (CKM) matrix element V_{cb}.

The HQET Lagrangian is superficially frame dependent. Since the SM is Lorentz invariant, the HQET Lagrangian must be independent of the choice of the frame vector v. Therefore, a shift in v accompanied by corresponding shifts of the fields h_v and of the covariant derivatives must leave the Lagrangian invariant. This reparametrization invariance is unaffected by renormalization and it relates coefficients with different powers in $1/m_Q$.

Soft-Collinear Effective Theory

HQET is not applicable in heavy-quark decays where some of the light particles in the final state have momenta of $O(m_Q)$, for example, for inclusive decays like $B\to X_s\gamma$ or exclusive ones like $B\to\pi\pi$. In recent years, a systematic heavy-quark expansion for heavy-to-light decays has been set up in the form of soft-collinear effective theory (SCET).

SCET is more complicated than HQET because now the low-energy theory involves more than one scale. In the SCET Lagrangian, a light quark or gluon field is represented by several effective fields. In addition to the soft fields h_v in [15], the so-called collinear fields enter that have large energy and carry large momentum in the direction of the light hadrons in the final state.

In addition to the frame vector v of HQET ($v=(1,0,0,0)$ in the heavy-hadron rest frame), SCET introduces a lightlike reference vector n in the direction of the jet of energetic light particles (for inclusive decays), for example, $n=(1,0,0,1)$. All momenta p are decomposed in terms of light-cone coordinates $(p_+, p_-, p_\perp)$ with

$$p^\mu = \frac{n\cdot p}{2}\bar{n}^\mu + \frac{\bar{n}\cdot p}{2}n^\mu + p_\perp^\mu = p_+^\mu + p_-^\mu + p_\perp^\mu \qquad [18]$$

where $\bar{n}=2v-n=(1,0,0,-1)$. For large energies, the three light-cone components are widely separated, with $p_-=O(m_Q)$ being large while $p_\perp$ and p_+ are small. Introducing a small parameter $\lambda\sim p_\perp/p_-$, the light-cone components of (hard-)collinear particles scale like $(p_+,p_-,p_\perp)=m_Q(\lambda^2,1,\lambda)$. Thus, there are three different scales in the problem compared to only two in HQET. For exclusive decays, the situation is even more involved.

The SCET Lagrangian is obtained from the full theory by an expansion in powers of λ. In addition to the heavy quark field h_v, one introduces soft as well as collinear quark and gluon fields by field

redefinitions so that the various fields have momentum components that scale appropriately with λ.

Similar to HQET, the leading-order Lagrangian of SCET exhibits again approximate symmetries that can lead to a reduction of form factors describing heavy-to-light decays. As in HQET, reparametrization invariance implements Lorentz invariance and results in stringent constraints on subleading corrections in SCET.

An important result of SCET is the proof of factorization theorems to all orders in α_s. For inclusive decays, the differential rate is of the form

$$d\Gamma \sim HJ \times S \qquad [19]$$

where H contains the hard corrections. The so-called jet function J sensitive to the collinear region is convoluted with the shape function S representing the soft contributions. At leading order, the shape function drops out in the ratio of weighted decay spectra for $B \to X_u l \nu_l$ and $B \to X_s \gamma$ allowing for a determination of the CKM matrix element V_{ub}. Factorization theorems have become available for an increasing number of processes, most recently also for exclusive decays of B into two light mesons.

Nonrelativistic QCD

In HQET the kinetic energy of the heavy quark appears as a small correction of $O(\Lambda^2_{\rm QCD}/m_Q)$. For systems with more than one heavy quark, the kinetic energy cannot be treated as a perturbation in general. For instance, the virial theorem implies that the kinetic energy in quarkonia $\bar{Q}Q$ is of the same order as the binding energy of the bound state.

NRQCD, the EFT for heavy quarkonia, is an extension of HQET. The Lagrangian for NRQCD coincides with HQET in the bilinear sector of the heavy-quark fields but it also includes quartic interactions between quarks and antiquarks. The relevant expansion parameter in this case is the relative velocity between Q and $\bar{Q}$. In contrast to HQET, there are at least three widely separate scales in heavy quarkonia: in addition to m_Q, the relative momentum of the bound quarks $p \sim m_Q v$ with $v \ll 1$ and the typical kinetic energy $E \sim m_Q v^2$. The main challenges are to derive the quark–antiquark potential directly from QCD and to describe quarkonium production and decay at collider experiments. In the abelian case, the corresponding EFT for quantum electrodynamics (QED) is called NRQED that has been used to study electromagnetically bound systems like the hydrogen atom, positronium, muonium, etc.

In NRQCD only the hard degrees of freedom with momenta $\sim m_Q$ are integrated out. Therefore, NRQCD is not enough for a systematic computation of heavy-quarkonium properties. Because the nonrelativistic fluctuations of order $m_Q v$ and $m_Q v^2$ have not been separated, the power counting in NRQCD is ambiguous in higher orders.

To overcome those deficiencies, two approaches have been put forward: potential NRQCD (pNRQCD) and velocity NRQCD (vNRQCD). In pNRQCD, a two-step procedure is employed for integrating out quark and gluon degrees of freedom:

$$\begin{array}{ll} \text{QCD} & \Lambda > m_Q \\ \Downarrow & \\ \text{NRQCD} & m_Q > \Lambda > m_Q v \\ \Downarrow & \\ \text{pNRQCD} & m_Q v > \Lambda > m_Q v^2 \end{array}$$

The resulting EFT derives its name from the fact that the four-quark interactions generated in the matching procedure are the potentials that can be used in Schrödinger perturbation theory. It is claimed that pNRQCD can also be used in the nonperturbative domain where $\alpha_s(m_Q v^2)$ is of order 1 or larger. The advantage would be that also charmonium becomes accessible to a systematic EFT analysis.

The alternative approach of vNRQCD is only applicable in the fully perturbative regime when $m_Q \gg m_Q v \gg m_Q v^2 \gg \Lambda_{\rm QCD}$ is valid. It separates the different degrees of freedom in a single step leaving only ultrasoft energies and momenta of $O(m_Q v^2)$ as continuous variables. The separation of larger scales proceeds in a similar fashion as in HQET via field redefinitions. A systematic nonrelativistic power counting in the velocity v is implemented.

The Standard Model at Low Energies

At energies below 1 GeV, hadrons – rather than quarks and gluons – are the relevant degrees of freedom. Although the strong interactions are highly nonperturbative in the confinement region, Green functions and amplitudes are amenable to a systematic low-energy expansion. The key observation is that the QCD Lagrangian with $N_f = 2$ or 3 light quarks,

$$\begin{aligned} \mathcal{L}_{\rm QCD} &= \bar{q}(i\not{D} - \mathcal{M}_q)q - \tfrac{1}{4}G^{\alpha}_{\mu\nu}G^{\alpha\mu\nu} + \mathcal{L}_{\text{heavy quarks}} \\ &= \overline{q_{\rm L}} i\not{D} q_{\rm L} + \overline{q_{\rm R}} i\not{D} q_{\rm R} - \overline{q_{\rm L}}\mathcal{M}_q q_{\rm R} \\ &\quad - \overline{q_{\rm R}}\mathcal{M}_q q_{\rm L} + \cdots \\ q_{\rm R,L} &= \tfrac{1}{2}(1 \pm \gamma_5)q, \quad q^{\rm T} = (ud[s]) \end{aligned} \qquad [20]$$

exhibits a global symmetry

$$\underbrace{\mathrm{SU}(N_f)_{\rm L} \times \mathrm{SU}(N_f)_{\rm R}}_{\text{chiral group } G} \times \mathrm{U}(1)_{\rm V} \times \mathrm{U}(1)_{\rm A} \qquad [21]$$

in the limit of N_f massless quarks ($\mathcal{M}_q = 0$). At the hadronic level, the quark number symmetry $U(1)_V$ is realized as baryon number. The axial $U(1)_A$ is not a symmetry at the quantum level due to the abelian anomaly.

Although not yet derived from first principles, there are compelling theoretical and phenomenological arguments that the ground state of QCD is not even approximately chirally symmetric. All evidence, such as the existence of relatively light pseudoscalar mesons, points to spontaneous chiral symmetry breaking $G \to SU(N_f)_V$, where $SU(N_f)_V$ is the diagonal subgroup of G. The resulting $N_f^2 - 1$ (pseudo-)Goldstone bosons interact weakly at low energies. In fact, Goldstone's theorem ensures that purely mesonic or single-baryon amplitudes vanish in the chiral limit ($\mathcal{M}_q = 0$) when the momenta of all pseudoscalar mesons tend to zero. This is the basis for a systematic low-energy expansion of Green functions and amplitudes. The corresponding EFT (type 3B in our classification) is called chiral perturbation theory (CHPT) (Weinberg 1979, Gasser and Leutwyler 1984, 1985).

Although the construction of effective Lagrangians with nonlinearly realized chiral symmetry is well understood, there are some subtleties involved. First of all, there may be terms in a chiral-invariant action that cannot be written as the four-dimensional integral of an invariant Lagrangian. The chiral anomaly for $SU(3) \times SU(3)$ bears witness of this fact and gives rise to the Wess–Zumino–Witten action. A general theorem to account for such exceptional cases is due to D'Hoker and Weinberg (1994). Consider the most general action for Goldstone fields with symmetry group G, spontaneously broken to a subgroup H. The only possible non-G-invariant terms in the Lagrangian that give rise to a G-invariant action are in one-to-one correspondence with the generators of the fifth cohomology group $\mathcal{H}^5(G/H;\mathbb{R})$ of the coset manifold G/H. For the relevant case of chiral $SU(N)$, the coset space $SU(N)_L \times SU(N)_R/SU(N)_V$ is itself an $SU(N)$ manifold. For $N \geq 3, \mathcal{H}^5(SU(N);\mathbb{R})$ has a single generator that corresponds precisely to the Wess–Zumino–Witten term.

At a still deeper level, one may ask whether chiral-invariant Lagrangians are sufficient (except for the anomaly) to describe the low-energy structure of Green functions as dictated by the chiral Ward identities of QCD. To be able to calculate such Green functions in general, the global chiral symmetry of QCD is extended to a local symmetry by the introduction of external gauge fields. The following invariance theorem (Leutwyler 1994) provides an answer to the above question. Except for the anomaly, the most general solution of the Ward identities for a spontaneously broken symmetry in Lorentz-invariant theories can be obtained from gauge-invariant Lagrangians to all orders in the low-energy expansion. The restriction to Lorentz invariance is crucial: the theorem does not hold in general in nonrelativistic effective theories.

Chiral Perturbation Theory

The effective chiral Lagrangian of the SM in the meson sector is displayed in **Table 1**. The lowest-order Lagrangian for the purely strong interactions is given by

$$\mathcal{L}_{p^2} = \frac{F^2}{4}\langle D_\mu U D^\mu U^\dagger \rangle + \frac{F^2 B}{2}\langle (s + ip)U^\dagger + (s - ip)U \rangle \qquad [22]$$

with a covariant derivative $D = \partial_\mu U - i(v_\mu + a_\mu)U + iU(v_\mu - a_\mu)$. The first term has the familiar form [7] of the gauged nonlinear σ model, with the matrix field $U(\phi)$ transforming as $U \to g_R U g_L^\dagger$ under chiral rotations. External fields v_μ, a_μ, s, p are introduced for constructing the generating functional of Green functions of quark currents. To implement explicit chiral symmetry breaking, the scalar field s is set equal to the quark mass matrix $\mathcal{M}_q$ at the end of the calculation.

The leading-order Lagrangian has two free parameters F, B related to the pion decay constant and to the quark condensate, respectively:

$$F_\pi = F[1 + O(m_q)]$$
$$\langle 0|\bar{u}u|0\rangle = -F^2 B[1 + O(m_q)] \qquad [23]$$

The Lagrangian [22] gives rise to $M_\pi^2 = B(m_u + m_d)$ at lowest order. From detailed studies of pion–pion scattering (Colangelo *et al.* 2001), we know that the leading term accounts for at least 94% of the pion mass. This supports the standard counting of CHPT,

Table 1 The effective chiral Lagrangian of the SM in the meson sector

$\mathcal{L}_{\text{chiral order}}$ (*# of LECs*)	*Loop order*
$\mathcal{L}_{p^2}(2) + \mathcal{L}^{\Delta S=1}_{G_F p^2}(2) + \mathcal{L}^{\text{em}}_{e^2 p^0}(1) + \mathcal{L}^{\text{emweak}}_{G_8 e^2 p^0}(1)$	$L=0$
$+\mathcal{L}_{p^4}(10) + \mathcal{L}^{\text{odd}}_{p^5}(32) + \mathcal{L}^{\Delta S=1}_{G_8 p^4}(22) + \mathcal{L}^{\Delta S=1}_{G_{27} p^4}(28)$	$L=1$
$+\mathcal{L}^{\text{em}}_{e^2 p^2}(14) + \mathcal{L}^{\text{emweak}}_{G_8 e^2 p^2}(14) + \mathcal{L}^{\text{leptons}}_{e^2 p}(5)$	
$+\mathcal{L}_{p^6}(90)$	$L=2$

The numbers in brackets refer to the number of independent couplings for $N_f = 3$.. The parameter-free Wess–Zumino–Witten action S_{WZW} that cannot be written as the four-dimensional integral of an invariant Lagrangian must be added.

with quark masses booked as $O(p^2)$ like the two-derivative term in [22].

The effective chiral Lagrangian in **Table 1** contains the following parts:

1. strong interactions: $\mathcal{L}_{p^2}, \mathcal{L}_{p^4}, \mathcal{L}^{\text{odd}}_{p^6}, \mathcal{L}_{p^6} + S_{\text{WZW}}$;
2. nonleptonic weak interactions to first order in the Fermi coupling constant G_F: $\mathcal{L}^{\Delta S=1}_{G_F p^2}$, $\mathcal{L}^{\Delta S=1}_{G_8 p^4}, \mathcal{L}^{\Delta S=1}_{G_{27} p^4}$;
3. radiative corrections for strong processes: $\mathcal{L}^{\text{em}}_{e^2 p^0}, \mathcal{L}^{\text{em}}_{e^2 p^2}$;
4. radiative corrections for nonleptonic weak decays: $\mathcal{L}^{\text{emweak}}_{G_8 e^2 p^0}, \mathcal{L}^{\text{emweak}}_{G_8 e^2 p^2}$; and
5. radiative corrections for semileptonic weak decays: $\mathcal{L}^{\text{leptons}}_{e^2 p}$.

Beyond the leading order, unitarity and analyticity require the inclusion of loop contributions. In the purely strong sector, calculations have been performed up to next-to-next-to-leading order. **Figure 1** shows the corresponding skeleton diagrams of $O(p^6)$, with full lowest-order tree structures to be attached to propagators and vertices. The coupling constants of the various Lagrangians in **Table 1** absorb the divergences from loop diagrams leading to finite renormalized Green functions with scale-dependent couplings, the so-called low-energy constants (LECs). As in all EFTs, the LECs parametrize the effect of "heavy" degrees of freedom that are not represented explicitly in the EFT Lagrangian. Determination of those LECs is a major task for CHPT. In addition to phenomenological information, further theoretical input is needed. Lattice gauge theory has already furnished values for some LECs. To bridge the gap between the low-energy domain of CHPT and the perturbative domain of QCD, large-N_c motivated interpolations with meson resonance exchange have been used successfully to pin down some of the LECs.

Especially in cases where the knowledge of LECs is limited, renormalization group methods provide valuable information. As in renormalizable quantum field theories, the leading chiral logs $(\ln M^2/\mu^2)^L$ with a typical meson mass M, renormalization scale μ and loop order L can in principle be determined from one-loop diagrams only. In contrast to the renormalizable situation, new derivative structures (and quark mass insertions) occur at each loop order preventing a straightforward resummation of chiral logs.

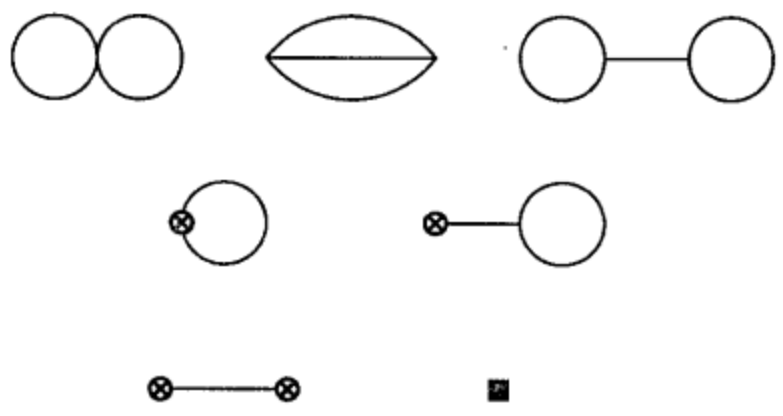


Figure 1 Skeleton diagrams of $O(p^6)$.. Normal vertices are from $\mathcal{L}_{p^2}$, crossed circles and the full square denote vertices from $\mathcal{L}_{p^4}$ and $\mathcal{L}_{p^6}$, respectively.

Among the many applications of CHPT in the meson sector are the determination of quark mass ratios and the analysis of pion–pion scattering where the chiral amplitude of next-to-next-to-leading order has been combined with dispersion theory (Roy equations). Of increasing importance for precision physics (CKM matrix elements, $(g-2)_\mu, \ldots$) are isospin-violating corrections including radiative corrections, where CHPT provides the only reliable approach in the low-energy region. Such corrections are also essential for the analysis of hadronic atoms like pionium, a $\pi^+\pi^-$ bound state.

CHPT has also been applied extensively in the single-baryon sector. There are several differences to the purely mesonic case. For instance, the chiral expansion proceeds more slowly and the nucleon mass m_N provides a new scale that does not vanish in the chiral limit. The formulation of heavy-baryon CHPT was modeled after HQET integrating out the nucleon modes of $O(m_N)$. To improve the convergence of the chiral expansion in some regions of phase space, a manifestly Lorentz-invariant formulation has been set up more recently (relativistic baryon CHPT). Many single-baryon processes have been calculated to fourth order in both approaches, for example, pion–nucleon scattering. With similar methods as in the mesonic sector, hadronic atoms like pionic or kaonic hydrogen have been investigated.

Nuclear Physics

In contrast to the meson and single-baryon sectors, amplitudes with two or more nucleons do not vanish in the chiral limit when the momenta of Goldstone mesons tend to zero. Consequently, the power counting is different in the many-nucleon sector. Multinucleon processes are treated with different EFTs depending on whether all momenta are smaller or larger than the pion mass.

In the very low energy regime $|\vec{p}| \ll M_\pi$, pions or other mesons do not appear as dynamical degrees of freedom. The resulting EFT is called "pionless EFT" and it describes systems like the deuteron, where the typical nucleon momenta are $\sim\sqrt{m_N B_d} \simeq 45\,\text{MeV}$ (B_d is the binding energy of the deuteron). The Lagrangian for the strong interactions between two nucleons has the form

$$\mathcal{L}_{NN} = C_0 \left(N^\top P_i N\right)^\dagger N^\top P_i N + \cdots \qquad [24]$$

where P_i are spin–isospin projectors and higher-order terms contain derivatives of the nucleon fields. The existence of bound states implies that at least part of the EFT Lagrangian must be treated nonperturbatively. Pionless EFT is an extension of effective-range theory that has long been used in nuclear physics. It has been applied successfully especially to the deuteron but also to more complicated few-nucleon systems like the Nd and $n\alpha$ systems. For instance, precise results for Nd scattering have been obtained with parameters fully determined from NN scattering. Pionless EFT has also been applied to the so-called halo nuclei, where a tight cluster of nucleons (like ^{4}He) is surrounded by one or more "halo" nucleons.

In the regime $|\vec{p}| > M_\pi$, the pion must be included as a dynamical degree of freedom. With some modifications in the power counting, the corresponding EFT is based on the approach of Weinberg (1990, 1991), who applied the usual rules of the meson and single-nucleon sectors to the nucleon–nucleon potential (instead of the scattering amplitude). The potential is then to be inserted into a Schrödinger equation to calculate physical observables. The systematic power counting leads to a natural hierarchy of nuclear forces, with only two-nucleon forces appearing up to next-to-leading order. Three- and four-nucleon forces arise at third and fourth order, respectively.

Significant progress has been achieved in the phenomenology of few-nucleon systems. The two- and n-nucleon ($3 \leq n \leq 6$) sectors have been pushed to fourth and third order, respectively, with encouraging signs of "convergence." Compton scattering off the deuteron, πd scattering, nuclear parity violation, solar fusion, and other processes have been investigated in the EFT approach. The quark mass dependence of the nucleon–nucleon interaction has also been studied.

See also: Anomalies; Electroweak Theory; High T_c Superconductor Theory; Noncommutative Geometry and the Standard Model; Operator Product Expansion in Quantum Field Theory; Perturbation Theory and its Techniques; Quantum Chromodynamics; Quantum Electrodynamics and its Precision Tests; Renormalization: General Theory; Seiberg–Witten Theory; Standard Model of Particle Physics; Symmetries and Conservation Laws; Symmetry Breaking in Field Theory.

Further Reading

Appelquist T and Carazzone J (1975) Infrared singularities and massive fields. *Physical Review D* 11: 2856.

Bedaque PF and van Kolck U (2002) Effective field theory for few nucleon systems. *Annual Review of Nuclear and Particle Science* 52: 339 (nucl-th/0203055).

Brambilla N, Pineda A, Soto J, and Vairo A (2004) Effective field theories for quarkonium. *Reviews of Modern Physics* (in press) (hep-ph/0410047).

Buchmüller W and Wyler D (1986) Effective Lagrangian analysis of new interactions and flavour conservation. *Nuclear Physics B* 268: 621.

Colangelo G, Gasser J, and Leutwyler H (2001) $\pi\pi$ scattering. *Nuclear Physics B* 603: 125 (hep-ph/0103088).

D'Hoker E and Weinberg S (1994) General effective actions. *Physical Review D* 50: 6050 (hep-ph/9409402).

Ecker G (1995) Chiral perturbation theory. *Progress in Particle and Nuclear Physics* 35: 1 (hep-ph/9501357).

Feruglio F (1993) The chiral approach to the electroweak interactions. *International Journal of Modern Physics A* 8: 4937 (hep-ph/9301281).

Gasser J and Leutwyler H (1984) Chiral perturbation theory to one loop. *Annals of Physics* 158: 142.

Gasser J and Leutwyler H (1985) Chiral perturbation theory: expansions in the mass of the strange quark. *Nuclear Physics B* 250: 465.

Hinchliffe I, Kersting N, and Ma YL (2004) Review of the phenomenology of noncommutative geometry. *International Journal of Modern Physics A* 19: 179 (hep/0205040).

Hoang AH (2002) Heavy quarkonium dynamics. In: Shifman M (ed.) *At the Frontier of Particle Physics/Handbook of QCD*, vol. 4. Singapore: World Scientific. (hep-ph/0204299).

Leutwyler H (1994) On the foundations of chiral perturbation theory. *Annals of Physics* 235: 165 (hep-ph/9311274).

Mannel T (2004) *Effective Field Theories in Flavour Physics*, Springer Tracts in Modern Physics, vol. 203. Berlin: Springer.

Manohar AV and Wise MB (2000) Heavy quark physics. *Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol.* 10: 1.

Meißner U (2005) Modern theory of nuclear forces. *Nuclear Physics A* 751: 149 (nucl-th/0409028).

Pich A (1999) Effective field theory. In: Gupta R *et al.* (eds.) *Proc. of Les Houches Summer School 1997: Probing the Standard Model of Particle Interactions*. Amsterdam: Elsevier. (hep-ph/9806303).

Scherer S (2002) Introduction to chiral perturbation theory. *Advances in Nuclear Physics* 27: 277 (hep-ph/0210398).

Weinberg S (1979) Phenomenological Lagrangians. *Physica A* 96: 327.

Weinberg S (1990) Nuclear forces from chiral Lagrangians. *Physics Letters B* 251: 288.

Weinberg S (1991) Effective chiral Lagrangians for nucleon–pion interactions and nuclear forces. *Nuclear Physics B* 363: 3.

Electric–Magnetic Duality

Tsou Sheung Tsun, University of Oxford, Oxford, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Classical electromagnetism is described by Maxwell's equations, which, in 3-vector notation and corresponding respectively to the laws of Coulomb, Ampère, Gauss, and Faraday, are given by eqns [1a]–[1d]:

$$\operatorname{div} \boldsymbol{E} = \rho \qquad [1a]$$

$$\operatorname{curl} \boldsymbol{B} - \frac{\partial \boldsymbol{E}}{\partial t} = \boldsymbol{J} \qquad [1b]$$

$$\operatorname{div} \boldsymbol{B} = 0 \qquad [1c]$$

$$\operatorname{curl} \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0 \qquad [1d]$$

Equivalently, in covariant 4-vector notation, these correspond to eqns [2a] and [2b]:

$$\partial_\nu F^{\mu\nu} = -j^\mu \qquad [2a]$$

$$\partial^*{}_\nu F^{\mu\nu} = 0 \qquad [2b]$$

In eqns [1], $\boldsymbol{E}$ and $\boldsymbol{B}$ are the electric and magnetic fields, respectively, ρ is the electric charge density, and $\boldsymbol{J}$ is the electric current. In eqns [2], $F_{\mu\nu}$ is the field tensor, ${}^*F_{\mu\nu}$ the dual field tensor, and j^μ is the 4-current, related to the previous vector quantities by the following relations:

$$F_{\mu\nu} = \begin{pmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & -B_3 & B_2 \\ -E_2 & B_3 & 0 & -B_1 \\ -E_3 & -B_2 & B_1 & 0 \end{pmatrix}$$

$${}^*F_{\mu\nu} = \begin{pmatrix} 0 & B_1 & B_2 & B_3 \\ -B_1 & 0 & E_3 & -E_2 \\ -B_2 & -E_3 & 0 & E_1 \\ -B_3 & E_2 & -E_1 & 0 \end{pmatrix}$$

$$j^\mu = (\rho, \boldsymbol{J})$$

Throughout this article, we shall denote the three spatial indices by lower-case Latin letters such as i, j, while Greek indices such as μ, ν denote spacetime indices running through $0, 1, 2, 3$. The Einstein summation convention is used, whereby repeated indices are summed. Spacetime indices are raised and lowered by the (flat) Minkowski metric $g_{\mu\nu} = \operatorname{diag}(1, -1, -1, -1)$. We also use units conventional in particle physics, in which the reduced Planck constant $\hbar$ and the speed of light c are both set to 1.

In terms of the totally skew symmetric symbol $\varepsilon_{\mu\nu\rho\sigma}$ (with $\varepsilon_{0123} = 1$), the two field tensors are related by eqn [3]:

$${}^*F_{\mu\nu} = -\tfrac{1}{2}\varepsilon_{\mu\nu\rho\sigma}F^{\rho\sigma} \qquad [3]$$

We say that ${}^*F_{\mu\nu}$ is the dual of $F_{\mu\nu}$, and eqn [3] is indeed a duality relation because eqn [4] holds, which means that up to a sign, $F_{\mu\nu}$ and ${}^*F_{\mu\nu}$ are duals of each other:

$${}^*({}^*F) = -F \qquad [4]$$

This duality is in fact the Hodge duality between p-forms and $(n-p)$-forms in an n-dimensional space. In our particular case, $p = 2$ and $n = 4$, so that both F and its dual are 2-forms. The minus sign in eqn [4] comes about because of the Lorentzian (or pseudo-Riemannian) signature of Minkowski spacetime.

The physical significance of this duality is that such a symmetry interchanges electric and magnetic fields (again up to sign) (eqn [5]), as can be seen from the matrix representation of $F_{\mu\nu}$ and ${}^*F_{\mu\nu}$ above:

$${}^*\colon \boldsymbol{E} \mapsto \boldsymbol{B},\ \boldsymbol{B} \mapsto -\boldsymbol{E} \qquad [5]$$

Now in the absence of electric charges and currents, one sees immediately that Maxwell's equations [1] or [2] are dual symmetric. This means that, *in vacuo*, whether we call an electromagnetic field electric or magnetic is a matter of convention. As far as the dynamics is concerned, there is no distinction.

On the other hand, eqns [1] and [2] as presented, that is, in the presence of matter, are manifestly not dual symmetric. The underlying reason for this asymmetry has been much studied both in physics and in mathematics. One of the two questions that this article addresses is precisely this. Following on this, we shall see what happens if we try somehow to restore this dual symmetry even in the presence of matter.

The second question that we wish to discuss is a generalization of this duality. Electromagnetism is a gauge theory, in which the gauge group is the abelian circle group U(1), representing the phase of wave-functions in quantum mechanics. A physically relevant generalization, in which the abelian U(1) is replaced by a nonabelian group (e.g., SU(2), SU(3))

is called Yang–Mills theory (Yang and Mills 1954), which is the theoretical basis of all modern particle physics. We shall show in this article how the concept of electric–magnetic duality can be generalized in the context of Yang–Mills theory.

Gauge Invariance, Sources, and Monopoles

Electric–magnetic duality, whether in the well-known abelian case or in the still somewhat open nonabelian case, is intimately connected with gauge invariance, sources, and monopoles, and also the dynamics as embodied in the gauge action. These questions in turn find their natural setting in differential geometry, particularly the geometry of fibre bundles.

Although classical electrodynamics can be fully described by the field tensor $F_{\mu\nu}$, one needs to introduce the electromagnetic (or gauge) potential A_μ if one considers quantum mechanics, as has been beautifully demonstrated by the Bohm–Aharonov experiment. The two quantities are related by eqn [6]:

$$F_{\mu\nu}(x) = \partial_\nu A_\mu(x) - \partial_\mu A_\nu(x) \qquad [6]$$

The fact that the phase of a wave function $\psi(x)$ (e.g., of the electron) is not a measurable quantity (although relative phases of course are) implies that we are free to make the following transformation:

$$\psi(x) \mapsto e^{ie\Lambda(x)}\psi(x) \qquad [7]$$

This in turn implies an unobservable transformation [8] on the gauge potential, where $\Lambda(x)$ is a real-valued function on spacetime:

$$A_\mu(x) \mapsto A_\mu(x) + \partial_\mu\Lambda(x) \qquad [8]$$

This invariance is called gauge invariance. Since in this abelian case $F_{\mu\nu}$ is gauge invariant, so are the Maxwell equations, for which we shall take from now on the covariant form [2]. Inasmuch as the Maxwell equations dictate the dynamics of electromagnetism, gauge invariance is an intrinsic ingredient even in the classical theory.

In Yang–Mills theory, the U(1) phase $e^{ie\Lambda(x)}$ is replaced by an element $S(x)$ of a nonabelian group G, so that eqns [7], [8], and [6] become, respectively, eqns [9], [10], and [11]:

$$\psi(x) \mapsto S(x)\psi(x) \qquad [9]$$

$$A_\mu(x) \mapsto S(x)A_\mu(x)S^{-1}(x) - \left(\frac{i}{g}\right)\partial_\mu S(x)S^{-1}(x) \qquad [10]$$

$$F_{\mu\nu}(x) = \partial_\nu A_\mu(x) - \partial_\mu A_\nu(x) + ig[A_\mu(x), A_\nu(x)] \qquad [11]$$

Here the electric coupling e is replaced by a general gauge coupling g. The quantities A_μ and $F_{\mu\nu}$ now take values in the Lie algebra of the Lie group G and the bracket is the Lie bracket. The wave function $\psi(x)$ takes values in a vector space on which an appropriate representation of G acts. Notice that now the field tensor $F_{\mu\nu}$ is no longer invariant, but only covariant:

$$F_{\mu\nu}(x) \mapsto S(x)F_{\mu\nu}(x)S^{-1}(x) \qquad [12]$$

Next we consider the charges of gauge theory. For the moment, we wish to distinguish between two types of charges: sources and monopoles. These are defined with respect to the gauge field, which in turn is derivable from the gauge potential.

Source charges are those charges that give rise to a nonvanishing divergence of the field. For example, the electric current j due to the presence of the electric charge e occurs on the right-hand side of the first Maxwell equation, and is given in the quantum case by eqn [13], where γ^μ is a Dirac gamma matrix, identifiable as a basis element of the Clifford algebra over spacetime:

$$j^\mu = e\bar{\psi}\gamma^\mu\psi \qquad [13]$$

In the Yang–Mills case, the first Maxwell equation is replaced by the Yang–Mills equation

$$D_\nu F^{\mu\nu} = -j^\mu, \qquad j^\mu = g\bar{\psi}\gamma^\mu\psi \qquad [14]$$

We define the covariant derivative D as in

$$D_\mu F^{\mu\nu} = \partial_\mu F^{\mu\nu} - ig[A_\mu, F^{\mu\nu}] \qquad [15]$$

Monopole charges, on the other hand, are topological obstructions specified geometrically by nontrivial G-bundles over every 2-sphere S^2 surrounding the charge. They are classified by elements of $\pi_1(G)$, the fundamental group of G. They are typified by the (abelian) magnetic monopole as first discussed by Dirac in 1931.

Let us go into a little more detail about the Dirac magnetic monopole. If the field tensor $F_{\mu\nu}$ does come from a gauge potential A_μ as in eqn [6], then simple algebra will tell us that this implies $\partial_\nu {}^*F^{\mu\nu} = 0$ as in eqn [2]. Hence, we conclude the following:

$$\exists \text{ monopole} \implies A_\mu \text{ cannot be well defined everywhere}$$

The result is actually stronger. Suppose there exists a magnetic monopole at a certain point in spacetime, and, without loss of generality, we shall consider a static monopole. If we surround this point by a (spatial) 2-sphere Σ, then the magnetic flux out of the sphere is given by

$$\iint_\Sigma \boldsymbol{B}\cdot d\boldsymbol{\sigma} = \iint_{\Sigma^N} \boldsymbol{B}\cdot d\boldsymbol{\sigma} + \iint_{\Sigma^S} \boldsymbol{B}\cdot d\boldsymbol{\sigma} \qquad [16]$$

Here Σ^N and Σ^S are the northern and southern hemispheres overlapping on the equator S. By Stokes' theorem, since $F_{\mu\nu}$ has no components $F_{0i} = E_i$, we have

$$\iint_{\Sigma^N} \boldsymbol{B} \cdot d\boldsymbol{\sigma} = \oint_S \boldsymbol{A} \cdot d\boldsymbol{s} \qquad [17a]$$

$$\iint_{\Sigma^S} \boldsymbol{B} \cdot d\boldsymbol{\sigma} = \oint_{-S} \boldsymbol{A} \cdot d\boldsymbol{s} \qquad [17b]$$

In eqn [17b], $-S$ means the equator with the opposite orientation. Hence, $\oint_S + \oint_{-S} = 0$. But this contradicts the assumption that there exists a magnetic monopole at the center of the sphere. Hence, we see that if a monopole exists, then A_μ will have at least a string of singularities leading out of it. This is the famous Dirac string.

The more mathematically elegant way to describe this is that the principal bundle corresponding to electromagnetism with a magnetic monopole is nontrivial, so that the gauge potential A_μ has to be patched (i.e., related by transition functions in the overlap). Consider the example of a static monopole of magnetic charge $\tilde{e}$. For any (spatial) sphere S_r of radius r surrounding the monopole, we cover it with two patches N, S as follows:

$$\text{(N):} \quad 0 \le \theta < \pi, \ 0 \le \phi \le 2\pi$$
$$\text{(S):} \quad 0 < \theta \le \pi, \ 0 \le \phi \le 2\pi$$

In each patch we define the following:

$$A_1^{(N)} = \frac{\tilde{e}y}{4\pi r(r+z)}$$
$$A_2^{(N)} = -\frac{\tilde{e}x}{4\pi r(r+z)}$$
$$A_3^{(N)} = 0$$
$$A_1^{(S)} = -\frac{\tilde{e}y}{4\pi r(r-z)}$$
$$A_2^{(S)} = \frac{\tilde{e}x}{4\pi r(r-z)}$$
$$A_3^{(S)} = 0$$

In the overlap (containing the equator), $A^{(N)}$ and $A^{(S)}$ are related by a gauge transformation:

$$A_i^{(N)} - A_i^{(S)} = \partial_i \Lambda$$
$$\Lambda = \left(\frac{\tilde{e}}{2\pi}\right) \tan^{-1}\left(\frac{y}{x}\right) = \frac{\tilde{e}\phi}{2\pi} \qquad [18]$$

Notice that $A_i^{(N)}$ has a line of singularity along the negative z-axis (which is the Dirac string in this case); similarly for $A_i^{(S)}$ along the positive z-axis.

Furthermore, the corresponding field strength is given by

$$\boldsymbol{E} = 0 \qquad [19a]$$

$$\boldsymbol{B} = \frac{\tilde{e}\boldsymbol{r}}{4\pi r^3} \qquad [19b]$$

If we now evaluate the "magnetic flux" out of S_r, we have

$$\iint_{S_r} \boldsymbol{B} \cdot d\boldsymbol{\sigma} = \oint_{\text{Equator}} \left(A_\mu^{(N)} - A_\mu^{(S)}\right) dx^\mu = \tilde{e} \qquad [20]$$

In other words, in the presence of a magnetic monopole, the second half of Maxwell's equations is modified according to eqn [21], with $\tilde{j}^\mu$ given by eqn [22].

$$\left.\begin{aligned} \operatorname{div} \boldsymbol{B} &= \tilde{\rho} \\ \operatorname{curl} \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} &= \tilde{\boldsymbol{J}} \end{aligned}\right\}, \qquad \partial_\nu {}^*F^{\mu\nu} = -\tilde{j}^\mu \qquad [21]$$

$$\tilde{j}^\mu = \tilde{e}\bar{\psi}\gamma^\mu\psi \qquad [22]$$

Furthermore, the form of eqn [21] tells us that a monopole of the $F_{\mu\nu}$ field can also be considered as a source of the ${}^*F_{\mu\nu}$ field. The two descriptions are equivalent.

How are the charges e and $\tilde{e}$ related? The gauge transformation $S = e^{ie\Lambda}$ relating $A_\mu^{(N)}$ and $A_\mu^{(S)}$ must be well defined; that is, if one goes round the equator once, $\phi = 0 \to 2\pi$, one should get the same S. This gives

$$e\tilde{e} = 2\pi n, \qquad n \in \mathbb{Z} \qquad [23]$$

In particular, the unit electric and magnetic charges are related by eqn [24], which is Dirac's quantization condition,

$$e\tilde{e} = 2\pi \qquad [24]$$

So, in principle, just as in the electric case, where we could have charges $e, 2e, \ldots$, here we could also have magnetic charges of $\tilde{e}, 2\tilde{e}, \ldots$. In other words, both charges are quantized.

Another way to look at this is to consider the classification of principal bundles over S^2. The reason for these topological 2-spheres is that we are interested in enclosing a point charge. For a nontrivial bundle, the patching is given by a function S defined in the overlap (the equator), in other words, a map $S^1 \to U(1)$. What this amounts to is a closed curve in the circle group U(1). Now, curves that can be continuously deformed into one another cannot give

distinct fibre bundles, so that one sees easily that there exists a one-to-one correspondence:

$$\{\text{principal U(1) bundles over } S^2\}$$
$$\updownarrow$$
$$\{\text{homotopy classes of closed curves in U(1)}\}$$

This last is $\pi_1(\mathrm{U}(1)) \cong \mathbb{Z}$. Hence, we recover Dirac's quantization condition.

So, for electromagnetism, there are two equivalent ways of defining the magnetic charge, as a source or as a monopole:

1. $\partial_\nu {}^*F^{\mu\nu} = -\tilde{j}^\mu \propto n\tilde{e} \neq 0$.
2. An element of $\pi_1(\mathrm{U}(1)) \cong \mathbb{Z}$.

The same goes for the electric charge. We also note that both definitions give us the fact that these charges are discrete (quantized) and conserved (invariant under continuous deformations).

We now want to apply similar considerations to the magnetic charges in the nonabelian case. For several (subtle) reasons the obvious expression $D_\nu {}^*F^{\mu\nu} \stackrel{?}{=} -\tilde{j}^\mu$ as a source (see **Table 1**) does not work. The quickest way to say this is that ${}^*F^{\mu\nu}$ in general has no corresponding potential $\tilde{A}_\mu$ and so is not a gauge field. Moreover, in contrast to the abelian case, the field tensor does not fully specify the physical field configuration, as demonstrated by Wu and Yang. We shall come back to this later.

But we have just seen that in the abelian case there is another equivalent definition, which is that a magnetic monopole is given by the gauge configuration corresponding to a nontrivial U(1) bundle over S^2. This can be generalized to the nonabelian case without any problem. Moreover, this definition automatically guarantees that a nonabelian monopole charge is quantized and conserved. This is the way monopoles are defined above.

Arguments similar to the abelian case easily yield the nonabelian analog of the Dirac quantization condition, eqn [25], the difference between the two cases being only a matter of conventional normalization.

$$g\tilde{g} = 4\pi \qquad [25]$$

Table 1 Definitions of charges

	Sources	*Monopoles*
Abelian	$\partial_\nu F^{\mu\nu} = -j^\mu$	$\partial_\nu {}^*F^{\mu\nu} = -\tilde{j}^\mu$
Nonabelian	$D_\nu F^{\mu\nu} = -j^\mu$	?

Abelian Duality and the Wu–Yang Criterion

We saw above the well-known fact that classical Maxwell theory is invariant under the duality operator. By this we mean that at any point in spacetime free of electric and magnetic charges we have the two dual symmetric Maxwell equations:

$$\partial_\nu {}^*F^{\mu\nu} = 0 \quad [\mathrm{d}F = 0] \qquad [26]$$

$$\partial_\nu F^{\mu\nu} = 0 \quad [\mathrm{d}{}^*F = 0] \qquad [27]$$

Displayed in square brackets are the equivalent equations in the language of differential forms. Then by the Poincaré lemma we deduce immediately the existence of potentials A and $\tilde{A}$ such that eqns [28] and [29] hold:

$$F_{\mu\nu}(x) = \partial_\nu A_\mu(x) - \partial_\mu A_\nu(x) \qquad [F = \mathrm{d}A] \qquad [28]$$

$${}^*F_{\mu\nu}(x) = \partial_\nu \tilde{A}_\mu(x) - \partial_\mu \tilde{A}_\nu(x) \qquad [{}^*F = \mathrm{d}\tilde{A}] \qquad [29]$$

The two potentials transform independently under independent gauge transformations Λ and $\tilde{\Lambda}$:

$$A_\mu(x) \mapsto A_\mu(x) + \partial_\mu \Lambda(x) \qquad [30]$$

$$\tilde{A}_\mu(x) \mapsto \tilde{A}_\mu(x) + \partial_\mu \tilde{\Lambda}(x) \qquad [31]$$

This means that the full symmetry of this theory is doubled to $\mathrm{U}(1) \times \tilde{\mathrm{U}}(1)$, where the tilde on the second circle group indicates that it is the symmetry of the dual potential $\tilde{A}$. It is important to note that the physical degrees of freedom remain the same. This is clear because F and ${}^*F$ are related by an algebraic equation [3]. As a consequence, the physical theory is the same: the doubled gauge symmetry is there all the time but is just not so readily detected.

As mentioned in the Introduction, this dual symmetry means that what we call "electric" or "magnetic" is entirely a matter of choice.

In the presence of electric charges, the Maxwell equations usually appear as

$$\partial_\nu {}^*F^{\mu\nu} = 0 \qquad [32]$$

$$\partial_\nu F^{\mu\nu} = -j^\mu \qquad [33]$$

The apparent asymmetry in these equations comes from the experimental fact that there is only one type of charges observed in nature, which we choose to regard as a source of the field F (or, equivalently but unconventionally, as a monopole of the field ${}^*F$). But as we see by dualizing eqns [32] and [33], that is, by interchanging the role of electricity and magnetism in relation to F, we could equally have thought of these instead as source charges of

the field ${}^*F$ (or, similarly to the above, as monopoles of F):

$$\partial_\nu {}^*F^{\mu\nu} = -\tilde{j}^\mu \qquad [34]$$

$$\partial_\nu F^{\mu\nu} = 0 \qquad [35]$$

If both electric and magnetic charges existed in nature, then we would have the dual symmetric pair:

$$\partial_\nu {}^*F^{\mu\nu} = -\tilde{j}^\mu \qquad [36]$$

$$\partial_\nu F^{\mu\nu} = -j^\mu \qquad [37]$$

This duality in fact goes much deeper, as can be seen if we use the Wu–Yang criterion to derive the Maxwell equations, although we should note that what we present here is not the textbook derivation of the Maxwell equations from an action, but we conisder this method to be much more intrinsic and geometric. Consider first pure electromagnetism. The free Maxwell action is given by

$$\mathcal{A}_F^0 = -\frac{1}{4}\int F_{\mu\nu}F^{\mu\nu} \qquad [38]$$

The true variables of the (quantum) theory are the A_μ, so in eqn [38] we should put in a constraint to say that $F_{\mu\nu}$ is the curl of A_μ [28]. This can be viewed as a topological constraint, because it is precisely equivalent to [26]. Using the method of Lagrange multipliers, we form the constrained action

$$\mathcal{A} = \mathcal{A}_F^0 + \int \lambda_\mu(\partial_\nu {}^*F^{\mu\nu}) \qquad [39]$$

We can now vary this with respect to $F_{\mu\nu}$, obtaining eqn [40], which implies [27]:

$$F^{\mu\nu} = 2\varepsilon^{\mu\nu\rho\sigma}\partial_\rho\lambda_\sigma \qquad [40]$$

Moreover, the Lagrange multiplier λ is exactly the dual potential $\tilde{A}$.

This derivation is entirely dual symmetric, since we can equally well use [27] as constraint for the action $\mathcal{A}_F^0$, now considered as a functional of ${}^*F^{\mu\nu}$ (eqn [41]), and obtain [26] as the equation of motion:

$$\mathcal{A}_F^0 = \frac{1}{4}\int {}^*F_{\mu\nu}{}^*F^{\mu\nu} \qquad [41]$$

This method applies to the interaction of charges and fields as well. In this case we start with the free field plus free particle action (eqn [42]), where we assume the free particle m to satisfy the Dirac equation,

$$\mathcal{A}^0 = \mathcal{A}_F^0 + \int \bar{\psi}(\mathrm{i}\partial_\mu\gamma^\mu - m)\psi \qquad [42]$$

To fix ideas, let us regard this particle carrying an electric charge e as a monopole of the potential $\tilde{A}_\mu$. Then the constraint we put in is [33], giving

$$\mathcal{A}' = \mathcal{A}^0 + \int \tilde{\lambda}_\mu(\partial_\nu F^{\mu\nu} + j^\mu) \qquad [43]$$

Variation with respect to ${}^*F$ gives eqn [32], and varying with respect to $\bar{\psi}$ gives

$$(\mathrm{i}\partial_\mu\gamma^\mu - m)\psi = -eA_\mu\gamma^\mu\psi \qquad [44]$$

So, the complete set of equations for a Dirac particle carrying an electric charge e in an electromagnetic field is [32], [33], and [44]. The duals of these equations will describe the dynamics of a Dirac magnetic monopole in an electromagnetic field.

We see from this that the Wu–Yang criterion actually gives us an intuitively clear picture of interactions. The assertion that there is a monopole at a certain spacetime point x means that the gauge field on a 2-sphere surrounding x has to have a certain topological configuration (e.g., giving a nontrivial bundle of a particular class), and if the monopole moves to another point then the gauge field will have to rearrange itself so as to maintain the same topological configuration around the new point. There is thus naturally a coupling between the gauge field and the position of the monopole, or, in physical language, a topologically induced interaction between the field and the charge (Wu and Yang, 1976). Furthermore, this treatment of interaction between field and matter is entirely dual symmetric.

As a side remark, consider that although the action $\mathcal{A}_F^0$ is not immediately identifiable as geometric in nature, the Wu–Yang criterion, by putting the topological constraint and the equation of motion on equal (or dual) footing, suggests that in fact it is geometric in a subtle manner not yet fully understood. Moreover, as pointed out, eqn [40] says that the dual potential is given by the Lagrange multiplier of the constrained action.

Nonabelian Duality Using Loop Variables

The next natural step is to generalize this duality to the nonabelian Yang–Mills case. Although there is no difficulty in defining ${}^*F^{\mu\nu}$, which is again given by [3], we immediately come to difficulties in the relation between field and potential; for example, as in eqn [11],

$$F_{\mu\nu}(x) = \partial_\nu A_\mu(x) - \partial_\mu A_\nu(x) + \mathrm{i}g[A_\mu(x), A_\nu(x)]$$

First of all, despite appearances the Yang–Mills equation [45] (in the free-field case) and the Bianchi

identity [46] are not dual-symmetric, because the correct dual of the Yang–Mills equation ought to be given by eqn [47], where $\tilde{D}_\nu$ is the covariant derivative corresponding to a dual potential:

$$D_\nu F^{\mu\nu} = 0 \qquad [45]$$

$$D_\nu {}^*F^{\mu\nu} = 0 \qquad [46]$$

$$\tilde{D}_\nu {}^*F^{\mu\nu} = 0 \qquad [47]$$

Secondly, the Yang–Mills equation, unlike its abelian counterpart [27], says nothing about whether the 2-form $^*F$ is closed or not. Nor is the relation [11] about exactness at all. In other words, the Yang–Mills equation does not guarantee the existence of a dual potential, in contrast to the Maxwell case. In fact, Gu and Yang have constructed a counterexample. Because the true variables of a gauge theory are the potentials and not the fields, this means that Yang–Mills theory is not symmetric under the Hodge star operation [3].

Nevertheless, electric–magnetic duality is a very useful physical concept, so one may wish to seek a more general duality transform ($\tilde{}$), satisfying the following properties:

1. $(\ )^{\tilde{}\tilde{}} = \pm(\ )$.
2. Electric field $F_{\mu\nu} \overset{\sim}{\longleftrightarrow}$ magnetic field $\tilde{F}_{\mu\nu}$.
3. Both A_μ and $\tilde{A}_\mu$ exist as potentials (away from charges).
4. Magnetic charges are monopoles of A_μ, and electric charges are monopoles of $\tilde{A}_\mu$.
5. $\tilde{}$ reduces to * in the abelian case.

One way to do this is to study the Wu–Yang criterion more closely. This reveals the concept of charges as topological constraints to be crucial even in the pure field case, as can be seen in **Figure 1**. The point to stress is that, in the above abelian case, the condition for the absence of a topological charge (a monopole) exactly removes the redundancy of the variables $F_{\mu\nu}$, and hence recovers the potential A_μ.

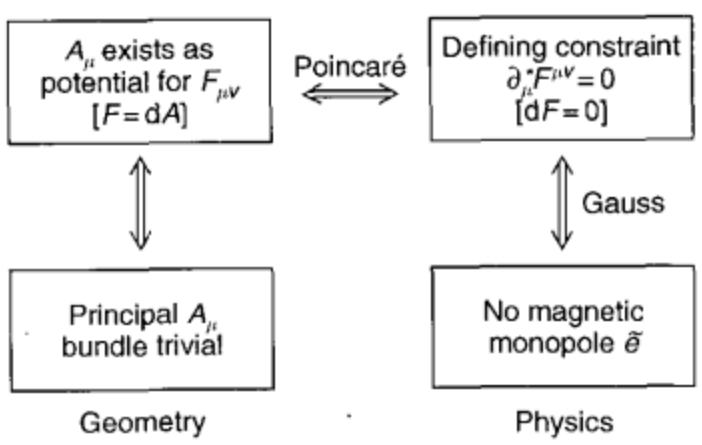


Figure 1

Now the nonabelian monopole charge was defined topologically as an element of $\pi_1(G)$, and this definition also holds in the abelian case of U(1), with $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$. So the first task is to write down a condition for the absence of a nonabelian monopole.

To fix ideas, let us consider the group SO(3), whose monopole charges are elements of $\mathbb{Z}_2$, which can be denoted by a sign $\pm$. The vacuum, charge (+) (that is, no monopole) is represented by a closed curve in the group manifold of even winding number, and the monopole charge (−) by a closed curve of odd winding number. It is more convenient, however, to work in SU(2), which is the double cover of SO(3) and which has the topology of S^3, as sometimes it is useful to identify the fundamental group of SO(3) with the center of SU(2) and hence consider the monopole charge as an element of this center. There the charge (+) is represented by a closed curve, and the charge (−) by a curve that winds an odd number of "half-times" round the sphere S^3. Since these charges are defined by closed curves, it is reasonable to try to write the constraint in terms of loop variables. The treatment presented below is not as rigorous as some others, but the latter are not so well adapted to the problem in hand. Furthermore, it is important to emphasize that this approach aims to generalize electric–magnetic duality to Yang–Mills theory in direct and close analogy to duality in electromagnetism, without any further symmetries with which it may be expedient to enrich the theory. Other approaches are referred to in the next section.

Consider the gauge-invariant Dirac phase factor (or holonomy) $\Phi(C)$ of a loop C, which can be written symbollically as a path-ordered exponential:

$$\Phi[\xi] = P_s \exp ig \int_0^{2\pi} ds\, A_\mu(\xi(s))\dot{\xi}^\mu(s) \qquad [48]$$

In eqn [48], we parametrize the loop C as is eqn [49] and a dot denotes differentiation with respect to the parameter s.

$$C: \quad \{\xi^\mu(s):\ s = 0 \to 2\pi, \xi(0) = \xi(2\pi) = \xi_0\} \qquad [49]$$

We thus regard loop variables in general as functionals of continuous piecewise smooth functions ξ of s. In this way, loop derivatives and loop integrals are just functional derivatives and functional integrals. This means that loop derivatives $\delta_\mu(s)$ are defined by a regularization procedure approximating delta functions with finite bump functions and then taking limits in a definite order. For functional integrals, there exist various regularization procedures, which are treated elsewhere in this Encyclopedia.

Polyakov (1980) introduces the logarithmic loop derivative of $\Phi[\xi]$:

$$F_\mu[\xi|s] = \frac{i}{g}\Phi^{-1}[\xi]\delta_\mu(s)\Phi[\xi] \qquad [50]$$

This acts as a kind of "connection" in loop space since it tells us how the phase of $\Phi[\xi]$ changes from one loop to a neighbouring loop. One can go a step further and define its "curvature" in direct analogy with $F_{\mu\nu}(x)$ by

$$G_{\mu\nu}[\xi|s] = \delta_\nu(s)F_\mu[\xi|s] - \delta_\mu(s)F_\nu[\xi|s] + ig[F_\mu[\xi|s], F_\nu[\xi|s]] \qquad [51]$$

It can be shown that by using the $F_\mu[\xi|s]$ we can rewrite the Yang–Mills action as eqn [52], where the normalization factor $\bar{N}$ is an infinite constant:

$$\mathcal{A}_F^0 = -\frac{1}{4\pi\bar{N}}\int \delta\xi \int_0^{2\pi} ds\, \mathrm{tr}\{F_\mu[\xi|s]F^\mu[\xi|s]\}|\dot{\xi}(s)|^{-2} \qquad [52]$$

However, the true variables of the theory are still the A_μ. They represent 4 functions of a real variable, whereas the loop connections represent 4 functionals of the real function $\xi(s)$. Just as in the case of the $F_{\mu\nu}$, these $F_\mu[\xi|s]$ have to be constrained so as to recover A_μ, but this time much more severely.

It turns out that, in pure Yang–Mills theory, the constraint that says there are no monopoles ([53]) also removes the redundancy of the loop variables, exactly as in the abelian case,

$$G_{\mu\nu}[\xi|s] = 0 \qquad [53]$$

That this condition is necessary is easy to see by simple algebra. The proof of the converse of this "extended Poincaré lemma" is fairly lengthy. Granted this, we can now apply the Wu–Yang criterion to the action [52] and derive the Polyakov equation [54], which is the loop version of the Yang–Mills equation:

$$\delta_\mu(s)F^\mu[\xi|s] = 0 \qquad [54]$$

In the presence of a monopole charge $(-)$, the constraint [53] will have a nonzero right-hand side,

$$G_{\mu\nu}[\xi|s] = -J_{\mu\nu}[\xi|s] \qquad [55]$$

The loop current $J_{\mu\nu}[\xi|s]$ can be written down explicitly. However, its global form is much easier to understand. Recall that $F^\mu[\xi|s]$ can be thought of as a loop connection, for which we can form its "holonomy." This is defined for a closed (spatial) surface Σ (enclosing the monopole), parametrized by a family of closed curves $\xi_t(s), t = 0 \to 2\pi$. The "holonomy" Θ_Σ is then the total change in phase of $\Phi[\xi_t]$ as $t \to 2\pi$, and thus equals the charge $(-)$.

To formulate an electric–magnetic duality that is applicable to nonabelian theory, one defines yet another set of loop variables. Instead of the Dirac phase factor $\Phi[\xi]$ for a complete curve [48], we consider the parallel phase transport for part of a curve from s_1 to s_2:

$$\Phi_\xi(s_2, s_1) = P_s \exp ig\int_{s_1}^{s_2} ds A_\mu(\xi(s))\dot{\xi}^\mu(s) \qquad [56]$$

Then the new variables are defined by [57].

$$E_\mu[\xi|s] = \Phi_\xi(s,0)F_\mu[\xi|s]\Phi_\xi^{-1}(s,0) \qquad [57]$$

These are not gauge invariant like $F_\mu[\xi|s]$ and may not be as useful in general, but seem more convenient for dealing with duality.

Using these variables, we now define their dual $\tilde{E}_\mu[\eta|t]$ according to

$$\omega^{-1}(\eta(t))\tilde{E}_\mu[\eta|t]\omega(\eta(t)) = -\frac{2}{\bar{N}}\varepsilon_{\mu\nu\rho\sigma}\dot{\eta}^\nu(t)\int \delta\xi\, ds E^\rho[\xi|s]\dot{\xi}^\sigma(s)\dot{\xi}^{-2}(s) \times \delta(\xi(s) - \eta(t)) \qquad [58]$$

In eqn [58], $\omega(x)$ is a (local) rotation matrix transforming from the frame in which the orientation in internal symmetry space of the fields $E_\mu[\xi|s]$ are measured to the frame in which the dual fields $\tilde{E}_\nu[\eta|t]$ are measured. It can be shown that this dual transform satisfies all five of the required conditions listed earlier.

Electric–magnetic duality in Yang–Mills theory is now fully reestablished using this generalized duality. We have the dual pairs of equations [59]–[60] and [61]–[62]:

$$\delta_\nu E_\mu - \delta_\mu E_\nu = 0 \qquad [59]$$

$$\delta^\mu E_\mu = 0 \qquad [60]$$

$$\delta^\mu \tilde{E}_\mu = 0 \qquad [61]$$

$$\delta_\nu \tilde{E}_\mu - \delta_\mu \tilde{E}_\nu = 0 \qquad [62]$$

Equation [59] guarantees that the potential A exists, and so is equivalent to [53], and hence is the nonabelian analog of [26]; while equation [60] is equivalent to the Polyakov version of Yang–Mills equation [54], and hence is the nonabelian analog of [27]. Equation [61] is equivalent by duality to [59] and is the dual Yang–Mills equation. Similarly equation [62] is equivalent to [60], and guarantees the existence of the dual potential $\tilde{A}$.

The treatment of charges using the Wu–Yang criterion also follows the abelian case, and will not be further elaborated here. For this and further details, the reader is referred to the orginal papers (Chan and Tsou 1993, 1999).

Also, just as in the abelian case, the gauge symmetry is doubled: from the group G we deduce that the full gauge symmetry is in fact $G \times \tilde{G}$, but that the physical degrees of freedom remain the same.

The above exposition establishes electric–magnetic duality in Yang–Mills theory only for classical fields. A hint that this duality persists at the quantum level comes from the work of 't Hooft (1978) on confinement. There he introduces two loop quantities $A(C)$ and $B(C)$ that are operators in the Hilbert space of quantum states satisfying the commutation relation [63] for an SU(N) gauge theory, where n is the linking number between the two (spatial) loops C and C':

$$A(C)B(C') = B(C')A(C)\exp(2\pi i n/N) \qquad [63]$$

The order or Wilson operator is given explicitly by $A(C) = \mathrm{tr}\,\Phi(C)$. These two operators play dual roles in the sense of electric–magnetic duality:

- $A(C)$ measures the magnetic flux through C and creates electric flux along C.
- $B(C)$ measures the electric flux through C and creates magnetic flux along C.

By defining the disorder operator $B(C)$ as the Wilson operator corresponding to the dual potential $\tilde{A}$ obtained above, one can prove the commutation relation [63], thus showing that these classical fields, when promoted to operators, retain their duality relation. Furthermore, there is a remarkable relation between the two (abstractly identical) gauge groups, in that if one is confined then the dual must be broken (that is, in the Higgs phase). This result is known as 't Hooft's theorem.

The doubling of gauge symmetry, together with 't Hooft's theorem, has been applied to the confined colour group SU(3) of quantum chromodynamics (QCD), in the Dualized Standard Model, to solve the puzzle of the existence of exactly three generations of fermions, with good observational support, by identifying the (necessarily broken) dual SU(3) with the generation symmetry (Chan and Tsou, 2002).

Other Treatments of Nonabelian Duality

Since Yang–Mills theory is not symmetric under the Hodge *-operation, there are several routes one can take to generalize the concept of electric–magnetic duality to the nonabelian case. What was presented in the last section is a modification of the *-operation so as to restore this symmetry for Yang–Mills theory, keeping to the original gauge structure as much as possible. However, Yang–Mills theory as used today in particle and field theories are usually embedded in theories with more structures. In the simplest case we have the Standard Model of Particle Physics, which describes all of particle interactions (except gravity) and which has the gauge group usually written as $\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)$, corresponding to the SU(3) of strong interaction and $\mathrm{SU}(2) \times \mathrm{U}(1)$ of electroweak interaction. [Strictly speaking, it is $(\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1))/\mathbb{Z}_6$, if we have the standard particle spectrum.] However, the former group is confined and the latter broken. The breaking is usually effected by introducing scalar fields called Higgs fields into the theory.

Besides the experimentally well-tested Standard Model, there are many theoretically popular models of gauge theory in which supersymmetry is postulated, thereby introducing extra symmetries into the theory. Many of these are remnants of string theory, and are usually envisaged as gauge theories in a spacetime dimension higher than 4.

Because of the extra structures and increased symmetries in these theories, there is quite a proliferation of concepts of duality, which could all be thought of as generalizations of abelian electric–magnetic duality (Schwarz, 1997). They come under the names of Seiberg–Witten duality, S-duality, T-duality, mirror symmetry, and so on. All these other aspects of duality have their own entries in this Encyclopedia.

See also: AdS/CFT Correspondence; Duality in Topological Quantum Field Theory; Four-Manifold Invariants and Physics; Large-N Dualities; Measure on Loop Spaces; Mirror Symmetry: a Geometric Survey; Nonperturbative and Topological Aspects of Gauge Theory; Seiberg–Witten theory; Standard Model of Particle Physics.

Further Reading

Chan Hong-Mo and Tsou Sheung Tsun (1993) *Some Elementary Gauge Theory Concepts*. Singapore: World Scientific.

Chan Hong-Mo and Tsou Sheung Tsun (1999) Nonabelian generalization of electric–magnetic duality – a brief review. *International Journal of Modern Physics* A14: 2139–2172.

Chan Hong-Mo and Tsou Sheung Tsun (2002) Fermion generations and mixing from dualized standard model. *Acta Physica Polonica* B33(12): 4041–4100.

Polyakov AM (1980) Gauge fields as rings of glue. *Nuclear Physics* B164: 171–188.

Schwarz John H (1997) Lectures on superstring and M theory dualities. *Nuclear Physics Proceedings Supplements* 55B: 1–32.

't Hooft G (1978) On the phase transition towards permanent quark confinement. *Nuclear Physics* B138: 1–25.

Wu Tai Tsun and Yang Chen Ning (1976) Dirac monopoles without strings: classical Lagrangian theory. *Physical Review* D14: 437–445.

Yang Chen Ning and Mills RL (1954) Conservation of isotopic spin and isotopic gauge invariance. *Physical Review* 96: 191–195.

Electroweak Theory

K Konishi, Università di Pisa, Pisa, Italy

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The discovery of the electroweak theory crowned long years of investigation on weak interactions. The key earlier developments included Fermi's phenomenological four-fermion interactions for the β-decay, discovery of parity violation and establishment of $V-A$ structure of the weak currents, the Feynman–Gell–Mann conserved vector current (CVC) hypothesis, current algebra and its beautiful applications in the 1960s, Cabibbo mixing and lepton–hadron universality, and finally, the proposal of intermediate vector bosons (IVBs) to mitigate the high-energy behavior of the pointlike Fermi's interaction theory.

It turned out that the scattering amplitudes in IVB theory still generally violated unitarity, due to the massive vector boson propagator,

$$\frac{-g^{\mu\nu}+q^{\mu}q^{\nu}/M^2}{q^2-M^2+\mathrm{i}\epsilon}$$

The electroweak theory, known as Glashow–Weinberg–Salam (GWS) theory (Weinberg 1967, Salam 1968, Taylor 1976), was born through the attempts to make the hypothesis of IVBs for the weak interactions such that it is consistent with unitarity.

The GWS theory contains, and is in a sense a generalization of, quantum electrodynamics (QED) which was earlier successfully established as the quantum theory of electromagnetism in interaction with matter. GWS theory describes the weak and electromagnetic interactions in a single, unified gauge theory with gauge group

$$\mathrm{SU_L}(2)\times\mathrm{U}(1) \qquad [1]$$

Part of this gauge symmetry is realized in the so-called "spontaneously broken" mode; only a $\mathrm{U_{EM}}(1)\subset\mathrm{SU_L}(2)\times\mathrm{U}(1)$ subgroup, corresponding to the usual local gauge symmetry of the electromagnetism, remains manifest at low energies, with a massless gauge boson (photon). The other three gauge bosons $W^{\pm},Z$, are massive, with masses $\sim$80.4 and 91.2 GeV, respectively.

The theory is renormalizable, as conjectured by S Weinberg and by A Salam, and subsequently proved by G 't Hooft (1971), and makes well-defined predictions order by order in perturbation theory.

Since the experimental observation of neutral currents (a characteristic feature of the Weinberg–Salam theory which predicts an extra, neutral massive vector boson, Z, as compared to the naive IVB hypothesis) at Gargamelle bubble chamber at CERN (1973), the theory has passed a large number of experimental tests. The first basic confirmation also included the discovery of various new particles required by the theory: the charm quark (SLAC, BNL, 1974), the bottom quark (Fermilab, 1977), and the tau (τ) lepton (SLAC, 1975). The heaviest top quark, having mass about two hundred times that of the proton, was found later (Fermilab, 1995). The direct observation of W and Z vector bosons was first made by UA1 and UA2 experiments at CERN (1983).

The GWS theory is today one of the most precise and successful theories in physics. Even more important, perhaps, together with quantum chromodynamics (QCD), which is a SU(3) (color) gauge theory describing the strong interactions (which bind quarks into protons and neutrons, and the latter two into atomic nuclei), it describes correctly – within the present experimental and theoretical uncertainties – all the presently known fundamental forces in Nature, except gravity. The $\mathrm{SU}(3)_{\mathrm{QCD}}\times(\mathrm{SU_L}(2)\times\mathrm{U}(1))_{\mathrm{GWS}}$ theory is known as the standard model (SM).

Both the electroweak (GSW) theory and QCD are gauge theories with a nonabelian (noncommutative) gauge group. This type of theories, known as Yang–Mills theories, can be constructed by generalizing the well-known gauge principle of QED to more general group transformations. It is a truly remarkable fact that all of the fundamental forces known today (apart from gravity) are described by Yang–Mills theories, and in this sense a very nontrivial unification can be said to underlie the basic laws of Nature (G 't Hooft).

There are further deep and remarkable conditions (anomaly cancellations), satisfied by the structure of the theory and by the charges of experimentally known spin-1/2 elementary particles (see **Tables 1** and **2**), which guarantees the consistency of the theory as a quantum theory.

It should be mentioned, however, that the recent discovery of neutrino oscillations (SuperKamiokande (1998), SNO, KamLAND, K2K experiments), which proved the neutrinos to possess nonvanishing masses, clearly indicates that the standard GWS theory must be extended, in an as yet unknown way.

Table 1 Quarks and their charges

Quarks	$SU_L(2)$	$U_Y(1)$	$U_{EM}(1)$
$\begin{pmatrix} u_L \\ d'_L \end{pmatrix}, \begin{pmatrix} c_L \\ s'_L \end{pmatrix}, \begin{pmatrix} t_L \\ b'_L \end{pmatrix}$	$\underline{2}$	$\frac{1}{3}$	$\begin{pmatrix} \frac{2}{3} \\ -\frac{1}{3} \end{pmatrix}$
u_R, c_R, t_R	$\underline{1}$	$\frac{4}{3}$	$\frac{2}{3}$
d_R, s_R, b_R	$\underline{1}$	$-\frac{2}{3}$	$-\frac{1}{3}$

The primes indicate that the mass eigenstates are different from the states transforming as multiplets of $SU_L(2) \times U_Y(1)$. They are linearly related by CKM mixing matrix.

Table 2 Leptons and their charges

Leptons	$SU_L(2)$	$U_Y(1)$	$U_{EM}(1)$
$\begin{pmatrix} \nu'_{eL} \\ e_L \end{pmatrix}, \begin{pmatrix} \nu'_{\mu L} \\ \mu_L \end{pmatrix}, \begin{pmatrix} \nu'_{\tau L} \\ \tau_L \end{pmatrix}$	$\underline{2}$	-1	$\begin{pmatrix} 0 \\ -1 \end{pmatrix}$
e_R, μ_R, τ_R	$\underline{1}$	-2	-1

The primes indicate again that the mass eigenstates are in different from the states transforming as multiplets of $SU_L(2) \times U_Y(1)$, as required by the observed neutrino oscillations.

The following is a brief summary of the GWS theory, its characteristic features, its implications to the symmetries of Nature, the status of the precision tests, and its possible extensions.

GWS Theory

All the presently known elementary particles (except for the gauge bosons $W^{\pm}, Z, \gamma$, the gluons, the graviton, possibly right-handed neutrinos) are listed in **Tables 1–3** together with their charges with respect to the $SU_L(2) \times U(1)$ gauge group.

A doublet of Higgs scalar particles is included even though the physical component (which should appear as an ordinary scalar particle) has not yet been experimentally observed.

The Lagrangian is given by

$$\mathcal{L} = \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{quarks}} + \mathcal{L}_{\text{leptons}} + \mathcal{L}_{\text{Higgs}} + \mathcal{L}_{\text{Yukawa}} + \mathcal{L}_{\text{g.f.}} + \mathcal{L}_{\text{ghosts}}$$

The gauge kinetic terms are

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{4}\sum_{a=1}^{3} F^a_{\mu\nu}F^{a\mu\nu} - \frac{1}{4}G_{\mu\nu}G^{\mu\nu}$$

Table 3 Higgs doublet scalars and their charges

Higgs doublet	$SU_L(2)$	$U_Y(1)$	$U_{EM}(1)$
$\begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}$	$\underline{2}$	1	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$

where

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g\epsilon^{abc}A^b_\mu A^c_\nu$$
$$G_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$$

are $SU_L(2) \times U(1)$ gauge field tensors; $\mathcal{L}_{\text{g.f.}}$ and $\mathcal{L}_{\text{FP}}$ are the so-called gauge-fixing term and Faddeev–Popov ghost term, needed to define the gauge-boson propagators appropriately and to eliminate certain unphysical contributions. The gauge invariance of the theory is ensured by a set of identities (A Slavnov, J C Taylor). The quark kinetic terms have the form

$$\mathcal{L}_{\text{quarks}} = \sum_{\text{quarks}} \bar{\psi} i\gamma^\mu \mathcal{D}_\mu \psi$$

where $\mathcal{D}_\mu$ are appropriate covariant derivatives,

$$\mathcal{D}_\mu q_L = \left(\partial_\mu - \frac{ig}{2}\tau \cdot A_\mu - \frac{ig'}{6}B_\mu\right) q_L$$

for the left-handed quark doublets,

$$\mathcal{D}_\mu u_R = \left(\partial_\mu - \frac{2ig'}{3}B_\mu\right) u_R$$
$$\mathcal{D}_\mu d_R = \left(\partial_\mu + \frac{ig'}{3}B_\mu\right) d_R$$

and similarly for other "up" quarks c_R (charm) and t_R (top), and "down" quarks, s_R (strange), and b_R (bottom). Analogously, the lepton kinetic terms are given by

$$\mathcal{L}_{\text{leptons}} = \sum_{i=1}^{3} \bar{\psi}^i i\gamma^\mu \mathcal{D}_\mu \psi^i$$
$$= \sum_{i=1}^{3} \bar{\psi}^i_L i\gamma^\mu \left(\partial_\mu - ig\frac{\tau^a A^a_\mu}{2} + \frac{ig'}{2}B_\mu\right)\psi^i_L$$
$$+ \sum_{i=1}^{3} \bar{\psi}^i_R i\gamma^\mu (\partial_\mu + ig'B_\mu)\psi^i_R$$

where $\psi^i (i = 1, 2, 3)$ indicate the e, μ, τ lepton families; finally, the parts involving the Higgs fields are

$$\mathcal{L}_{\text{Higgs}} = \mathcal{D}_\mu \phi * \mathcal{D}^\mu \phi + V(\phi, \phi^\dagger)$$
$$V(\phi, \phi^\dagger) = -\mu^2 \phi^\dagger \phi - \lambda(\phi^\dagger \phi)^2$$

and

$$\mathcal{L}_{\text{Yukawa}} = \sum_{i,j=1}^{3} \left[g_d^{ij} \bar{q}^i_L \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} d^j_R \right.$$
$$\left. + g_u^{ij} \bar{q}^i_L \begin{pmatrix} \phi^{0*} \\ -\phi^- \end{pmatrix} u^j_R \right] + \text{h.c.}$$
$$+ \sum_{i,j=1}^{3} \left[g_e^j \bar{\psi}^i_L \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \psi^j_R \right] + \text{h.c.} \qquad [2]$$

For $\mu^2 < 0$, the Higgs potential has a minimum at

$$\langle \phi^\dagger \phi \rangle = \langle |\phi^+|^2 + |\phi^0|^2 \rangle = -\frac{\mu^2}{2\lambda} \equiv \frac{v^2}{2} \neq 0$$

By choosing conveniently the direction of the Higgs field, its vacuum expectation value (VEV) is expressed as

$$\left\langle \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \right\rangle = \begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix}, \quad v \equiv \sqrt{-\frac{\mu^2}{\lambda}} \qquad [3]$$

The physical properties of Higgs and gauge bosons are best seen by choosing the so-called unitary gauge,

$$\Phi(x) = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = e^{i\zeta^a(x)\tau^2/v} \begin{pmatrix} 0 \\ (v + \eta(x))/\sqrt{2} \end{pmatrix} \equiv U(\zeta)\Phi'(x)$$

$$\psi_L = U(\zeta)\psi'_L, \qquad \psi_R = \psi'_R$$

$$A_\mu = U(\zeta)\left(A'_\mu + \frac{i}{g}\partial_\mu\right)U^{-1}(\zeta), \qquad A_\mu \equiv \frac{\tau^a A^a_\mu}{2}$$

and expressing everything in terms of primed variables. It is easy to see that

1. There is one physical scalar (Higgs) particle with mass,

$$m_\eta = \sqrt{-2\mu^2} \qquad [4]$$

2. The Higgs kinetic term $(\mathcal{D}\phi'^\dagger)(\mathcal{D}\phi')$ produces the gauge-boson masses

$$M^2_{W^\pm} = \frac{g^2 v^2}{4}, \qquad M^2_Z = \frac{v^2}{4}(g^2 + g'^2) \qquad [5]$$

3. The physical gauge bosons are the charged $W^\pm$, and two neutral vector bosons described by the fields

$$Z_\mu = \cos\theta_W A_{3\mu} - \sin\theta_W B_\mu,$$
$$A_\mu = \sin\theta_W A_{3\mu} + \cos\theta_W B_\mu$$

where the mixing angle

$$\theta_W = \tan^{-1}\frac{g'}{g} \quad \left(\sin\theta_W = \frac{g'}{\sqrt{g^2 + g'^2}}\right)$$

is known as the Weinberg angle. The massless A_μ field describes the photon.

Fermi Interactions and Neutral Currents

The fermions interact with gauge bosons through the charge and neutral currents

$$\mathcal{L} = \frac{g}{2}\left(J_{-\mu}W^\mu_+ + J_{+\mu}W^\mu_-\right) + \mathcal{L}^{\text{n.c.}} \qquad [6]$$

$$\mathcal{L}^{\text{n.c.}} = gJ^3_\mu A^{3\mu} + \frac{g'}{2}J^Y_\mu B^\mu$$
$$= eJ^\mu_{\text{em}}A_\mu + \frac{g}{\cos\theta_W}J^0_\mu Z^\mu \qquad [7]$$

where

$$J_{+\mu} = \sum \bar{\psi}_L \gamma_\mu \tau^+ \psi_L$$
$$= \frac{1}{2}\sum \bar{\psi}\gamma_\mu\tau^+(1 - \gamma_5)\psi$$
$$\equiv \frac{1}{2}J^{V-A}_{+\mu} \qquad [8]$$

corresponds to the standard charged current, and

$$J^0_\mu = J^3_\mu - \sin^2\theta_W J^{\text{em}}_\mu \qquad [9]$$

is the neutral current to which the Z boson is coupled ($J^3_\mu = (1/2)\sum \bar{\psi}_L\gamma_\mu\tau^3\psi_L$ and J^{em}_μ is the electromagnetic current). The model thus predicts the existence of neutral current processes, mediated by the Z boson, such as $\nu_\mu e \to \nu_\mu e$ or $\bar{\nu}_\mu e \to \bar{\nu}_\mu e$, with cross section of the same order of that for the charged current process, $\bar{\nu}_e e \to \bar{\nu}_e e$, but with a characteristic L–R asymmetric couplings depending on the Weinberg angle. By eqn [9] appropriate ratios of cross sections, such as $\sigma(\nu_\mu e \to \nu_\mu e)/\sigma(\bar{\nu}_\mu e \to \bar{\nu}_\mu e)$, can be used to measure $\sin^2\theta_W$.

The exchange of heavy W bosons generates an effective current–current interaction at low energies:

$$\mathcal{L}^{\text{c.c}}_{\text{eff}} = -\frac{g^2}{2M^2_W}J_{-\mu}J^\mu_+$$

the well-known Fermi–Feynman–Gell-Mann Lagrangian $-\frac{G_F}{\sqrt{2}}J^\dagger_{V-A\mu}J^\mu_{V-A}$, with

$$\frac{G_F}{\sqrt{2}} = \frac{g^2}{8M^2_W}$$

This means that the Higgs VEV must be taken to be

$$v = 2^{-1/4}G_F^{-1/2} \simeq 246\ \text{GeV} \qquad [10]$$

Masses

It is remarkable that "all" known masses of the elementary particles – except perhaps those of the neutrino masses – are generated in GWS theory through the spontaneous breakdown of $SU_L(2) \times U(1)$ symmetry, through the Higgs VEV (eqns [3] and [10]). The boson masses are given by [4] and [5]. Note that the relation

$$\rho = \frac{M^2_W}{M^2_Z\cos^2\theta_W} = 1 + O(\alpha)$$

reflects an accidental SO(3) symmetry present (note the SO(4) symmetry of the Higgs potential in the limit $\alpha \to 0$, before the spontaneous breaking) in the model, called custodial symmetry. This is a characteristic, model-dependent feature of the minimal model, not

necessarily required by the gauge symmetry. This relation is well met experimentally, although a quantitative discussion requires the choice of the renormalization scheme (including the definition of sin θ_W itself) and check of consistency with various other data.

The fermions get mass through the Yukawa interactions (eqn [2]); the fermion masses are arbitrary parameters of the model and cannot be predicted within the GWS theory. An important feature of this mechanism is that the coupling of the physical Higgs particle to each fermion is proportional to the mass of the latter. This should give a clear, unambiguous experimental signature for the Higgs scalar of the minimal GWS model.

The recent discovery of nonvanishing neutrino masses requires the theory to be extended. Actually, there is a natural way to incorporate such masses in the standard GWS model, by a minimal extension. As the right-handed neutrinos, if they exist, are entirely neutral with respect to the $SU_L(2) \times U(1)$ gauge symmetry, they do not need its breaking to have mass. In other words, ν_R may get Majorana masses, $\sim M_R \nu_R \nu_R$, by some yet unknown mechanism, much larger than those of other fermions (such a mechanism is quite naturally present in some grand unified models). If now the Yukawa couplings are introduced as for the quarks and for the down leptons, then the Dirac mass terms result upon condensation of the Higgs field, and the neutrino mass matrix would take the form, for one flavor (in the space of $(\nu_L, \bar{\nu}_R)$):

$$\begin{pmatrix} 0 & m_D \\ m_D & M_R \end{pmatrix} \qquad [11]$$

Table 4 Quark masses

u (MeV)	c (GeV)	t (GeV)	d (MeV)	s (MeV)	b (GeV)
1.5–4	1.15–1.35	174.3 ± 5.1	4–8	80–130	4.1–4.4

Table 5 Leptons masses

ν_e(eV)	ν_μ(MeV)	ν_τ(MeV)
<3	<0.19	<18.2
e (MeV)	μ (MeV)	τ (MeV)
$0.510\,998\,92 \pm 4 \times 10^{-8}$	$105.658369 \pm 9 \times 10^{-6}$	1776.99 ± 0.26

Table 6 Gauge-boson masses

Photon	*Gluons*	$W^{\perp}$(GeV)	Z (GeV)
0	0	80.425 ± 0.038	91.1876 ± 0.0021

If the Dirac masses are assumed to be of the same order of those of the quarks and if the right-handed Majorana masses M_R are far larger, for example, of the order of the grand unified scale, $O(10^{16}\,\text{GeV})$, then diagonalization of the mass matrix would give, for the physical masses of the left-handed neutrinos, $\sim m_D^2/M_R \ll m_D$, much smaller than other fermion masses, quite naturally ("see-saw" mechanism).

CKM Quark Mixing As there is *a priori* no reason why the weak-interaction eigenstates should be equal to the mass eigenstates, the Yukawa couplings in eqn [2] are in general nondiagonal matrices in the flavor. Suppose that the the weak base for the quarks is given in terms of the mass eigenstates (in which quark masses are made diagonal), by unitary transformations

$$u_{Li} = \sum_j V_{ij}^{up} \tilde{u}_{Lj}, \qquad d_{Li} = \sum_j V_{ij}^{down} \tilde{d}_{Lj}$$

then the interaction terms with $W^{\pm}$ bosons [6] can be cast in the form (Kobayashi and Maskawa 1972)

$$\mathcal{L}^{W\text{-exc}} = \bar{u}_L^i \gamma^\mu W_\mu^+ U_{ij}^{(CKM)} d_L^j + \bar{d}_L^k \gamma^\mu W_\mu^- U_{k\ell}^{(CKM)\dagger} u_L^\ell \qquad [12]$$

where $U_{ij}^{CKM} \equiv (V^{up\dagger} \cdot V^{down})_{ij}$ is called Cabibbo–Kobayashi–Maskawa (CKM) matrix. It can be parametrized in terms of three Euler angles and one phase

$$U = \begin{pmatrix} U_{ud} & U_{us} & U_{ub} \\ U_{cd} & U_{cs} & U_{cb} \\ U_{td} & U_{ts} & U_{tb} \end{pmatrix} = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta_{13}} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta_{13}} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta_{13}} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta_{13}} & c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta_{13}} & c_{23}c_{13} \end{pmatrix} \qquad [13]$$

where $c_{12} = \cos\theta_{12}, s_{23} = \sin\theta_{23}$, etc. The requirement that charge–current weak processes are all described by these matrix elements, satisfying the unitarity relation,

$$\sum_\ell U_{i\ell}^{CKM} U_{\ell k}^{CKM\dagger} = \delta_{ik} \qquad [14]$$

gives a very stringent test for the validity of the model.

CP Violation

CP (product of charge conjugation and parity transformation) invariance is an approximate symmetry of Nature. Although it is known to be broken by very tiny amounts only, the exact extent and the nature of *CP* violation can have far-reaching consequences.

CP violation has first been discovered by Cronin and Fitch (BNL, 1964) in the *K*-meson system; more precise information on the nature of *CP* violation from the neutral kaon decays has been obtained more recently (2000) in NA48 (CERN) and KTeV (Fermilab) experiments. *CP* violation has been established in the *B*-meson systems as well, very recently (2002), by Babar experiments at SLAC and Belle experiments at KEK.

Through the so-called *CPT* theorem, *CP* invariance (or violation) is closely related to the *T* (time-reversal invariance) symmetry. Also, *CP* noninvariance is one of the conditions needed in the cosmological baryon number generation (baryogenesis).

In the GWS theory, with three families of quark flavors (six quarks), there is just one source of *CP* violation: the phase δ_{13} appearing in the CKM matrix (eqn [13]). For $\delta \neq 0, \pi$, *W*-exchange interactions [12] induce *CP* violation. The earlier and more recent experimental data on $K^0 - \bar{K}^0$ mixing and $K_{L,S}$ decay data appear to be compatible with the CKM mechanism for *CP* violation, but a quantitative comparison with the SM remains somewhat hindered by the difficulty of estimating certain strong interaction effects. The recent confirmation of *CP* violation in *B* systems is made in the context of a global fit with the SM predictions such as the "unitarity triangle" relations, for example,

$$1 + \frac{U_{ud}\,U^*_{ub}}{U_{cd}\,U^*_{cb}} + \frac{U_{td}\,U^*_{tb}}{U_{cd}\,U^*_{cb}} = 0 \qquad [15]$$

(eqn [14]), and by combining data from kaon deays, charmed meson decays, *B* meson decay and mixings, etc., and is a part of direct tests of the GWS model, with nonvanishing *CP* violation CKM phase (eqn [16] and **Figure 1**). Recent evidence for nonzero neutrino masses and mixings opens the way to possible *CP* violation in the leptonic processes as well.

Finally, within the SM including strong interactions, there is one more source of *CP* violation: the so-called θ (vacuum) parameter of QCD.

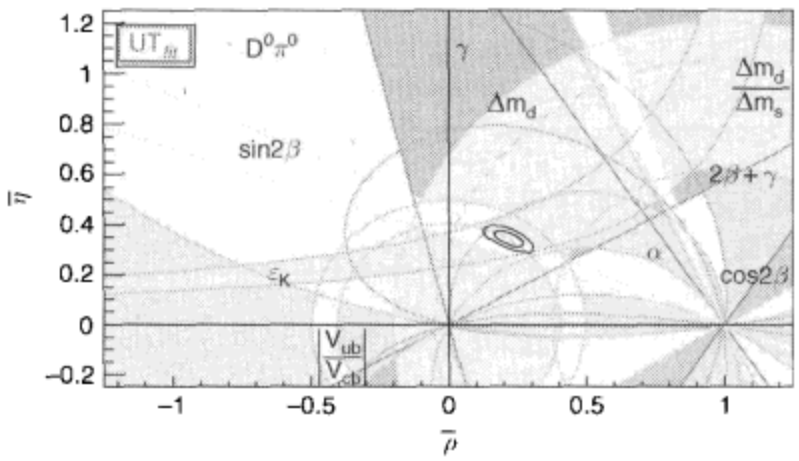


Figure 1 Unitarity triangle test (Eq. (15)). The small ellipses represent 68% and 95% probability zones for the apex corresponding to $U_{ud}U^*_{ub}/U_{cd}U^*_{cb}$. Reproduced from M. Bona *et al.* (2005) The 2004 UTfit collaboration report on the status of the unitarity triangle in the standard model. Journal of High Energy Physics. 0507: 028–059 (hep-ph/0501199), with permission from IoP Publishing Ltd and the UTfit collaboration.

B and *L* Nonconservation

Another set of approximate symmetries in Nature are the baryon and lepton number conservations. In the electroweak theory, these global symmetries are exact to all orders of perturbation theory. Nonperturbative effects (a sort of barrier penetration in gauge field space) however violate both *B* and *L*; the combination *B*–*L* is conserved even nonperturbatively though. The nonperturbative electroweak baryon number violation is an extremely tiny effect, the amplitude being proportional to the typical tunneling factor $e^{-2\pi/\alpha}$, but the process is unsuppressed at finite temperatures as might have been experienced by the universe at some early stage after big bang.

B or *L* nonconservation can also arise naturally at high energy scales, if the electroweak theory is embedded as the low-energy approximation in a grand unified model. The experimental lower limit of proton lifetime, $\tau_p \geq 10^{32}$ years, from Kamiokande experiments, however severely restricts acceptable models of this type (the simplest SU(5) model is already ruled out).

On the other hand, cosmological baryogenesis requires sufficient amount of baryon number violation, at least in some stage of cosmological expansion. Detailed analyses suggest that the standard electroweak transition might not in itself explain the baryon number $n_p/n_\gamma \sim 10^{-10}$ observed in the present universe. Recent observations of neutrino oscillations suggest the right-handed Majorana-type neutrino masses to be present, which violate the lepton number *L*. In such a case it might be possible that the correct amount of baryon number excess would be generated, through the leptogenesis.

Global Fit

Various relations exist at the tree level among the masses, scattering cross sections, decay rates, various asymmetries, etc., which can be read off or calculated from the formulas given earlier. These quantities receive corrections at higher orders, and the experimental checks of these modified relations provide precision tests of the model on the one hand, and possibly a hint for new physics, if there is any discrepancy with the prediction. Very often the amplitudes of interest

receive important contributions due to strong interactions, which are difficult to estimate.

The basic parameters of the model, apart from the Higgs mass, and fermion masses and mixing parameters, can be taken to be (1) the fine structure constant, $\alpha = 1/137.035\,999\,11(46)$; (2) the Fermi constant $G_F = 1.166\,37 \times 10^{-5}\,\mathrm{GeV}^{-2}$ (which can be determined from the muon lifetime), and the Z-boson mass, $M_Z = 91.1876 \pm 0.0021\,\mathrm{GeV}$ (observed directly at LEP). M_W and $\sin^2\theta_W$ are then calculable numbers, in terms of these quantities, and depending on m_t (measured independently by CDF and DØ experiments at Fermilab) and on the unknown M_H.

Such precision tests of the GWS model are being made, combining the analyses of various decay rates and asymmetries in B-meson systems at B factories and in colliders, production and decays of Z and W bosons, elastic νe or $\bar{\nu} e$ scatterings, elastic νp or $\bar{\nu} p$ scatterings, deep inelastic lepton nucleon (or deuteron) scatterings, the muon anomalous magnetic moment, atomic parity violation experiments, etc.

An overall fit to the data gives an excellent agreement, with the input parameters

$$M_H = 113^{+56}_{-40}\,\mathrm{GeV}, \quad m_t = 176.9 \pm 4.0\,\mathrm{GeV},$$
$$\alpha_s(M_Z) = 0.1213 \pm 0.0018$$

For instance (in GeV),

$$M_W = 80.390 \pm 0.018 \quad \text{vs.} \quad 80.412 \pm 0.042$$
(exp. value (LEP))
$$\Gamma_Z = 2.4972 \pm \pm 0.0012 \quad \text{vs.} \quad 2.4952 \pm 0.0023$$
(exp. value)

For $\sin^2\theta_W$ (defined in the so-called $\overline{MS}$ scheme) all data give consistently the value

$$\sin^2\theta_W = 0.231\,20 \pm 0.000\,15$$

(a slightly larger value is reported by an νN experiment at Fermilab).

The unitarity-triangle tests of the SM and determination of CKM matrix have already been mentioned. The results of global fit can be summarized in **Figure 1**, and by the angles

$$\begin{aligned} s_{12} &= 0.2243 \pm 0.0016 \\ s_{23} &= 0.0413 \pm 0.0015 \\ s_{13} &= 0.037 \pm 0.0005 \\ \delta_{13} &= 60^\circ \pm 14^\circ \end{aligned} \qquad [16]$$

For the muon anomalous gyromagnetic ratio $(g-2)$, the experimental data

$$a_\mu^{\mathrm{exp}} = \frac{g_\mu - 2}{2} = (1.116\,5920(37) \pm 0.78) \times 10^{-9}$$

is to be compared with the theoretical prediction

$$a_\mu^{\mathrm{th}} = (1.1165918(83) \pm 0.49) \times 10^{-9}$$

which is slightly smaller (1.9σ), where the largest theoretical uncertainty comes from the two-loop hadronic contribution $a_\mu^{\mathrm{had}} \simeq (69.63 \pm 0.72) \times 10^{-9}$ (the QED corrections to $O(\alpha^5)$ are included).

For further details of the analyses and the present status of experimental tests of the electroweak theory, see the reviews by J Erler and P Langacker, and by F J Gilman *et al.*, cited in "Further reading" (most of numbers cited here come from these two reviews).

Need for Extension of the Model

In spite of such an impressive experimental confirmation, there are reasons to believe that the electroweak theory, in its standard minimal form, is not a complete story. As already mentioned, neutrino oscillations, predicted earlier by Pontecorvo, have recently been experimentally confirmed, giving uncontroversial evidence for nonvanishing neutrino masses and their mixing. This is a clear signal that the theory must be extended. If the mass is instead taken in the form of eqn [11] but with three neutrinos families, the diagonalization in general yields a mixing for the light neutrinos, as for the quarks. Some of the experimental data on the neutrinos are summarized in **Table 7**.

In addition, the Higgs sector of the theory (the part of the interactions responsible for spontaneous breaking $SU_L(2) \times U(1) \hookrightarrow U_{EM}(1)$) is still largely untested. The theory predicts a physical scalar particle, the Higgs particle, of unknown mass. The present-day expectation for its mass, which combines the experimental lower limit and an indirect upper limit following from the analysis of various radiative corrections, is

$$114\,(\mathrm{GeV}) < m_H < 250\,(\mathrm{GeV})$$

This particle should be observable either in the Tevatron at Fermilab or in the coming LHC

Table 7 Neutrino mass square differences and mixing

ν_e		ν_μ		ν_τ
$\Delta_{12}m^2$	=	$(6-9)$	$\times$	$10^{-5}\,\mathrm{eV}^2$
$\Delta_{23}m^2$	=	$(1-3)$	$\times$	$10^{-3}\,\mathrm{eV}^2$

Solar neutrinos and reactor (SNO, SuperKamiokande, KamLAND) experiments give the first results. Atmospheric neutrino data and the long baseline experiment (SuperKamiokande, K2K) provide the second. The mixing angle relevant to the solar and reactor neutrino oscillation is large, $\tan^2\theta_{12} \sim 0.40^{+0.10}_{-0.07}$, while the one related to the atmospheric neutrino data is maximal, $\sin^2 2\theta_{23} \sim 1$. Cosmological considerations give $\sum m_{\nu_i} < O(1\,\mathrm{eV})$.

experiments at CERN; negative results would force upon us a substantial modification of the electroweak theory.

Last, but not least, there are a few theoretical motivations for an extension of the model to be considered necessary. First, the structure of the GWS theory is not entirely determined by the gauge principle. The form of the Higgs self-interactions, as well as their number and the Yukawa couplings of the Higgs scalar to the fermions, are unconstrained by any principle, and the particular, minimal form assumed by Weinberg and Salam is yet to be confirmed experimentally.

In addition, the theory is not really a unified gauge theory: $SU_L(2)$ and U(1) gauge couplings are distinct. One possibility is that the $SU(3)_{QCD} \times SU_L(2) \times U(1)$ theory of the SM is actually a low-energy manifestation of a truly unified gauge theory – grand unified theory (GUT) – defined at some higher mass scale. The simplest version of GUT models based on SU(5) or SO(10) gauge groups has however a difficulty with the proton decay rates, and with the coupling-constant unification itself. Supersymmetric GUTs appear to be more acceptable both from the coupling-constant unification and from the proton lifetime constraints.

A more subtle, but perhaps more severe theoretical problem, is the so-called naturalness problem. At the quantum level, due to the quadratic divergences in the scalar mass, the structure of the theory turns out to be quite peculiar. If the ultraviolet cutoff of the theory is taken to be the Planck mass scale, $\Lambda_{UV} \sim m_{Pl} \sim 10^{19}\,\text{GeV}$, at which gravity becomes strongly coupled, the theory at Λ_{UV} would have to possess parameters which are fine-tuned with an excessive precision. The problem is known also as a "hierarchy" problem.

A way to avoid having such a difficulty is to introduce supersymmetry. In a supersymmetric version of the standard theory – in fact, there are phenomenologically well-acceptable models such as the minimal supersymmetric standard model (MSSM) – this problems is absent due to the cancellation of bosonic and fermionic loop contributions typical of supersymmetric theories. As a result, the properties of the theory at low energies are much less sensitive to those of the theory at the Planck mass scale. Experiments at LHC (expected to be performed after 2008, CERN) should be able to produce a whole set of new particles associated with supersymmetry, if this is a part of the physical law beyond TeV energies.

At a deeper level, however, the hierarchy problem in a more general sense persists, even in supersymmetric models: why the masses of the order of O(100 GeV) should appear at all in a theory with a natural cutoff of the order of the Planck mass? Furthermore, if the masses of the neutrinos turn out to be of the order of $O(10^{-3}–10^{0})\,\text{eV}$, we are left with the problem of understanding the large disparities among the quark and lepton masses, spanning the range of more than 13 orders of magnitudes: another "hierarchy" problem.

It is also possible that the spacetime the physical world lives in is actually higher dimensional: the usual four-dimensional Minkowski spacetime times either compactified or uncompactified "extra dimensions." In theories of this type, some of the difficulties mentioned above might find a natural solution. It is yet to be seen whether a consistent theory of this type can be constructed that correctly account for the properties of the universe we inhabit.

Bibliographic Notes

A short but comprehensive introduction to the Weinberg–Salam theory is found in Taylor (1976); see also Abers and Lee (1973) and 't Hooft and Veltman (1973).

The reprint collection edited by Taylor (2001) contains many of fundamental papers, e.g., on Yang–Mills theories (by CN Yang, RL Mills, R Shaw), on spontaneous symmetry breaking and its application to gauge theories (by Y Nambu, J Schwinger, PW Anderson, PW Higgs, F Englert, R Brout, TWB Kibble) and on renormalization of Yang–Mills theories and application to the electroweak theory (LD Faddeev, VN Popov, G 't Hooft).

For up-to-date review on precision tests of the GWS theory, and details of the analyses, see Erler and Langacker (2004), and references cited therein.

For a recent review on neutrino experiments, see Shirai (2005). For theory on neutrinos, see Fukugita and Yanagida (2003).

For the unitarity triangle test of the GWS model and determination of CKM matrix elements, see results from the CKM fitter Group (Bret *et al.* 2005), the UTfit Collaboration (Bona *et al.* 2005), and a review by Gilman *et al.* (2005), and references cited therein.

See also: Abelian and Nonabelian Gauge Theories using Differential Forms; Current Algebra; Effective Field Theories; Finite Group Symmetry Breaking; Noncommutative Tori, Yang–Mills, and String Theory; Quantum Chromodynamics; Quantum Electrodynamics and Its Precision Tests; Quantum Field Theory: A Brief Introduction; Renormalization: General Theory; Standard Model of Particle Physics; Symmetries and Conservation Laws; Symmetry and Symmetry Breaking in Dynamical Systems; Symmetry Breaking in Field Theory.

Further Reading

Abers ES and Lee BW (1973) Gauge theories. *Physics reports* C9: 1–141.

Bona M, Ciuchina M, and Franco E *et al.* (2005) The 2004 UTfit collaboration report on the status of the unitarity triangle in the standard model. *Journal of High Energy Physics* 0507: 028–059 (hep-ph/0501199).

Charles J, Hocker A, Lacker H *et al.* (2005) CP violation and the CKM matrix: assesing the impact of the asymmetric B factories. *European Physical Journal C* 41: 1–131 (hep-ph/0406184).

Erler J and Langacker P (2004) Electroweak Model and constraints on new physics. In: Eidelman S, Hayes KG, Olive KA *et al.* (ed.) *Review of Particle Physics*, Particle Data Group, Physics Letters B vol. 592, pp. 1–1109, (http://pdglbl.gov/2005).

Fukugita M and Yanagida T (2003) *Physics of Neutrinos and Applications to Astrophysics*. Berlin: Springer.

Gilman FJ, Kleinknecht K, and Renk B The Cabibbo–Kobayashi–Maskawa quark-mixing matrix. In Eidelman S, Hayes KG, Olive KA *et al.* (eds.) PDG *Review of Particle Physics*, Particle Data Group, Physics Letters B, vol. 592, pp. 1–1109 (http://pdg.lbl.gov/2005).

Glashow SL, Iliopoulos J, and Maiani L (1970) Weak interactions with Lepton–Hadron symmetry. *Physical Review D*, 2: 1285–1292.

Kobayashi M and Maskawa T (1972) *CP* violation in renormalizable theory of weak interaction. *Progress in Theoretical Physics* 49: 652–657.

Salam A (1968) Weak and electromagnetic interactions. In: Svartholm N (ed.) *Elementary Particle Theory*, p. 367. Lerum, Sweden: Almqvist Forlag AB.

Shirai J (2005) Neutrino experiments: review of recent results. *Nuclear Physics B (Proc. Suppl.)* 144: 286–296.

Taylor JC (1976) *Gauge Theories of Weak Interactions*. Cambridge: Cambridge University Press.

Taylor JC (ed.) (2001) *Gauge Theories in the Twentieth Century*. London: Imperial College Press.

't Hooft G (1971) Renormalizable Lagrangians for massive Yang–Mills fields. *Nuclear Physics* B35: 167–188.

't Hooft G and Veltman M (1973) Diagrammar. *CERN Yellow Report* 73–9.

Weinberg S (1967) Theory of Leptons. *Physical Review Letters* 19: 1264–1266.

Exact Renormalization Group

P K Mitter, Université de Montpellier 2, Montpellier, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The renormalization group (RG) in its modern form was invented by K G Wilson in the context of statistical mechanics and Euclidean quantum field theory (EQFT). It offers the deepest understanding of renormalization in quantum field theory (QFT) by connecting EQFT with the the theory of second-order phase transition and associated critical phenomena. Thermodynamic functions of many statistical mechanical models (the prototype being the Ising model in two or more dimensions) exhibit power-like singularities as the temperature approaches a critical value. One of the major triumphs of the Wilson RG was the prediction of the exponents (known as critical exponents) associated to these singularities. Wilson's fundamental contribution was to realize that many length scales begin to cooperate as one approaches criticality and that one should disentangle them and treat them one at a time. This leads to an iterative procedure known as the "renormalization group." Singularities and critical exponents then arise from a limiting process. Ultraviolet singularities of field theory can also be understood in the same way. Wilson reviews this (Wilson and Kogut 1974) and gives the historical genesis of his ideas (Wilson 1983).

The early work in the subject was heuristic, in the sense that clever but uncontrolled approximations were made to the exact equations often with much success. Subsequently, authors with mathematical bent began to use the underlying ideas to prove theorems. Benfatto, Cassandro, Gallavotti, Nicolo, Olivieri *et al.* pioneered the rigorous use of Wilson's renormalization group in the construction of super-renormalizable QFTs, (see Benfatto and Gallavotti (1995) and references therein). The subject saw further mathematical development in the work of Gawedzki and Kupiainen (1984, 1986) and that of Balaban (1982), and references therein. Balaban in a series of papers ending in Bałaban (1989) proved a basic result on the continuum limit of Wilson's lattice gauge theory. Brydges and Yau (1990) simplified the mathematical treatment of the renormalization group for a class of models and this has led to further systemization and simplification in the work of Brydges *et al.* (1998, 2003). Another method which has been intensely developed during the same historical period is based on phase cell expansions: Feldman, Magnen, Rivasseau, and Sénéor developed the early phase cell ideas of

Notes:

Glimm and Jaffe and were able to prove independently many of the results cited earlier (see Rivasseau (1991) and references therein). Although these methods share many features of the Wilson RG, they are different in methodology and thus remain outside of the purview of the present exposition.

A somewhat different line of development has been the use of the RG to give simple proofs of perturbative renormalizability of various QFTs: Gallavotti and Nicolo, via iterative methods (see Benfatto and Gallavotti (1995) and references therein), and Polchinski (1984), who exploited a continuous version of the RG for which Wilson (1974) had derived a nonlinear differential equation. These early works were devoted to the standard $(\phi^4)_4$ scalar field theory, but subsequently Polchinski's work has been extended to a large class of models, including four-dimensional nonabelian gauge theories (*see* Kopper and Muller (2000) and references therein).

Finally, it should be mentioned that apart from QFT and statistical mechanics, the RG method has proved fruitful in other domains. An example is the study of interacting fermion systems in condensed matter physics (*see* Fermionic Systems and Renormalization: Statistical Mechanics and Condensed Matter). In the rest of this article, our focus will be on EQFT and statistical mechanics.

The RG as a Discrete Semigroup

We will first define a discrete version of the RG and consider its continuous version later. As we will see, the RG is really a semigroup, so calling it a group is a misnomer.

Let ϕ be a Gaussian random field (see, e.g., Gelfand and Vilenkin (1964) for a discussion of random fields) in $\mathbf{R}^d$. Associated to it there is a positive-definite function which is identified as its covariance. In QFT one is interested in the covariance

$$E(\phi(x)\phi(y)) = \text{const.}\,|x-y|^{-2[\phi]} = \int_{\mathbf{R}^d} \mathrm{d}p\, \mathrm{e}^{\mathrm{i}p.(x-y)} \frac{1}{|p|^{d-2[\phi]}} \qquad [1]$$

Here $[\phi] > 0$ is the (canonical) dimension of the field, which for the standard massless free field is $[\phi] = (d-2)/2$. The latter is positive for $d > 2$. However, other choices are possible but in EQFT they are restricted by the Osterwalder–Schrader positivity. It is assured if $[\phi] = (d-\alpha)/2$, with $0 < \alpha \leq 2$. If $\alpha < 2$, we get a generalized free field.

Observe that the covariance is singular for $x = y$ and this singularity is responsible for the ultraviolet divergences of QFT. This singularity has to be initially cut off and there are many ways to do this. A simple way is as follows. Let $u(x)$ be a smooth, rotationally invariant, positive-definite function of fast decrease. Examples of such functions are legion. Observe that

$$|x-y|^{-2\phi} = \text{const.} \int_0^\infty \frac{\mathrm{d}l}{l} l^{-2[\phi]}\, u\left(\frac{x-y}{l}\right) \qquad [2]$$

as can be seen by scaling in l. We define the unit ultraviolet cutoff covariance C by cutting off at the lower end point of the l integration (responsible for the singularity at $x = y$) at $l = 1$,

$$C(x-y) = \int_1^\infty \frac{\mathrm{d}l}{l}\, l^{-2[\phi]}\, u\left(\frac{x-y}{l}\right) \qquad [3]$$

$C(x-y)$ is positive-definite and everywhere smooth. Being positive-definite, it qualifies as the covariance of a Gaussian probability measure denoted μ_C on a function space Ω (which it is not necessary to specify any further). The covariance C being smooth implies that the sample fields of the measure are μ_C almost everywhere sufficiently differentiable.

Remark Note that, more generally, we could have cut off the lower end point singularity in [1] at any $\epsilon > 0$. The ϵ-cutoff covariance is related to the unit cutoff covariance by a scale transformation (defined below) and we will exploit this relation later.

Let $L > 1$ be any real number. We define a scale transformation S_L on fields ϕ by

$$S_L\phi(x) = L^{-[\phi]}\, \phi\left(\frac{x}{L}\right) \qquad [4]$$

on covariances by

$$S_L C(x-y) = L^{-2[\phi]} C\left(\frac{x-y}{L}\right) \qquad [5]$$

and on functions of fields $F(\phi)$ by

$$S_L F(\phi) = F(S_L\phi) \qquad [6]$$

The scale transformations form a multiplicative group: $S_L^n = S_{L^n}$.

Now define a fluctuation covariance Γ_L:

$$\Gamma_L(x-y) = \int_1^L \frac{\mathrm{d}l}{l} l^{-2[\phi]}\, u\left(\frac{x-y}{l}\right) \qquad [7]$$

$\Gamma_L(x-y)$ is smooth, positive-definite and of fast decrease on scale L. It generates a key scaling decomposition

$$C(x-y) = \Gamma_L(x-y) + S_L C(x-y) \qquad [8]$$

Iterating this, we get

$$C(x-y) = \sum_{n=0}^{\infty} \Gamma_n(x-y) \qquad [9]$$

where

$$\Gamma_n(x-y) = S_{L^n}\Gamma_L(x-y) = L^{-2n[\phi]}\Gamma_L\left(\frac{x-y}{L^n}\right) \quad [10]$$

The functions $\Gamma_n(x-y)$ are of fast decrease on scale L^{n+1}.

Thus, [9] achieves the decomposition into a sum over increasing length scales as desired. Being positive definite, Γ_n qualify as covariances of Gaussian probability measures, and therefore $\mu_C = \bigotimes_{n=0}^{\infty} \mu_{\Gamma_n}$. Correspondingly introduce a family of independent Gaussian random fields ζ_n, called fluctuation fields, distributed according to μ_{Γ_n}. Then

$$\phi = \sum_{n=0}^{\infty} \zeta_n \quad [11]$$

Note that the fluctuation fields ζ_n are slowly varying over length scales L^n. In fact, an easy estimate using a Tchebycheff inequality shows that, for any $\gamma > 0$,

$$|x-y| \leq L^n$$
$$\Rightarrow \mu_C(|\zeta_n(x) - \zeta_n(y)| \geq \gamma) \leq \text{const.}\, \gamma^{-2} \quad [12]$$

which reveals the slowly varying nature of ζ_n on scale L^n. Equation [11] is an example of a multiscale decomposition of a Gaussian random field.

The above implies that the μ_C integral of a function can be written as a multiple integral over the fields ζ_n. We calculate it by integrating out the fluctuation fields ζ_n step by step, going from shorter to longer length scales. This can be accomplished by the iteration of a single transformation T_L, a renormalization group transformation, as follows. Let $F(\phi)$ be a function of fields. Then we define a RG transformation $F \to T_L F$ by

$$(T_L F)(\phi) = S_L \mu_{\Gamma_L} * F(\phi)$$
$$= \int \mathrm{d}\mu_{\Gamma_L}(\zeta)\, F(\zeta + S_L\phi) \quad [13]$$

Thus the renormalization group transformation consists of a convolution with the fluctuation measure followed by a rescaling.

Semigroup Property

The discrete RG transformations form a semigroup:

$$T_L T_{L^n} = T_{L^{n+1}} \quad \text{for all } n \geq 0 \quad [14]$$

To prove this, we must first see how scaling commutes with convolution with a measure. We have the property

$$\mu_{\Gamma_L} * S_L F = S_L \mu_{S_L\Gamma_L} * F \quad [15]$$

To see this, observe first that if ζ is a Gaussian random field distributed with covariance Γ_L then the Gaussian field $S_L\zeta$ is distributed according to $S_L\Gamma_L$. This can be checked by computing the covariance of $S_L\zeta$. Now the left-hand side of [15] is just the integral of $F(S_L\zeta + S_L\phi)$ with respect to $\mathrm{d}\mu_{\Gamma_L}(\zeta)$. By the previous observation, this is the integral of $F(\zeta + S_L\phi)$ with respect to $\mathrm{d}\mu_{S_L\Gamma_L}(\zeta)$, and the latter is the right-hand side of [15]. Now we can check the semigroup property trivially:

$$\begin{aligned} T_L T_{L^n} F &= S_L \mu_{\Gamma_L} * S_{L^n} \mu_{\Gamma_{L^n}} * F \\ &= S_L S_{L^n} \mu_{S_{L^n}\Gamma_L} * \mu_{\Gamma_{L^n}} * F \\ &= S_{L^{n+1}} \mu_{\Gamma_{L^n} + S_{L^n}\Gamma_L} * F \\ &= S_{L^{n+1}} \mu_{\Gamma_{L^{n+1}}} * F \\ &= T_{L^{n+1}} F \end{aligned} \quad [16]$$

We have used the fact that $\Gamma_{L^n} + S_{L^n}\Gamma_L = \Gamma_{L^{n+1}}$. This is because $S_{L^n}\Gamma_L$ has the representation [7] with integration interval changed to $[L^n, L^{n+1}]$.

We note some properties of T_L. T_L has an unique invariant measure, namely μ_C: for any bounded function F,

$$\int \mathrm{d}\mu_C\, T_L F = \int \mathrm{d}\mu_C\, F \quad [17]$$

To understand [17], recall the earlier observation that if ϕ is distributed according to the covariance C, then $S_L\phi$ is distributed according to $S_L C$. By [8], $\Gamma_L + S_L C = C$. Therefore,

$$\int \mathrm{d}\mu_C T_L F = \int \mathrm{d}\mu_C S_L \mu_{\Gamma_L} * F$$
$$= \int \mathrm{d}\mu_{S_L C} \mu_{\Gamma_L} * F$$
$$= \int \mathrm{d}\mu_C F \quad [18]$$

The uniqueness of the invariant measure follows from the fact that the semigroup T_L is realized by a convolution with a probability measure and, therefore, is positivity improving:

$$F \geq 0, \mu_C \text{ a.e.} \Rightarrow T_L F > 0, \mu_C \text{ a.e.}$$

Finally, note that T_L is a contraction semigroup on $L^p(\mathrm{d}\mu_C)$ for $1 \leq p < \infty$. To see this, note that since T_L is a convolution with a probability measure $T_L F = \mu_{S_{L^{-1}}\Gamma_L} * S_L F$, we have, via Hölder's inequality, $|T_L F|^p \leq T_L |F|^p$. Then use the fact that μ_C is an invariant measure.

Eigenfunctions

Let $:p_{n,m}:(\phi(x))$ be a C^∞ Wick-ordered local monomial of m fields with n derivatives. Define

$$P_{n,m}(X) = \int_X \mathrm{d}x :p_{n,m}:_C(x)$$

The functions $P_{n,m}(X)$ play the role of eigenfunctions of the RG transformation T_L up to a scaling of volume:

$$T_L P_{n,m}(X) = L^{d-m[\phi]-n} P_{n,m}(L^{-1}X) \qquad [19]$$

Because of the scaling in volume, $P_{n,m}(X)$ are not true eigenfunctions. Nevertheless, they are very useful because they play an important role in the analysis of the evolution of the dynamical system which we will later associate with T_L. They are classified as expanding (relevant), contracting (irrelevant) or central (marginal), depending on whether the exponent of L on the right-hand side of [19] is positive, negative, or zero, respectively. This depends, of course, on the space dimension d and the field dimension $[\phi]$.

Gaussian measures are of limited interest. But we can create new measures by perturbing the Gaussian measure μ_C with local interactions. We cannot study directly the situation where the interactions are in infinite volume. Instead, we put them in a very large volume which will eventually go to infinity. We have a ratio of two length scales, one from the size of the diameter of the volume and the other from the ultraviolet cutoff in μ_C, and this ratio is enormous. The RG is useful whenever there are two length scales whose ratio is very large. It permits us to do a scale-by-scale analysis and at each step the volume is reduced at the cost of changing the interactions. The largeness of the ratio is reflected in the large number of steps to be accomplished, this number tending eventually to infinity. This large number of steps has to be controlled mathematically.

Perturbation of the Gaussian Measure

Let $\Lambda_N = [-L^N/2, L^N/2]^d \subset \mathbf{R}^d$ be a large cube in $\mathbf{R}^d$. For any $X \subset \Lambda_N$, let $V_0(X, \phi)$ be a local semibounded function where the fields are restricted to the set X. Here "local" means that if X, Y are sets with disjoint interiors then $V_0(X \cup Y, \phi) = V_0(X) + V_0(Y)$. Consider the integral (known as the partition function in QFT and statistical mechanics)

$$Z(\Lambda_N) = \int \mathrm{d}\mu_C(\phi) z_0(\Lambda_N, \phi) \qquad [20]$$

where

$$z_0(X, \phi) = \mathrm{e}^{-V_0(X,\phi)} \qquad [21]$$

and

$$\mathrm{d}\mu^{(0)}(\Lambda_N, \phi) = \frac{1}{Z(\Lambda_N)} \mathrm{d}\mu_C(\phi) \mathrm{e}^{-V_0(\Lambda_N,\phi)} \qquad [22]$$

is the corresponding probability measure. V_0 is typically not quadratic in the fields and therefore leads to a non-Gaussian perturbation. For example,

$$V_0(X, \phi) = \int_X \mathrm{d}x (\xi |\nabla\phi(x)|^2 + g_0\phi^4(x) + \mu_0\phi^2(x)) \qquad [23]$$

where we take $g_0 > 0$. The integral [20] is well defined because the sample fields are smooth.

We now proceed to the scale-by-scale analysis mentioned earlier. Because μ_C is an invariant measure of T_L, we have the partition function $Z(\Lambda_N)$ in the volume Λ_N as

$$Z(\Lambda_N) = \int \mathrm{d}\mu_C(\phi) z_0(\Lambda_N, \phi) = \int \mathrm{d}\mu_C(\phi) T_L z_0(\Lambda_N, \phi) \qquad [24]$$

The integrand on the right-hand side is a new function of fields which, because of the final scaling, live in the smaller volume Λ_{N-1}. This leads to the following definition:

$$z_1(\Lambda_{N-1}, \phi) = T_L z_0(\Lambda_N, \phi) \qquad [25]$$

Because V_0 is local, z_0 has a factorization property for unions of sets with disjoint interiors. This is no longer the case for z_1. Wilson noted that, nevertheless, the integral is well approximated by an integrand which does, but the approximator has new coupling constants. The phrase "well approximated" is what all the rigorous work is about and this was not evident in the early Wilson era. The idea is to extract out a local part and also consider the remainder. The local part leads to a flow of coupling constants and the (unexponentiated) remainder is an irrelevant term. This operation and its mathematical control is an essential feature of RG analysis.

Iterating the above transformation, we get, for all $0 \le n \le N$,

$$z_{n+1}(\Lambda_{N-n-1}, \phi) = T_L z_n(\Lambda_{N-n}, \phi) \qquad [26]$$

After N iterations, we get

$$Z(\Lambda_N) = \int \mathrm{d}\mu_C(\phi) z_N(\Lambda_0, \phi) \qquad [27]$$

where Λ_0 is the unit cube. To take the limit as $N \to \infty$, we have to control the infinite sequence of iterations. We cannot hope to control the infinite sequence at the level of the entire partition function. Instead, one chooses representative coordinates for which the infinite sequence has a chance of having a meaning. The coordinates are provided by the coupling constants of the extracted local part and the irrelevant terms (an

approximate calculation of the flow of coupling constants is given in the next section). The existence of a global trajectory for such coordinates helps us to control the limit for moments of the probability measure (correlation functions). The question of coordinates and the representation of the irrelevant terms will be taken up in the section "Rigorous RG analysis."

Ultraviolet Cutoff Removal

The next issue is ultraviolet cutoff removal in field theory. This problem can be put into the earlier framework as follows. Let ϵ_N be a sequence of positive numbers which tend to 0 as $N \to \infty$. Following the remark after [3], we replace the unit cutoff covariance C by the covariance C_{ϵ_N} defined by taking ϵ_N (instead of 1) as the lower end point in the integral [3]. Thus, ϵ_N acts as a short-distance or ultraviolet cutoff. It is easy to see that

$$C_{\epsilon_N}(x-y) = S_{\epsilon_N} C(x-y) \qquad [28]$$

Consider the partition function $Z_{\epsilon_N}(\Lambda)$ in a cube $\Lambda = [-R/2, R/2]^d$:

$$Z_{\epsilon_N}(\Lambda) = \int \mathrm{d}\mu_{C_{\epsilon_N}}(\phi) \mathrm{e}^{-V_0(\Lambda, \phi, \tilde{\xi}_N, \tilde{g}_N, \tilde{\mu}_N)} \qquad [29]$$

where V_0 is given by [23] with g_0, μ_0 replaced by $\tilde{g}_N, \tilde{\mu}_N$, respectively. By dimensional analysis we can write

$$\begin{aligned} \tilde{\xi}_N &= \epsilon_N^{(2[\phi]-d+2)} \xi, \quad \tilde{g}_N = \epsilon_N^{(4[\phi]-d)} g, \\ \tilde{\mu}_N &= \epsilon_N^{(2[\phi]-d)} \mu \end{aligned} \qquad [30]$$

where g, ξ, μ are dimensionless parameters. Now ϕ distributed according to C_{ϵ_N} equals in distribution $S_{\epsilon_N}\phi$ distributed according to C. Therefore, choosing $\epsilon_N = L^{-N}$, we get

$$\begin{aligned} Z_{\epsilon_N}(\Lambda) &= \int \mathrm{d}\mu_C(\phi) \mathrm{e}^{-V_0(\Lambda, S_{\epsilon_N}\phi, \tilde{\xi}_N, \tilde{g}_N, \tilde{\mu}_N)} \\ &= \int \mathrm{d}\mu_C(\phi) \mathrm{e}^{-V_0(\Lambda_N, \phi, \xi g, \mu)} \end{aligned} \qquad [31]$$

where $\Lambda_N = [-L^N R/2, L^N R/2]^d$. Thus, the field theory problem of removing the ultraviolet cutoff, that is, taking the limit $\epsilon_N \to 0$, has been reduced to the study of a statistical mechanical model in a very large volume. The latter has to be analyzed via RG iterations as before.

Critical Field Theories

As mentioned earlier, we have to study the flow of local interactions as well as that of irrelevant terms. Together they constitute the RG trajectory and we have to prove that it exists globally. In general, the trajectory will tend to explode after a large number of iterations due to growing relevant terms (characterized in terms of the expanding Wick monomials mentioned earlier). Wilson pointed out that the saving factor is to exploit fixed points and their invariant manifolds by tuning the initial interaction so that the RG has a global trajectory. This leads to the notion of a critical manifold which can be defined as follows. A fixed point will have contracting and/or marginal attractive directions besides the expanding ones. In the language of dynamical systems, the critical manifold is the stable or center stable manifold of the fixed point in question. This is determined by a detailed study of the discrete flow. In the examples above, it amounts to fixing the initial "mass" parameter $\mu_0 = \mu_c(g_0)$ with a suitable function μ_c such that the flow remains bounded in an invariant set. The critical manifold is then the graph of a function from the space of contracting and marginal variables to the space of μ's which remains invariant under the flow. Restricted to it the flow will now converge to a fixed point. All references to initial coupling constants have disappeared. The result is known as a critical theory.

Critical theories have been rigorously constructed in a number of cases. Take the standard ϕ^4 in d dimensions. Then $[\phi] = (d-2)/2$. For $d > 5$ the ϕ^4 interaction is irrelevant and the Gaussian fixed point is attractive with one unstable direction (corresponding to μ). In this case one can prove that the interactions converge exponentially fast to the Gaussian fixed point on the critical manifold. For $d = 4$ the interaction is marginal and the Gaussian fixed point attractive for $g > 0$. The critical theory has been constructed by Gawedzki and Kupainen (1984) starting with a sufficiently small coupling constant. The fixed point is Gaussian (interactions vanish in the limit) and the convergence rate is logarithmic. This is thus a mean-field theory with logarithmic corrections, as expected on heuristic grounds. The mathematical construction of the critical theory in $d = 3$ is an open problem. (It is expected to exist with a non-Gaussian fixed point, and this is indicated by the perturbative ϵ expansion of Wilson and Fisher in $4 - \epsilon$ dimensions.) However, the critical theory for $d = 3$ for $[\phi] = (3-\epsilon)/4$ for $\epsilon > 0$ held very small has been rigorously constructed by Brydges *et al.* (2003). This theory has a nontrivial hyperbolic fixed point of $O(\epsilon)$. The stable manifold is constructed in a small neighborhood of the fixed point. Note that the covariance without cutoff is Osterwalder–Schrader positive and thus this is a candidate for a nontrivial EQFT. For $\epsilon = 1$ we have the standard situation in $d = 3$, and

this remains open, as mentioned earlier. A very simplified picture of the above is furnished by the perturbative computation in the next section.

Unstable Fixed Points

We may attempt to construct field theories around unstable fixed points. In this case the initial parameters have to be adjusted as functions of the cutoff in such a way as to stabilize the flow in the neighborhood of the fixed point. This may be called a genuine renormalization. A famous example of this is pure Yang–Mills theory in $d=4$, where the Gaussian fixed point has only marginal unstable directions. Bałaban in a series of papers ending in Bałaban (1989) considered Wilson's lattice cutoff version of Yang–Mills theory in $d=4$ with initial coupling fixed by the two-loop asymptotic freedom formula. He proved, by lattice RG iterations, that in the weak-coupling regime the free energy per unit volume is bounded above and below by constants independent of the lattice spacing. Instability of the flow is expected to lead to mass generation for observables but this is a famous open problem. Another example is the standard nonlinear sigma model for $d=2$. Here too the flow is unstable around the Gaussian fixed point and we can set the initial coupling constant by the two-loop asymptotic freedom formula. Although much is known via approximation methods (as well as by methods based on integrable systems) this theory remains to be rigorously constructed as an EQFT.

Let us now consider a relatively simpler example, that of constructing a massive super-renormalizable scalar field theory. This has been studied in $d=3$, with $[\phi]=(d-2)/2=1/2$. We get $\xi=\tilde{\xi}, g=L^{-N}\tilde{g}, \mu=L^{-2N}\tilde{\mu}$, and $\tilde{g}$ is taken to be small. ξ is marginal, whereas g,μ are relevant parameters and grow with the iterations. After N iterations, they are brought up to $\tilde{g},\tilde{\mu}$ together with remainders. This realizes the so-called massive continuum ϕ^4 theory in $d=3$, and this has been mathematically controlled in the exact RG framework. This was proved by Brydges, Dimock, and Hurd and earlier by Benfatto, Cassandro, Gallavotti, and others, (see the references in Brydges *et al.* (1998) and Benfatto and Gallavotti (1995)).

The Exact RG as a Continuous Semigroup

The discrete semigroup defined in [13] of the previous section has a natural continuous counterpart. Just take L to be a continuous parameter, $L=e^t, t\geq 0$, and write by abuse of notation T_t, S_t, Γ_t instead of T_{e^t}, etc. The continuous transformations T_t,

$$T_t F = S_t \mu_{\Gamma_t} * F \qquad [32]$$

give a semigroup

$$T_t T_s = T_{t+s} \qquad [33]$$

of contractions on $L^2(d\mu_C)$ with μ_C as invariant measure. One can show that T_t is strongly continuous and, therefore, has a generator which we will call $\mathcal{L}$. This is defined by

$$\mathcal{L}F = \lim_{t\to 0^+} \frac{T_t - 1}{t} F \qquad [34]$$

whenever this limit exists. This restricts F to a suitable subspace $\mathcal{D}(\mathcal{L}) \subset L^2(d\mu_C)$. $\mathcal{D}(\mathcal{L})$ contains, for example, polynomials in fields as well as twice-differentiable bounded cylindrical functions. The generator $\mathcal{L}$ can be easily computed. To state it, we need some definitions. Define $(D^n F)(\phi; f_1, \ldots, f_n)$ as the nth tangent map at ϕ along directions $f_1, \ldots, f_n$. The functional Laplacian $\Delta_{\dot{\Gamma}}$ is defined by

$$\Delta_{\dot{\Gamma}} F(\phi) = \int d\mu_{\dot{\Gamma}}(\zeta)(D^2 F)(\phi; \zeta, \zeta) \qquad [35]$$

where $\dot{\Gamma}=u$. Define an infinitesimal dilatation operator

$$\mathcal{D}\phi(x) = x \cdot \nabla\phi(x) \qquad [36]$$

and a vector field $\mathcal{X}$,

$$\mathcal{X}F = -[\phi](DF)(\phi;\phi) - (DF)(\phi; \mathcal{D}\phi) \qquad [37]$$

Then, an easy computation gives

$$\mathcal{L} = \tfrac{1}{2}\Delta_{\dot{\Gamma}} + \mathcal{X} \qquad [38]$$

T_t is a semigroup with $\mathcal{L}$ as generator. Therefore, $T_t = e^{t\mathcal{L}}$. Let $F_t(\phi) = T_t F(\phi)$. Then F_t satisfies the linear PDE

$$\frac{\partial F_t}{\partial t} = \mathcal{L} F_t \qquad [39]$$

with the initial condition $F_0 = F$. This evolution equation assumes a more familiar form if we write $F_t = e^{-V_t}$, V_t being known as the effective potential. We get

$$\frac{\partial V_t}{\partial t} = \mathcal{L} V_t - \frac{1}{2}(V_t)_\phi \cdot (V_t)_\phi \qquad [40]$$

where

$$(V_t(\phi))_\phi \cdot (V_t(\phi))_\phi = \int d\mu_{\dot{\Gamma}}(\zeta)((DV_t)(\phi;\zeta))^2 \qquad [41]$$

and $V_0 = V$. This infinite-dimensional nonlinear PDE is a version of Wilson's flow equation. Note that the linear semigroup T_t acting on

functions induces a semigroup $\mathcal{R}_t$ acting nonlinearly on effective potentials giving a trajectory $V_t = \mathcal{R}_t V_0$.

Equations like the above are notoriously difficult to control rigorously, especially for large times. However, they may be solved in formal perturbation theory when the initial V_0 is small via the presence of small parameters. In particular, they give rise easily to perturbative flow equations for coupling constants. They can be obtained to any order but then there is the remainder. It is hard to control the remainder from the flow equation for effective potentials in bosonic field theories. They require other methods based on the discrete RG. Nevertheless, these approximate perturbative flows are very useful for getting a preliminary view of the flow. Moreover, their discrete versions figure as an input in further nonperturbative analysis.

Perturbative Flow

It is instructive to see this in second-order perturbation theory. We will simplify by working in infinite volume (no infrared divergences can arise because $\dot\Gamma(x-y)$ is of fast decrease). Now suppose that we are in standard ϕ^4 theory with $[\phi] = (d-2)/2$ and $d > 2$. We want to show that

$$V_t = \int \mathrm{d}x \Big(\xi_t : |\nabla\phi(x)|^2 : + g_t : \phi(x)^4 : + \mu_t : \phi(x)^2 : \Big) \qquad [42]$$

satisfies the flow equation in second order modulo irrelevant terms provided the parameters flow correctly. We will ignore field-independent terms. The Wick ordering is with respect to the covariance C of the invariant measure. The reader will notice that we have ignored a ϕ^6 term which is actually relevant in $d=3$ for the above choice of $[\phi]$. This is because we will only discuss the $d=3$ case for the model discussed at the end of this section and for this case the ϕ^6 term is irrelevant. We will assume that ξ_t, μ_t are of order $O(g^2)$. Plug in the above in the flow equation. The quantity $\lambda_t^{n,m} : P_{n,m} :$ represents one of the terms above with m fields and n derivatives. Because $\mathcal{L}$ is the generator of the semigroup T_t we have

$$\left(\frac{\partial}{\partial t} - \mathcal{L}\right) \lambda_t^{n,m} : P_{n,m} : = \left(\frac{\mathrm{d}\lambda_t^{n,m}}{\mathrm{d}t} - (d - m[\phi] - n)\lambda_t^{n,m} \right) : P_{n,m} : \qquad [43]$$

Next turn to the nonlinear term in the flow equation and insert the ϕ^4 term (the others are already of order $O(g^2)$). This produces a double integral of $\dot\Gamma(x-y) : \phi(x)^3 :: \phi(y)^3 :$, which after complete Wick ordering, gives

$$-\frac{g_t^2}{2} 16 \int \mathrm{d}x\, \mathrm{d}y\, \dot\Gamma(x-y) \Big(: \phi(x)^3 \phi(y)^3 : + 9C(x-y) : \phi(x)^2 \phi(y)^2 : + 36 C(x-y)^2 : \phi(x)\phi(y) : + 6C(x-y)^2 \Big) \qquad [44]$$

Consider the nonlocal ϕ^4 term. We can localize it by writing

$$: \phi(x)^2 \phi(y)^2 : = \tfrac{1}{2} : \Big(\phi(x)^4 + \phi(y)^4 - (\phi(x)^2 - \phi(y)^2)^2 \Big) : \qquad [45]$$

The local part gives a ϕ^4 contribution and the last term above gives rise to an irrelevant contribution because it produces additional derivatives. The coefficients are well defined because $C, \dot\Gamma$ are smooth and $\dot\Gamma(x-y)$ is of fast decrease. Now the nonlocal ϕ^2 term is similarly localized. It gives a relevant local ϕ^2 contribution as well as a marginal $|\nabla\phi|^2$ contribution. Finally, the same principle applies to the nonlocal ϕ^6 contribution and gives rise to further irrelevant terms. Then it is easy to see by matching that the flow equation is satisfied in second order up to irrelevant terms (these would have to be compensated by adding additional terms in V_t) provided

$$\begin{aligned} \frac{\mathrm{d}g_t}{\mathrm{d}t} &= (4-d)g_t - a g_t^2 + O(g_t^3) \\ \frac{\mathrm{d}\mu_t}{\mathrm{d}t} &= 2\mu_t - b g_t^2 + O(g_t^3) \\ \frac{\mathrm{d}\xi_t}{\mathrm{d}t} &= c g_t^2 + O(g_t^3) \end{aligned} \qquad [46]$$

where a, b, c are positive constants. We see from the above formulas that, up to second order in g^2, as $t \to \infty, g_t \to 0$ for $d \geq 4$. In fact, for $d \geq 5$ the decay rate is $O(\mathrm{e}^{-t})$ and for $d=4$ the rate is $O(t^{-1})$. However, to see if V_t converges, we also have to discuss the μ_t, ξ_t flows. It is clear that in general the μ_t flow will diverge. This is fixed by choosing the initial μ_0 to be the bare critical mass. This is obtained by integrating up to time t and then expressing μ_0 as a function of the entire g trajectory up to time t. Assume that μ_t is uniformly bounded and take $t \to \infty$. This gives the critical mass as

$$\mu_0 = b \int_0^\infty \mathrm{d}s\, \mathrm{e}^{-2s} g_s^2 = \mu_c(g_0) \qquad [47]$$

This integral converges for all cases discussed above. With this choice of μ_0 we get

$$\mu_t = b \int_0^\infty \mathrm{d}s\, \mathrm{e}^{-2s} g_{s+t}^2 \qquad [48]$$

and this exists for all t and converges as $t \to \infty$. Now consider the perturbative ξ flow. It is easy to see from the above that for $d \geq 4, \xi_t$ converges as $t \to \infty$.

We have not discussed the $d = 3$ case because the perturbative g fixed point is of order $O(1)$. But suppose we take, in the $d = 3$ case, $[\phi] = (3 - \epsilon)/4$ with $\epsilon > 0$ held small as in Brydges *et al.* (2003). Then the above perturbative flow equations are easily modified (by taking account of [43]) and we get, to second order, an attractive fixed point $g_* = O(\epsilon)$ of the g flow. The critical bare mass μ_0 can be determined as before and the ξ_t flow converges. The qualitative picture obtained above has a rigorous justification.

Rigorous RG Analysis

We will give a brief introduction to rigorous RG analysis in the discrete setup in the section "The RG as a Discrete Semigroup" concentrating on the principal problems encountered and how one attempts to solve them. Our approach is borrowed from Brydges *et al.* (2003). It is a simplification of the methods initiated by Brydges and Yau in (1990) and developed further by Brydges *et al.* (1998). The reader will find other approaches to rigorous RG methods in the selected references, such as those of Bałaban, Gawedzki and Kupiainen, Gallavotti, and others. We will take as a concrete example the scalar field model introduced earlier.

At the core of the analysis is the choice of good coordinates for the partition function density, z, of the section "The RG as a Discrete Semigroup". This is provided by a polymer representation (defined below) which parametrizes z by a couple (V, K), where V is a local potential and K is a set function also depending on the fields. Then the RG transformation T_L maps (V, K) to a new (V, K). (V, K) remain good coordinates as the volume tends to ∞, whereas z(volume) diverges. There exist norms which are suited to the fixed-point analysis of (V, K) to new (V, K). Now comes the important point: z does not uniquely specify the representation (V, K). Therefore, we can take advantage of this nonuniqueness to keep K small in norm and let most of the action of T_L reside in V. This process is called extraction in Brydges *et al.* (2003). It makes sure that K is an irrelevant term, whereas the local flow of V gives rise to discrete flow equations in coupling constants. We will not discuss extraction any further. In the following, we introduce the polymer representation and explain how the RG transformation acts on it.

To proceed further, we first introduce a simplification in the setup used in the section "The RG as a Discrete Semigroup." Recall that the function u introduced in [3] was smooth, positive definite, and of rapid decrease. We will simplify further by imposing the stronger property that it is actually of finite range: $u(x) = 0$ for $|x| \geq 1$. We say that u is of finite range 1. It is easy to construct such functions. For example, if g is any smooth function of finite range 1/2, then $u = g * g$ is a smooth positive-definite function of finite range 1. This implies that the fluctuation covariance Γ_L of [7] has finite range L. As a result, Γ_n in [10] has finite range L^{n+1} and the corresponding fluctuation fields $\zeta_n(x)$ and $\zeta_n(y)$ are independent when $|x - y| \geq L^{n+1}$.

Polymer Representation

Pave $\mathbf{R}^d$ with closed cubes of side length 1 called 1-blocks or unit blocks denoted by Δ, and suppose that Λ is a large cube consisting of unit blocks. A connected polymer $X \subset \Lambda$ is a closed connected subset of these unit blocks. A polymer activity $K(X, \phi)$ is a map $X, \phi \to \mathbf{R}$ where the fields ϕ depend only on the points of X. We will set $K(X, \phi) = 0$ if X is not connected. A generic form of the partition function density $z(\Lambda, \phi)$ after a certain number of RG iterations is

$$z(\Lambda) = \sum_{N=0}^{\infty} \frac{1}{N!} \sum_{X_1, \ldots, X_N} e^{-V(X_c)} \prod_{j=1}^{N} K(X_j) \qquad [49]$$

Here $X_j \subset \Lambda$ are disjoint polymers, $X = \bigcup X_j$, and $X_c = \Lambda \setminus X$. V is a local potential of the form [23] with parameters ξ, g, μ. We have suppressed the ϕ-dependence. Initially, the activities $K_j = 0$, but they will arise under RG iterations and the form [49] remains stable, as we will see. The partition function density is thus parametrized as a couple (V, K).

Norms for Polymer Activities

Polymer activities $K(X, \phi)$ are endowed with a norm $\|K(X)\|$, which must satisfy two properties:

$$\dot{X} \cap \dot{Y} = 0 \Rightarrow \|K_1(X)K_2(Y)\| \leq \|K_1(X)\| \, \|K_1(Y)\|$$
$$\|T_L K(X)\| \leq c^{|X|} \|K(X)\| \qquad [50]$$

where $\dot{X}$ is the interior of X and $|X|$ is the number of blocks in X. c is a constant of order $O(1)$. The norm measures (Fréchét) differentiability properties of the activity $K(X, \phi)$ with respect to the field ϕ as well as its admissible growth in ϕ. The growth is admissible if it is μ_C integrable. The second property above ensures the stability of the norm under RG iteration. For a fixed polymer X, the norm is such that it gives rise to a Banach

space of activities $K(X)$. The final norm $\|\cdot\|_{\mathcal{A}}$ incorporates the previous one and washes out the set dependence,

$$\|K\|_{\mathcal{A}} = \sup_{\Delta} \sum_{X \supset \Delta} A(X)\, \|K(X)\| \qquad [51]$$

where $A(X) = L^{(d+2)|X|}$. This norm essentially ensures that large polymers have small activities. The details of the above norms can be found in Brydges *et al.* (2003).

The RG operation map f is a composition of two maps. The RG iteration map $z \to T_L z$ induces a map $V \to \tilde{V}'_L$ and a nonlinear map $\tilde{T}_L : K \to \tilde{K} = \tilde{T}_L(K)$. We then compose this with a (nonlinear) extraction map $\mathcal{E}$ which takes out the expanding (relevant) parts of $\tilde{K} \to \mathcal{E}(\tilde{K}) = K'$ and compensates the local potential $\tilde{V}'_L \to V'$ such that $T_L z$ remains invariant. We denote by f the composition of these two maps with

$$V \to V' = f_V(V, K), \quad K \to K' = f_K(V, K) \qquad [52]$$

The Map $\tilde{T}_L$

Consider applying the RG map T_L to [49]. The map consists of a convolution $\mu_{\Gamma_L} *$ followed by the rescaling S_L. In the integration over the fluctuation field ζ, we will exploit the independence of $\zeta(x)$ and $\zeta(y)$ when $|x - y| \geq L$. To do this, we pave Λ by closed blocks of side L, called L-blocks, so that each L-block is a union of 1-blocks. Let $\bar{X}^L$ be the L-closure of a set X, namely the smallest union of L-blocks containing X. The polymers will be combined into L-polymers which are, by definition, connected unions of L blocks. The combination is performed in such a way that the new polymers are associated to independent functionals of ζ.

Let $\tilde{V}(X, \phi))$, to be chosen later, be a local potential independent of ζ. For a coupling constant sufficiently small, there is a bound

$$\|e^{-V(Y)}\| \leq 2^{|Y|} \qquad [53]$$

We assume that $\tilde{V}$ is so chosen that the same bound holds when V is replaced by $\tilde{V}$. Define

$$P(\Delta, \zeta, \phi) = e^{-V(\Delta, \zeta+\phi)} - e^{-\tilde{V}(\Delta, \phi)} \qquad [54]$$

Then we have

$$e^{-V(X_c, \zeta+\phi)} = e^{-V(\bar{X}_c, \zeta+\phi)} = \prod_{\Delta \subset \bar{X}_c} (e^{-\tilde{V}(\Delta, \phi)} + P(\Delta, \zeta, \phi)) \qquad [55]$$

where $\bar{X}_c$ is the closure of X_c. Expand out the product and insert into the representation [49] for $z(\Lambda, \zeta + \phi)$). We then rewrite the resulting sum in terms of L-polymers. The sum splits into a sum over connected components. Define, for every connected L-polymer Y,

$$\mathcal{B}K(Y) = \sum_{N+M \geq 1} \frac{1}{N!M!} \times \sum_{(X_j), (\Delta_i) \to Y} e^{-\tilde{V}(X_0)} \prod_{j=1}^{N} K(X_j) \prod_{i=1}^{M} P(\Delta_i) \qquad [56]$$

where $X_0 = Y \setminus (\cup X_j) \cup (\cup \Delta_j)$ and the sum over the distinct Δ_i, and disjoint 1-polymers X_j is such that their L-closure is Y. Equation [49] now becomes

$$z(\Lambda) = \sum \frac{1}{N!} \sum_{Y_1, \ldots, Y_N} e^{-\tilde{V}(Y_c)} \prod_{j=1}^{N} \mathcal{B}K(Y_j) \qquad [57]$$

where the sum is over disjoint, connected closed L-polymers. We now perform the fluctuation integration over ζ followed by the rescaling. Now $\tilde{V}(Y_c)$ is independent of ζ. The ζ-integration sails through and then factorizes because the Y_j, being disjoint closed L-polymers, are separated from each other by a distance $\geq L$. The rescaling brings us back to 1-polymers and reduces the volume from Λ to $L^{-1}\Lambda$. Therefore,

$$z'(L^{-1}\Lambda) = T_L z(\Lambda) = \sum \frac{1}{N!} \sum_{X_1, \ldots, X_N} e^{-\tilde{V}_L(X_c)} \prod_{j=1}^{N} (T_L \mathcal{B}K)(X_j) \qquad [58]$$

where the sum is over disjoint 1-polymers, $X_c = L^{-1}\Lambda \setminus X$. By definition $\tilde{V}_L(\Delta) = S_L \tilde{V}(L\Delta)$ and $(T_L \mathcal{B}K)(Z) = S_L \mu_{\Gamma_L} * \mathcal{B}K(LZ)$. This shows that the representation [49] is stable under iteration and, furthermore, gives us the map

$$\begin{aligned} V &\to \tilde{V}_L \\ K &\to \tilde{K} = \tilde{T}_L(K) = T_L \mathcal{B}K \end{aligned} \qquad [59]$$

The norm boundedness of K implies that $\tilde{T}_L(K)$ is norm bounded. We see from the above that a variation in the choice of $\tilde{V}$ is reflected in the corresponding variation of $\tilde{K}$. The extraction map $\mathcal{E}$ now takes out from $\tilde{K}$ the expanding parts and then compensates it by a change of $\tilde{V}_L$ in such a way that the representation [58] is left invariant by the simultaneous replacement $\tilde{V}_L \to V', \tilde{K} \to K' = \mathcal{E}(\tilde{K})$. The extraction map is nonlinear. Its linearization is a subtraction operation and this dominates in norm the nonlinearities, (Brydges *et al.* 1998).

The map $V \to \tilde{V}_L \to V'$ leads to a discrete flow of the coupling constants in V. It is convenient to write $K = K^{\text{pert}} + R$, where R is the remainder. Then the coupling constant flow is a discrete version of the

continuous flows encountered in the last section, together with remainders which are controlled by the size of R. In addition, we have the flow of K. The discrete flow of the pair (coupling constants, K) can be studied in a Banach space norm. Once one proves that the nonlinear parts satisfy a Lipshitz property, the discrete flow can be analyzed by the methods of stable-manifold theory of dynamical systems in a Banach space context. The reader is referred to the article by Brydges *et al.* (2003) for details of the extraction map and the application of stable-manifold theory in the construction of a global RG trajectory.

Further Topics

Lattice RG Methods

Statistical mechanical systems are often defined on a lattice. Moreover, the lattice provides an ultraviolet cutoff for Euclidean field theory compatible with Osterwalder–Schrader positivity. The standard lattice RG is based on Kadanoff–Wilson block spins. Its mathematical theory and applications have been developed by Bałaban, and Gawedzki and Kupiainen (see Gawedzki and Kupiainen (1986) and references therein). This leads to multiscale decompositions of the Gaussian lattice field as a sum of independent fluctuation fields on increasing length scales. Brydges *et al.* (2004) have shown that standard Gaussian lattice fields have multiscale decompositions as a sum of independent fluctuation fields with the finite-range property introduced in the last section. This permits the development of rigorous lattice RG theory in the spirit of the continuum framework of the previous section.

Fermionic Field Theories

Field theories of interacting fermions are often simpler to handle than bosonic field theories. Because of statistics, fermion fields are bounded and perturbation series converges in finite volume in the presence of an ultraviolet cutoff. The notion of studying the RG flow at the level of effective potentials makes sense. At any given scale, there is always an ultraviolet cutoff and the fluctuation covariance being of fast decrease provides an infrared cutoff. This is illustrated by the work of Gawedzki and Kupiainen (1985), who gave a nonperturbative construction in the weak effective coupling regime of the RG trajectory for the Gross–Neveu model in two dimensions. This is an example of a model with an unstable Gaussian fixed point where the initial coupling has to be adjusted as a function of the ultraviolet cutoff consistent with ultraviolet asymptotic freedom so as to stabilize the flow.

Acknowledgment

The author is grateful to David Brydges for very helpful comments on a preliminary version of this article.

See also: Operator Product Expansion in Quantum Field Theory; Renormalization: Statistical Mechanics and Condensed Matter.

Further Reading

Bałaban T (1982) $(Higgs)_{2,3}$ quantum fields in a finite volume II, an upper bound. *Communications in Mathematical Physics* 86: 555–594.

Bałaban T (1989) Large field renormalization. 2: Localization, exponentiation and bounds for the R operation. *Communications in Mathematical Physics* 122: 355–392.

Benfatto G and Gallavotti G (1995) *Renormalization Group.* Princeton, NJ: Princeton University Press.

Brydges DC and Yau HT (1990) Grad ϕ perturbations of massless Gaussian fields. *Communication in Mathematical Physics* 129: 351–392.

Brydges D, Dimock J, and Hurd T (1998) Estimates on renormalization group transformations. *Canadian Journal of Mathematics* 50(4): 756–793.

Brydges DC, Mitter PK, and Scoppola B (2003) Critical $(\Phi^4)_{3,\epsilon}$. *Communications in Mathematical Physics* 240: 281–327.

Brydges D, Guadagni G, and Mitter PK (2004) Finite range decomposition of Gaussian processes. *Journal of Statistical Physics* 115: 415–449.

Gawedzki K and Kupiainen A (1984) Massless lattice ϕ^4 theory in four dimensions: a nonperturbative control of a renormalizable model. *Communications in Mathematical Physics* 99: 197–252.

Gawedzki K and Kupiainen A (1985) Gross–Neveu models through convergent perturbation expansions. *Communications in Mathematical Physics* 102: 1–30.

Gawedzki K and Kupiainen A (1986) Asymptotic freedom beyond perturbation theory. In: Osterwalder K and Stora R (eds.) *Critical Phenomena, Random Systems, Gauge Theories.* North-Holland: Les Houches.

Gelfand IM and Vilenkin NYa (1964) *Generalized Functions*, vol. 4. New York: Academic Press.

Kopper C and Muller V (2000) Renormalization proof for spontaneously broken Yang–Mills theory with flow equations. *Communications in Mathematical Physics* 209: 477–516.

Polchinski J (1984) Renormalization and effective Lagrangians. *Nuclear Physics B* 231: 269–295.

Rivasseau V (1991) *From Perturbative to Constructive Renormalization.* Princeton NJ: Princeton University Press.

Wilson KG (1983) Renormalization group and critical phenomena. *Reviews of Modern Physics 55*: 583–599.

Wilson KG and Kogut J (1974) Renormalization group and the ϵ expansion. *Physics Reports* 12C: 75–200.

Gauge Theories from Strings

P Di Vecchia, Nordita, Copenhagen, Denmark

© 2006 Elsevier Ltd. All rights reserved.

Introduction

One of the most exciting properties of string theory, which led ten years ago to the formulation of the M theory as the unique theory unifying all interactions, has been the discovery that type II theories, besides a perturbative spectrum consisting of closed-string excitations, contain also a nonperturbative one consisting of "solitonic" p-dimensional objects called Dp branes. They are characterized by two important properties. They are coupled to closed-string states as the graviton, the dilaton, and the R–R $(p+1)$-form potential, and are described by a classical solution of the low-energy string effective action. Their dynamics is, on the other hand, described by open strings having the endpoints attached to their world volume and therefore satisfying Dirichlet boundary conditions in the directions transverse to their world volume. This is the reason why they are called D (Dirichlet) branes. Since the lightest open-string excitation corresponds to a gauge field, they have a gauge theory living on their world volume. This twofold description of D-branes has opened the way to study both the perturbative and nonperturbative properties of the gauge theory living on their world volume from their dynamics in terms of closed strings. With the addition of the decoupling limit, these two properties have led to the Maldacena (1998) conjecture of the equivalence between the maximally supersymmetric and conformal $\mathcal{N}=4$ super Yang–Mills and type IIB string theory on $AdS_5 \times S^5$.

They have also been successfully applied to less supersymmetric and nonconformal gauge theories

Notes:

that live on the world volume of fractional and wrapped branes. For general reviews of various approaches see Bertolini *et al.* (2000), Herzog *et al.* (2001), Bertolini (2003), Bigazzi *et al.* (2002), and Di Vecchia and Liccardo (2003). Also in these cases, one has constructed a classical solution of the supergravity equations of motion corresponding to these more sophisticated branes. These equations contain not only the supergravity fields present in the bulk ten-dimensional action but also boundary terms corresponding to the location of the branes. It turns out that in general the classical solution develops a naked singularity of the repulson type at short distances from the branes. This means that at short distances, it does not provide a reliable description of the branes. In the case of $\mathcal{N}=2$ supersymmetry, this can be explicitly seen because of the appearance of an enhançon located at distances slightly higher than the naked singularity (Johnson *et al.* 2000). The enhançon radius corresponds, in supergravity, to the distance where a brane probe becomes tensionless, and, in the gauge theory living on the branes, to the dynamically generated scale Λ_{QCD}. Then, since short distances in supergravity correspond to large distances in the gauge theory, as implied by holography, the presence of the enhançon and of the naked singularity does not allow to get any information on the nonperturbative large-distance behavior of the gauge theory living on the D-branes. Above the radius of the enhançon, instead, the classical solution provides a good description of the branes and therefore it can be used to get information on the perturbative behavior of the gauge theory. This shows that, if we want to use the D-branes for studying the nonperturbative properties of the gauge theory living on their world volume, we must construct a classical solution that has no naked singularity at short distances in supergravity. We will see in a specific example that it will be possible to deform the classical solution, eliminating the naked singularity, and use it to describe nonperturbative properties as the gaugino condensate.

In this article, we review some of the results obtained by using fractional D3 branes of some orbifold and D5 branes wrapped on 2-cycles of some Calabi–Yau manifold. The analysis of the supersymmetric gauge theories living on the world volume of these D-branes will be based on the gauge/gravity relations that relate the gauge coupling constant and the θ-angle to the supergravity fields (see, e.g., reference Di Vecchia *et al.* (2005) for a derivation of them):

$$\frac{4\pi}{g_{YM}^2}=\frac{1}{g_s(2\pi\sqrt{\alpha'})^2}\int_{C_2} d^2\xi e^{-\phi}\sqrt{\det(G_{AB}+B_{AB})} \qquad [1]$$

and

$$\theta_{YM}=\frac{1}{2\pi\alpha' g_s}\int_{C_2}(C_2+C_0B_2) \qquad [2]$$

where C_2 is the 2-cycle where the branes are wrapped.

In the next section, we will describe the case of the fractional D3 branes of the orbifold C^2/Z_2 and show that the classical solution corresponding to a system of N D3 and M D7 branes reproduces the perturbative behavior of $\mathcal{N}=2$ super-QCD. Then, we will consider D5 branes wrapped on 2-cycles of a Calabi–Yau manifold described by the Maldacena–Núñez classical solution (Maldacena and Núñez 2001, Chamseddine and Volkov 1997) and show that in this case we are able to reproduce the phenomenon of gaugino condensate and to construct the complete β-function of $\mathcal{N}=1$ super Yang–Mills.

Fractional D3 Branes of the Orbifold C^2/Z_2 and $\mathcal{N}=2$ Super-QCD

In this section, we consider fractional D3 and D7 branes of the noncompact orbifold C^2/Z_2 in order to study the properties of $\mathcal{N}=2$ super-QCD. We group the coordinates of the directions $(x^4,\ldots,x^9)$ transverse to the world volume of the D3 brane where the gauge theory lives, into three complex quantities: $z_1=x^4+ix^5$, $z_2=x^6+ix^7$, $z_3=x^8+ix^9$. The nontrivial generator h of Z_2 acts as $z_2\to -z_2$, $z_3\to -z_3$, leaving z_1 invariant. This orbifold has one fixed point, located at $z_2=z_3=0$ and corresponding to a vanishing 2-cycle. Fractional D3 branes are D5 branes wrapped on the vanishing 2-cycle and therefore are, unlike bulk branes, stuck at the orbifold fixed point. By considering N fractional D3 and M fractional D7 branes of the orbifold C^2/Z_2, we are able to study $\mathcal{N}=2$ super-QCD with M hypermultiplets. In order to do that, we need to determine the classical solution corresponding to the previous brane configuration. For the case of the orbifold C^2/Z_2, the complete classical solution is found in Bertolini *et al.* (2002b); see also references therein and Bertolini *et al.* (2000) for a review on fractional branes. In the following, we write it explicitly for a system of N fractional D3 branes with their world volume along the directions x^0,x^1,x^2, and x^3 and M fractional D7 branes containing the D3 branes in their world volume and having the remaining four world-volume directions along the orbifolded ones. The metric,

the 5-form field strength, the axion, and the dilaton are given by

$$ds^2 = H^{-1/2}\eta_{\alpha\beta}\,dx^\alpha\,dx^\beta + H^{1/2}\left(\delta_{\ell m}\,dx^\ell\,dx^m + e^{-\phi}\delta_{ij}\,dx^i\,dx^j\right) \quad [3]$$

$$\tilde{F}_{(5)} = d\left(H^{-1}\,dx^0\wedge\cdots\wedge dx^3\right) + {}^*d\left(H^{-1}\,dx^0\wedge\cdots\wedge dx^3\right) \quad [4]$$

$$\tau \equiv C_0 + ie^{-\phi} = i\left(1 - \frac{Mg_s}{2\pi}\log\frac{z}{\epsilon}\right) \quad [5]$$
$$z \equiv x^4 + ix^5 = ye^{i\theta}$$

where the self-dual field strength $\tilde{F}_{(5)}$ is given in terms of the NS–NS and R–R 2-forms B_2 and C_2 and of the 4-form potential C_4 by $\tilde{F}_{(5)} = dC_4 + C_2 \wedge dB_2$. The warp factor H is a function of the coordinates $(x^4,\ldots,x^9)$ and ϵ is an infrared cutoff. We denote by α and β the four directions corresponding to the world volume of the fractional D3 brane, by ℓ and m those along the four orbifolded directions x^6, x^7, x^8, and x^9, and by i and j the directions x^4 and x^5 that are transverse to both the D3 and the D7 branes. The twisted fields are instead given by $B_2 = \omega_2 b$, $C_2 = \omega_2 c$ where ω_2 is the volume form of the vanishing 2-cycle and

$$be^{-\phi} = \frac{(2\pi\sqrt{\alpha'})^2}{2}\left[1 + \frac{2N-M}{\pi}g_s\log\frac{y}{\epsilon}\right] \quad [6]$$
$$c + C_0 b = -2\pi\alpha'\theta g_s(2N-M)$$

The expression of H (Kirsch and Vaman 2005) shows that the previous solution has a naked singularity of the repulson type at short distances. On the other hand, if we use a brane probe approaching from infinity the stack of branes, described by the previous classical solution, it can also be seen that the tension of the probe vanishes at a distance that is larger than that of the naked singularity. The point where the probe brane becomes tensionless is called "enhançon" (Johnson *et al.* 2000) and at this point the classical solution does not describe anymore the stack of fractional branes.

Let us now use the gauge/gravity relations given in the introduction, to determine the coupling constants of the world-volume theory from the supergravity solution. In the case of fractional D3 branes of the orbifold $\mathbb{C}^2/\mathbb{Z}_2$, that is characterized by one single vanishing 2-cycle $\mathcal{C}_2$, the gauge coupling constant given in eqn [1] reduces to

$$\frac{1}{g^2_{\rm YM}} = \frac{1}{4\pi g_s(2\pi\sqrt{\alpha'})^2}\int_{\mathcal{C}_2} e^{-\phi}B_2 \quad [7]$$

By inserting the classical solution in eqns [7] and [2], we get the following expressions for the gauge coupling constant and the $\theta_{\rm YM}$ angle (Bertolini *et al.* 2002b):

$$\frac{1}{g^2_{\rm YM}} = \frac{1}{8\pi g_s} + \frac{2N-M}{16\pi^2}\log\frac{y^2}{\epsilon^2} \quad [8]$$
$$\theta_{\rm YM} = -\theta(2N-M)$$

Notice that the gauge coupling constant appearing in the previous equation is the "bare" gauge coupling constant computed at the scale $m \sim y/\alpha'$, while the square of the bare gauge coupling constant computed at the cutoff $\Lambda \sim \epsilon/\alpha'$ is equal to $8\pi g_s$.

In the case of an $\mathcal{N}=2$ supersymmetric gauge theory, the gauge multiplet contains a complex scalar field Ψ that corresponds to the complex coordinate z transverse to both the world volume of the D3 brane and the four orbifolded directions: $\Psi \sim z/2\pi\alpha'$. This is another example of holographic identification between a quantity, Ψ, peculiar of the gauge theory living on the fractional D3 branes and another one, the coordinate z, peculiar of supergravity. It allows one to obtain the gauge theory anomalies from the supergravity background. In fact, since we know how the scale and U(1) transformations act on Ψ, from the previous gauge/gravity relation we can deduce how they act on z, namely

$$\Psi \to se^{2i\alpha}\Psi \Longleftrightarrow z \to se^{2i\alpha}z \Longrightarrow y \to sy \quad [9]$$
$$\theta \to \theta + 2\alpha$$

Those transformations do not leave invariant the supergravity background in eqn [6] and when we use them in eqns [7] and [2], they generate the anomalies of the gauge theory living on the fractional D3 branes. In fact, by acting with those transformations in eqns [8], we get

$$\frac{1}{g^2_{\rm YM}} \to \frac{1}{g^2_{\rm YM}} + \frac{2N-M}{8\pi^2}\log s \quad [10]$$
$$\theta_{\rm YM} \to \theta_{\rm YM} - 2\alpha(2N-M)$$

The first equation generates the β-function of $\mathcal{N}=2$ super-QCD with M hypermultiplets given by

$$\beta(g_{\rm YM}) = -\frac{2N-M}{16\pi^2}g^3_{\rm YM} \quad [11]$$

while the second one reproduces the chiral U(1) anomaly (Klebanov *et al.* 2002, Bertolini *et al.* 2002a). In particular, if we choose $\alpha = 2\pi/(2(2N-M))$, then $\theta_{\rm YM}$ is shifted by a factor 2π. But since $\theta_{\rm YM}$ is periodic of 2π, this means that the subgroup $Z_{2(2N-M)}$ is not anomalous in perfect agreement with the gauge theory results.

Wrapped D5 Branes and $\mathcal{N}=1$ Super Yang–Mills

In this section, we will consider the classical solution corresponding to N D5 branes wrapped on a 2-cycle of a noncompact Calabi–Yau space and we use it to study the properties of the gauge theory living on their world volume that can be shown to be $\mathcal{N}=1$ super Yang–Mills.

We start by writing the classical solution found in Maldacena and Núñez (2001) and Chamseddine and Volkov (1997). It has a nontrivial metric:

$$\begin{aligned} ds_{10}^2 = e^{\Phi}\left[dx_{1,3}^2 + \frac{e^{2h}}{\lambda^2}\left(d\tilde{\theta}^2 + \sin^2\tilde{\theta}\, d\tilde{\varphi}^2\right)\right] \\ + \frac{e^{\Phi}}{\lambda^2}\left[d\rho^2 + \sum_{a=1}^{3}(\sigma^a - \lambda A^a)^2\right] \end{aligned} \quad [12]$$

a 2-form R–R potential

$$\begin{aligned} C^{(2)} = \frac{1}{4\lambda^2}\Big[(\psi+\psi_0)\left(\sin\theta'\, d\theta'\wedge d\phi - \sin\tilde{\theta}\, d\tilde{\theta}\wedge d\tilde{\varphi}\right) \\ - \cos\theta'\,\cos\tilde{\theta}\, d\phi\wedge d\tilde{\varphi}\Big] \\ + \frac{a}{2\lambda^2}\left[d\tilde{\theta}\wedge\sigma^1 - \sin\tilde{\theta}\, d\tilde{\varphi}\wedge\sigma^2\right] \end{aligned} \quad [13]$$

and a dilaton

$$e^{2\Phi} = \frac{\sinh 2\rho}{2e^h} \quad [14]$$

where

$$\begin{aligned} e^{2h} &= \rho\coth 2\rho - \frac{\rho^2}{\sinh^2 2\rho} - \frac{1}{4} \\ e^{2k} &= e^h\,\frac{\sinh 2\rho}{2} \\ a &= \frac{2\rho}{\sinh 2\rho} \end{aligned} \quad [15]$$

and

$$\begin{aligned} A^1 &= -\frac{1}{2\lambda}a(r)\, d\tilde{\theta} \\ A^2 &= \frac{1}{2\lambda}a(r)\sin\tilde{\theta}\, d\tilde{\varphi} \\ A^3 &= -\frac{1}{2\lambda}\cos\tilde{\theta}\, d\tilde{\varphi} \end{aligned} \quad [16]$$

with $\rho\equiv\lambda r$ and $\lambda^{-2} = Ng_s\alpha'$. The left-invariant 1-forms of S^3 are

$$\begin{aligned} \sigma^1 &= \frac{1}{2}\left[\cos\psi\, d\theta' + \sin\theta'\sin\psi\, d\phi\right] \\ \sigma^2 &= -\frac{1}{2}\left[\sin\psi\, d\theta' - \sin\theta'\cos\psi\, d\phi\right] \\ \sigma^3 &= \frac{1}{2}\left[d\psi + \cos\theta'\, d\phi\right] \end{aligned} \quad [17]$$

with $0\le\theta'\le\pi, 0\le\phi\le 2\pi$, and $0\le\psi\le 4\pi$. The variables $\tilde{\theta}$ and $\tilde{\varphi}$ describe a two-dimensional sphere and vary in the range $0\le\tilde{\theta}\le\pi$ and $0\le\tilde{\varphi}\le 2\pi$. Before proceeding, here we want to stress the fact that the presence of the function $a(\rho)\neq 0$ makes the solution regular everywhere. This will allow us to use it later on to describe the nonperturbative gaugino condensate property of $\mathcal{N}=1$ super Yang–Mills.

We can now use the previous solution for computing the running coupling constant and the θ parameter of $\mathcal{N}=1$ super Yang–Mills (see Di Vecchia *et al.* (2002), Bertolini and Merlatti (2003), and Mück (2003) reviewed in Bertolini (2003), Di Vecchia and Liccardo (2003), and Imeroni (2003)). In order to do that, we have to fix the cycle on which to perform the integrals in eqns [1] and [2]. It turns out that this 2-cycle is specified by

$$\tilde{\theta} = \theta' \cdot \tilde{\varphi} = -\phi, \quad \psi = 0 \quad [18]$$

keeping ρ fixed. If we now compute the gauge couplings on the previous cycle with $B_2 = C_0 = 0$, we get

$$\frac{4\pi^2}{Ng_{\mathrm{YM}}^2} = \rho\coth 2\rho + \tfrac{1}{2}a(\rho)\cos\psi \quad [19]$$

and

$$\theta_{\mathrm{YM}} = \frac{1}{2\pi g_s\alpha'}\int_{S^2} C_2 = -N(\psi + a(\rho)\sin\psi + \psi_0) \quad [20]$$

where we have kept $\psi\neq 0$ for reasons that will become clear in a moment. Equation [19] shows that the coupling constant is running as a function of the distance ρ from the branes. In order to obtain the correct running of the gauge theory, we have to find a relation between ρ and the renormalization group scale μ. This can be obtained with the following considerations. If we look at the previous solution, it is easy to see that the metric in eqn [12] is invariant under the following transformations:

$$\begin{aligned} &\psi \to \psi + 2\pi \quad \text{if } a\neq 0 \\ &\psi \to \psi + 2\epsilon \quad \text{if } a = 0 \end{aligned} \quad [21]$$

where c is an arbitrary constant. On the other hand, C_2 is not invariant under the previous transformations, but its flux, that is exactly equal to θ_{YM} in eqn [20], changes by an integer multiple of 2π:

$$\begin{aligned} \theta_{\mathrm{YM}} = \frac{1}{2\pi\alpha' g_s}\int_{C_2} C_2 \to \theta_{\mathrm{YM}} \\ + \begin{cases} -2\pi N, & \text{if } a\neq 0 \\ -2N\epsilon, & \text{if } a = 0, \quad \epsilon = \frac{\pi k}{N} \end{cases} \end{aligned} \quad [22]$$

But since the physics does not change when $\theta_{\mathrm{YM}}\to\theta_{\mathrm{YM}} + 2\pi$, one gets that the transformation in eqn [22] is an invariance. Notice that also eqn [19] for

the gauge coupling constant is invariant under the transformation in eqn [21]. The previous considerations show that the classical solution and also the gauge couplings are invariant under the Z_2 transformation if $a \neq 0$, while this symmetry becomes Z_{2N} if a is taken to be zero. As a consequence, since in the ultraviolet $a(\rho)$ is exponentially small, we can neglect it and we have a Z_{2N} symmetry, while in the infrared we cannot neglect $a(\rho)$ anymore and we have only a Z_2 symmetry left. This fits very well with the fact that $\mathcal{N}=1$ super Yang–Mills has a nonzero gaugino condensate $<\lambda\lambda>$ that is responsible for the breaking of Z_{2N} into Z_2. Therefore, it is natural to identify the gaugino condensate precisely with the function $a(\rho) \neq 0$ that makes the classical solution regular also at short distances in supergravity (Di Vecchia *et al.* 2002, Apreda *et al.* 2002):

$$<\lambda\lambda> \sim \Lambda^3 = \mu^3 a(\rho) \qquad [23]$$

This provides the relation between the renormalization group scale μ and the supergravity spacetime parameter ρ. In the ultraviolet (large ρ) $a(\rho)$ is exponentially suppressed and in eqns [19] and [20] we can neglect it obtaining

$$\frac{4\pi^2}{Ng_{\mathrm{YM}}^2} = \rho \coth 2\rho \qquad [24]$$
$$\theta_{\mathrm{YM}} = -N(\psi + \psi_0)$$

The chiral anomaly can be obtained by performing the transformation $\psi \to \psi + 2\epsilon$ and getting

$$\theta_{\mathrm{YM}} \to \theta_{\mathrm{YM}} - 2N\epsilon \qquad [25]$$

This implies that the Z_{2N} transformations corresponding to $\epsilon = \pi k/N$ are symmetries because they shift θ_{YM} by multiples of 2π.

In general, however, eqns [19] and [20] are only invariant under the Z_2 subgroup of Z_{2N} corresponding to the transformation

$$\psi \to \psi + 2\pi \qquad [26]$$

that changes θ_{YM} in eqn [20] as follows:

$$\theta_{\mathrm{YM}} \to \theta_{\mathrm{YM}} - 2N\pi \qquad [27]$$

leaving invariant the gaugino condensate:

$$<\lambda^2> = \mu^3 \frac{16\pi^2}{3Ng_{\mathrm{YM}}^2} e^{-8\pi^2/Ng_{\mathrm{YM}}^2} e^{i\theta_{\mathrm{YM}}/N} \qquad [28]$$

Therefore, the chiral anomaly and the breaking of Z_{2N} to Z_2 are encoded in eqns [19] and [20]. Finally, if we put $\psi = 0$ in eqn [19], we get

$$\frac{4\pi^2}{Ng_{\mathrm{YM}}^2} = \rho \coth 2\rho - \tfrac{1}{2}a(\rho) = \rho \tanh \rho \qquad [29]$$

This equation taken together with eqn [23] allows us to determine the running coupling constant as a function of μ. From it, we get (Di Vecchia *et al.* 2002, Di Vecchia and Liccardo 2003) the Novikov–Shifman–Vainshtein–Zacharov (NSVZ) β-function plus nonperturbative corrections due to fractional instantons:

$$\beta(g_{\mathrm{YM}}) = -\frac{3Ng_{\mathrm{YM}}^3}{16\pi^2} \frac{1 + \frac{4\pi^2}{Ng_{\mathrm{YM}}^2} \sinh^{-2} \frac{4\pi^2}{Ng_{\mathrm{YM}}^2}}{1 - \frac{Ng_{\mathrm{YM}}^2}{8\pi^2} + \frac{1}{2}\sinh^{-2} \frac{4\pi^2}{Ng_{\mathrm{YM}}^2}} \qquad [30]$$

where in the ultraviolet we have approximated ρ with $4\pi^2/(Ng_{\mathrm{YM}}^2)\coth 4\pi^2/(Ng_{\mathrm{YM}}^2)$.

See also: AdS/CFT Correspondence; Anomalies; BF Theories; Brane Construction of Gauge Theories; Gauge Theory: Mathematical Applications; Noncommutative Geometry from Strings; Nonperturbative and Topological Aspects of Gauge Theory; Perturbation Theory and its Techniques; Seiberg–Witten Theory; Superstring Theories.

Further Reading

Apreda R, Bigazzi F, Cotrone AL, Petrini M, and Zaffaroni A (2002) Some comments on $\mathcal{N}=1$ gauge theories from wrapped branes. *Physics Letters B* 536: 161–168 (hep-th/0112236).

Bertolini M (2003) Four lectures on the gauge/gravity correspondence. *International Journal of Modern Physics* A18: 5647–5712 (hep-th/0303160).

Bertolini M, Di Vecchia P, Frau M, Lerda A, and Marotta R (2002a) More anomalies from fractional branes. *Physics Letters B* 540: 104–110 (hep-th/0202195).

Bertolini M, Di Vecchia P, Frau M, Lerda A, and Marotta R (2002b) $N=2$ gauge theories on systems of fractional D3/D7 branes. *Nuclear Physics B* 621: 157–178 (hep-th/0107057).

Bertolini M, Di Vecchia P, and Marotta R (2000) $N=2$ four-dimensional gauge theories from fractional branes. In: Olshanetsky M and Vainshtein A (eds.) *Multiple Facets of Quantization and Supersymmetry*, pp. 730–773. (hep-th/0112195). Singapore: World Scientific.

Bertolini M and Merlatti P (2003) A note on the dual of $\mathcal{N}=1$ super Yang–Mills theory. *Physics Letters B* 556: 80–86 (hep-th/0211142).

Bigazzi F, Cotrone AL, Petrini M, and Zaffaroni A (2002) Supergravity duals of supersymmetric four dimensional gauge theories. *Rivista del Nuovo Cimento* 25N12: 1–70 (hep-th/0303191).

Chamseddine AH and Volkov MS (1997) Non-abelian BPS monopoles in $\mathcal{N}=4$ gauged supergravity. *Physical Review Letters* 79: 3343–3346 (hep-th/9707176).

Di Vecchia P, Lerda A, and Merlatti P (2002) $\mathcal{N}=1$ and $\mathcal{N}=2$ super Yang–Mills theories from wrapped branes. *Nuclear Physics B* 646: 43–68 (hep-th/0205204).

Di Vecchia P and Liccardo A (2003) Gauge theories from D branes, Lectures given at the School "Frontiers in Number Theory, Physics and Geometry," Les Houches, hep-th/0307104.

Di Vecchia P, Liccardo A, Marotta R, and Pezzella F (2005) On the gauge/gravity correspondence and the open/closed string duality. *International Journal of Modern Physics A* 20: 4699–4796. hep-th/0503156.

Herzog CP, Klebanov IR, and Ouyang P (2001) D-branes on the conifold and $\mathcal{N}=1$ gauge/gravity dualities. Based on I.R.K.'s Lectures at the Les Houches Summer School Session 76, "Gravity, Gauge Theories, and Strings," hep-th/0205100.

Imeroni E (2003) *Studies of the Gauge/String Theory Correspondence*. Ph.D. thesis, hep-th/0312070.

Johnson CV, Peet AW, and Polchinski J (2000) Gauge theory and the excision of repulson singularities. *Physical Review D* 61: 086001 (hep-th/9911161).

Kirsch I and Vaman D (2005) The D3/D7 background and flavour dependence of Regge trajectories, hep-th/0505164.

Klebanov IR, Ouyang P, and Witten E (2002) A gravity dual of the chiral anomaly. *Physical Review D* 65: 105007 (hep-th/0202056).

Maldacena J (1998) The large N limit of superconformal field theories and supergravity. *Advances in Theoretical and Mathematical Physics* 2: 231–252 (hep-th/9711200).

Maldacena J and Núñez C (2001) Towards the large N limit of $N=1$ super Yang Mills. *Physical Review Letters* 86: 588–591 (hep-th/0008001).

Mück W (2003) Perturbative and nonperturbative aspects of pure $\mathcal{N}=1$ super Yang–Mills theory from wrapped branes. *Journal of High Energy Physics* 0302: 013 (hep-th/0301171).

Gauge Theory: Mathematical Applications

S K Donaldson, Imperial College, London, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

This article surveys some developments in pure mathematics which have, to varying degrees, grown out of the ideas of gauge theory in mathematical physics. The realization that the gauge fields of particle physics and the connections of differential geometry are one and the same has had wide-ranging consequences, at different levels. Most directly, it has led mathematicians to work on new kinds of questions, often shedding light later on well-established problems. Less directly, various fundamental ideas and techniques, notably the need to work with the infinite-dimensional gauge symmetry group, have found a place in the general world-view of many mathematicians, influencing developments in other fields. Still less directly, the work in this area – between geometry and mathematical physics – has been a prime example of the interaction between these fields which has been so fruitful since the 1970s.

The body of this article is divided into three sections: roughly corresponding to analysis, geometry, and topology. However, the different topics come together in many different ways: indeed the existence of these links between the topics is one of the most attractive features of the area.

Gauge Transformations

For a review of the usual foundational material on connections, curvature, and related differential geometric constructions, the reader is referred to standard texts. We will, however, briefly recall the notions of gauge transformations and gauge fixing. The simplest case is that of abelian gauge theory – connections on a U(1)-bundle, say over R^3. In that case the connection form, representing the connection in a local trivialization, is a pure imaginary 1-form A, which can also be identified with a vector field A. The curvature of the connection is the 2-form dA. Changing the local trivialization by a U(1)-valued function $g = e^{i\chi}$ changes the connection form to

$$\tilde{A} = A - dgg^{-1} = A - id\chi$$

The forms $A, \tilde{A}$ are two representations of the same geometric object: just as the same metric can be represented by different expressions in different coordinate systems. One may want to fix this choice of representation, usually by choosing A to satisfy the Coulomb gauge condition $d^*A = 0$ (equivalently $\mathrm{div}\, A = 0$), supplemented by appropriate boundary conditions. Here we are using the standard Euclidean metric on R^3. (Throughout this article we will work with positive-definite metrics, regardless of the fact that – at least at the classical level – the Lorentzian signature may have more obvious bearing on physics.) Arranging this choice of gauge involves solving a linear partial differential equation (PDE) for χ.

The case of a general structure group G is not much different. The connection form A now takes values in the Lie algebra of G and the curvature is given by the expression

$$F = dA + \tfrac{1}{2}[A, A]$$

The change of bundle trivialization is given by a G-valued function and the resulting change in the connection form is

$$\tilde{A} = gAg^{-1} - dgg^{-1}$$

(Our notation here assumes that G is a matrix group, but this is not important.) Again, we can seek to impose the Coulomb gauge condition $d^*A = 0$, but now we cannot linearize this equation as before.

We can carry the same ideas over to a global problem, working on a G-bundle P over a general

Notes:

Riemannian manifold M. The space of connections on P is an affine space $\mathcal{A}$: any two connections differ by a bundle-valued 1-form. Now the gauge group $\mathcal{G}$ of automorphisms of P acts on $\mathcal{A}$ and, again, two connections in the same orbit of this action represent essentially the same geometric object. Thus, in a sense we would really like to work on the quotient space $\mathcal{A}/\mathcal{G}$. Working locally in the space of connections, near to some A_0, this is quite straightforward. We represent the nearby connections as $A_0 + a$, where a satisfies the analog of the coulomb condition

$$\mathrm{d}_A^* a = 0$$

Under suitable hypotheses, this condition picks out a unique representative of each nearby orbit. However, this gauge-fixing condition need not single out a unique representative if we are far away from A_0: indeed, the space $\mathcal{A}/\mathcal{G}$ typically has, unlike $\mathcal{A}$, a complicated topology which means that it is impossible to find any such global gauge-fixing condition. As noted above, this is one of the distinctive features of gauge theory. The gauge group $\mathcal{G}$ is an infinite-dimensional group, but one of a comparatively straightforward kind – much less complicated than the diffeomorphism groups relevant in Riemannian geometry for example. One could argue that one of the most important influences of gauge theory has been to accustom mathematicians to working with infinite-dimensional symmetry groups in a comparatively simple setting.

Analysis and Variational Methods

The Yang–Mills Functional

A primary object brought to mathematicians attention by physics is the Yang–Mills functional

$$\mathrm{YM}(A) = \int_M |F_A|^2 \, \mathrm{d}\mu$$

Clearly, $\mathrm{YM}(A)$ is non-negative and vanishes if and only if the connection is flat: it is broadly analogous to functionals such as the area functional in minimal submanifold theory, or the energy functional for maps. As such, one can fit into a general framework associated with such functionals. The Euler–Lagrange equations are the Yang–Mills equations

$$\mathrm{d}_A^* F_A = 0$$

For any solution (a Yang–Mills connection), there is a "Jacobi operator" H_A such that the second variation is given by

$$\mathrm{YM}(A + ta) = \mathrm{YM}(A) + t^2 \langle H_A a, a \rangle + O(t^3)$$

The omnipresent phenomenon of gauge invariance means that Yang–Mills connections are never isolated, since we can always generate an infinite-dimensional family by gauge transformations. Thus, as explained in the last section, one imposes the gauge-fixing condition $\mathrm{d}_A^* a = 0$. Then the operator H_A can be written as

$$H_A a = \Delta_A a + [F_A, a]$$

where Δ_A is the bundle-valued "Hodge Laplacian" $\mathrm{d}_A d_A^* + \mathrm{d}_A^* d_A$ and the expression $[F_A, a]$ combines the bracket in the Lie algebra with the action of Λ^2 on Λ^1. This is a self-adjoint elliptic operator and, if M is compact, the span of the negative eigenspaces is finite dimensional, the dimension being defined to be the index of the Yang–Mills connection A.

In this general setting, a natural aspiration is to construct a "Morse theory" for the functional. Such a theory should relate the topology of the ambient space to the critical points and their indices. In the simplest case, one could hope to show that for any bundle P there is a Yang–Mills connection with index 0, giving a minimum of the functional. More generally, the relevant ambient space here is the quotient $\mathcal{A}/\mathcal{G}$ and one might hope that the rich topology of this is reflected in the solutions to the Yang–Mills equations.

Uhlenbeck's Theorem

The essential foundation needed to underpin such a "direct method" in the calculus of variations is an appropriate compactness theorem. Here the dimension of the base manifold M enters in a crucial way. Very roughly, when a connection is represented locally in a Coulomb gauge, the Yang–Mills action combines the L^2-norm of the derivative of the connection form A with the L^2-norm of the quadratic term $[A, A]$. The latter can be estimated by the L^4-norm of A. If $\dim M \leq 4$, then the Sobolev inequalities allow the L^4-norm of A to be controlled by the L^2-norm of its derivative, but this is definitely not true in higher dimensions. Thus, $\dim M = 4$ is the "critical dimension" for this variational problem. This is related to the fact that the Yang–Mills equations (and Yang–Mills functional) are conformally invariant in four dimensions. For any nontrivial Yang–Mills connection over the 4-sphere, one generates a one-parameter family of Yang–Mills connections, on which the functional takes the same value, by applying conformal transformations corresponding to dilations of $\boldsymbol{R}^4$. In such a family of connections the integrand $|F_A|^2$ – the "curvature density" – converges to a δ-function

at the origin. More generally, one can encounter sequences of connections over 4-manifolds for which YM is bounded but which do not converge, the Yang–Mills density converging to δ-functions. There is a detailed analogy with the theory of the harmonic maps energy functional, where the relevant critical dimension (for the domain of the map) is 2.

The result of Uhlenbeck (1982), which makes these ideas precise, considers connections over a ball $B^n \subset \boldsymbol{R}^n$. If the exponent $p \geq 2n$, then there are positive constants $\epsilon(p,n), C(p,n) > 0$ such that any connection with $\|F\|_{L^p(B^n)} \leq \epsilon$ can be represented in Coulomb gauge over the ball, by a connection form which satisfies the condition $d^*A = 0$, together with certain boundary conditions, and

$$\|A\|_{L^p_1} \leq C\|F\|_{L^p}$$

In this Coulomb gauge, the Yang–Mills equations are elliptic and it follows readily that, in this setting, if the connection A is Yang–Mills one can obtain estimates on all derivatives of A.

Instantons in Four Dimensions

This result of Uhlenbeck gives the analytical basis for the direct method of the calculus of variations for the Yang–Mills functional over base manifolds M of dimension ≤ 3. For example, any bundle over such a manifold must admit a Yang–Mills connection, minimizing the functional. Such a statement is definitely false in dimensions ≥ 5. For example, an early result of Bourguignon and Lawson (1981) and Simons asserts that there is no minimizing connection on any bundle over S^n for $n \geq 5$. The proof exploits the action of the conformal transformations of the sphere. In the critical dimension 4, the situation is much more complicated. In four dimensions, there are the renowned "instanton" solutions of the Yang–Mills equation. Recall that if M is an oriented 4-manifold the Hodge $*$-operation is an involution of $\Lambda^2 T^*M$ which decomposes the two forms into self-dual and anti-self-dual parts, $\Lambda^2 T^*M = \Lambda^+ \oplus \Lambda^-$. The curvature of a connection can then be written as

$$F_A = F_A^+ + F_A^-$$

and a connection is a self-dual (respectively anti-self-dual) instanton if F_A^- (respectively F_A^+) is 0. The Yang–Mills functional is

$$\mathrm{YM}(A) = \|F_A^+\|^2 + \|F_A^-\|^2$$

while the difference $\|F_A^+\|^2 - \|F_A^-\|^2$ is a topological invariant $\kappa(P)$ of the bundle P, obtained by evaluating a four-dimensional characteristic class on $[M]$. Depending on the sign of $\kappa(P)$, the self-dual or anti-self-dual connections (if any exist) minimize the Yang–Mills functional among all connections on P. These instanton solutions of the Yang–Mills equations are analogous to the holomorphic maps from a Riemann surface to a Kähler manifold, which minimize the harmonic maps energy functional in their homotopy class.

Moduli Spaces

The instanton solutions typically occur in "moduli spaces." To fix ideas, let us consider bundles with structure group SU(2), in which case $\kappa(P) = -8\pi^2 c_2(P)$. For each $k > 0$, we have a moduli space $\mathcal{M}_k$ of anti-self-dual instantons on a bundle $P_k \to M^4$, with $c_2(P_k) = k$. It is a manifold of dimension $8k - 3$. The general goal of the calculus of variations in this setting is to relate three things:

1. the topology of the space $\mathcal{A}/\mathcal{G}$ of equivalence classes of connections on P_k;
2. the topology of the moduli space $\mathcal{M}_k$ of instantons; and
3. the existence and indices of other, nonminimal, solutions to the Yang–Mills equations on P_k.

In this direction, a very influential conjecture was made by Atiyah and Jones (1978). They considered the case when $M = S^4$ and, to avoid certain technicalities, work with spaces of "framed" connections, dividing by the restricted group $\mathcal{G}_0$ of gauge transformations equal to the identity at infinity. Then, for any k, the quotient $\mathcal{A}/\mathcal{G}_0$ is homotopy equivalent to the third loop space $\Omega^3 S^3$ of based maps from the 3-sphere to itself. The corresponding "framed" moduli space $\tilde{\mathcal{M}}_k$ is a manifold of dimension $8k$ (a bundle over $\mathcal{M}_k$ with fiber SO(3)). Atiyah and Jones conjectured that the inclusion $\tilde{\mathcal{M}}_k \to \mathcal{A}/\mathcal{G}_0$ induces an isomorphism of homotopy groups π_l in a range of dimensions $l \leq l(k)$, where $l(k)$ increases with k. This would be consistent with what one might hope to prove by the calculus of variations if there were no other Yang–Mills solutions, or if the indices of such solutions increased with k.

The first result along these lines was due to Bourguignon and Lawson (1981), who showed that the instanton solutions are the only local minima of the Yang–Mills functional over the 4-sphere. Subsequently, Taubes (1983) showed that the index of an non-instanton Yang–Mills connection P_k is at least $k + 1$. Taubes' proof used ideas related to the action of the quaternions and the hyper-Kähler structure on the $\tilde{\mathcal{M}}_k$ (see the section on hyper-Kähler quotients). Contrary to some expectations, it was shown by

Sibner *et al.* (1989) that nonminimal solutions do exist; some later constructions were very explicit (Sadun and Segert 1992). Taubes' index bound gave ground for hope that an analytical proof of the Atiyah–Jones conjecture might be possible, but this is not at all straightforward. The problem is that in the critical dimension 4 a mini–max sequence for the Yang–Mills functional in a given homotopy class may diverge, with curvature densities converging to sums of δ-functions as outlined above. This is related to the fact that the $\tilde{\mathcal{M}}_k$ are not compact. In a series of papers culminating in a framework for Morse theory for Yang–Mills functional, Taubes (1998) succeeded in proving a partial version of the Atiyah–Jones conjecture, together with similar results for general base manifolds M^4. Taubes showed that, if the homotopy groups of the moduli spaces stabilize as $k \to \infty$, then the limit must be that predicted by Atiyah and Jones. Related analytical techniques were developed for other variational problems at the critical dimension involving "critical points at infinity." The full Atiyah–Jones conjecture was established by Boyer *et al.* but using geometrical techniques: the "explicit" description of the moduli spaces obtained from the Atiyah–Drinfeld–Hitchin–Manin (ADHM) construction (see below). A different geometrical proof was given by Kirwan (1994), together with generalizations to other gauge groups.

There was a parallel story for the solutions of the Bogomolony equation over R^3, which we will not recount in detail. Here the base dimension is below the critical case but the analytical difficulty arises from the noncompactness of R^3. Taubes succeeded in overcoming this difficulty and obtained relations between the topology of the moduli space, the appropriate configuration space and the higher critical points. Again, these higher critical points exist but their index grows with the numerical parameter corresponding to k. At about the same time, Donaldson (1984) showed that the moduli spaces could be identified with spaces of rational maps (subsequently extended to other gauge groups). The analog of the Atiyah–Jones conjecture is a result on the topology of spaces of rational maps proved earlier by Segal, which had been one of the motivations for Atiyah and Jones.

Higher Dimensions

While the scope for variational methods in Yang–Mills theory in higher dimensions is very limited, there are useful analytical results about solutions of the Yang–Mills equations. An important monotonicity result was obtained by Price (1983). For simplicity, consider a Yang–Mills connection over the unit ball $B^n \subset R^n$. Then Price showed that the normalized energy

$$E(A, B(r)) = \frac{1}{r^{n-4}} \int_{|x| \leq r} |F|^2 \, d\mu$$

decreases with r. Nakajima (1988) and Uhlenbeck used this monotonicity to show that for each n there is an ϵ such that if A is a Yang–Mills connection over a ball with $E(A, B(r)) \leq \epsilon$ then all derivatives of A, in a suitable gauge, can be controlled by $E(A, B(r))$. Tian (2000) showed that if A_i is a sequence of Yang–Mills connections over a compact manifold M with bounded Yang–Mills functional, then there is a subsequence which converges away from a set Z of Haussdorf codimension at least 4 (extending the case of points in a 4-manifold). Moreover, the singular set Z is a minimal subvariety, in a suitably generalized sense.

In higher dimensions, important examples of Yang–Mills connections arise within the framework of "calibrated geometry." Here, we consider a Riemannian n-manifold M with a covariant constant calibrating form $\Omega \in \Omega_M^{n-4}$. There is then an analog of the instanton equation

$$F_A = \pm * (\Omega \wedge F_A)$$

whose solutions minimize the Yang–Mills functional. This includes the Hermitian Yang–Mills equation over a Kähler manifold (see the section on moment maps) and also certain equations over manifolds with special holonomy groups (Donaldson and Thomas 1998). For these "higher-dimensional instantons," Tian shows that the singular sets Z that arise are calibrated varieties.

Gluing Techniques

Another set of ideas from PDEs and analysis which has had great impact in gauge theory involves the construction of solutions to appropriate equations by the following general scheme:

1. constructing an "approximate solution," formed from some standard models using cutoff functions;
2. showing that the approximate solution can be deformed to a true solution by means of an implicit function theorem.

The heart of the second step usually consists of estimates for the relevant linear differential operator. Of course, the success of this strategy depends on the particular features of the problem. This approach, due largely to Taubes, has been particularly effective in finding solutions to the first-order instanton equations and their relatives. (The applicability of the approach is connected to the fact that

such solutions typically occur in moduli spaces and one can often "see" local coordinates in the moduli space by varying the parameters in the approximate solution.) Taubes applied this approach to the Bogomolny monopole equation over $\boldsymbol{R}^3$ (Jaffe and Taubes 1980) and to construct instantons over general 4-manifolds (Taubes 1982). In the latter case, the approximate solutions are obtained by transplanting standard solutions over $\boldsymbol{R}^4$ – with curvature density concentrated in a small ball – to small balls on the 4-manifold, glued to the trivial flat connection over the remainder of the manifold. These types of techniques have now become a fairly standard part of the armory of many differential geometers, working both within gauge theory and other fields. An example of a problem where similar ideas have been used is Joyce's construction of constant of manifolds with exceptional holonomy groups (Joyce 1996). (Of course, it is likely that similar techniques have been developed over the years in many other areas, but Taubes' work in gauge theory has done a great deal to bring them into prominence.)

Geometry: Integrability and Moduli Spaces

The Ward Correspondence

Suppose that S is a complex surface and ω is the 2-form corresponding to a Hermitian metric on S. Then S is an oriented Riemannian 4-manifold and ω is a self-dual form. The orthogonal complement of ω in Λ^+ can be identified with the real parts of forms of type $(0,2)$. Hence, if A is an anti-self-dual instanton connection on a principle $\mathrm{U}(r)$-bundle over S the $(0,2)$ part of the curvature of A vanishes. This is the integrability condition for the $\bar{\partial}$-operator defined by the connection, acting on sections of the associated vector bundle $E \to S$. Thus, in the presence of the connection, the bundle E is naturally a holomorphic bundle over S.

The Ward correspondence (Ward 1877) builds on this idea to give a complete translation of the instanton equations over certain Riemannian 4-manifolds into holomorphic geometry. In the simplest case, let A be an instanton on a bundle over $\boldsymbol{R}^4$. Then, for any choice of a linear complex structure on $\boldsymbol{R}^4$, compatible with the metric, A defines a holomorphic structure. The choices of such a complex structure are parametrized by a 2-sphere; in fact, the unit sphere in $\Lambda^+(\boldsymbol{R}^4)$. So, for any $\lambda \in S^2$ we have a complex surface S_λ and a holomorphic bundle over S_λ. These data can be viewed in the following way. We consider the projection $\pi : \boldsymbol{R}^4 \times S^2 \to \boldsymbol{R}^4$ and the pull-back $\pi^*(E)$ to $\boldsymbol{R}^4 \times S^2$. This pullback bundle has a connection which defines a holomorphic structure along each fiber $S_\lambda \subset \boldsymbol{R}^4 \times S^2$ of the other projection. The product $\boldsymbol{R}^4 \times S^2$ is the twistor space of $\boldsymbol{R}^4$ and it is in a natural way a three-dimensional complex manifold. It can be identified with the complement of a line L_∞ in $\mathbf{CP}^3$ where the projection $\boldsymbol{R}^4 \times S^2 \to S^2$ becomes the fibration of $\mathbf{CP}^3 \setminus L_\infty$ by the complex planes through L_∞. One can see then that $\pi^*(E)$ is naturally a holomorphic bundle over $\mathbf{CP}^3 \setminus L_\infty$. The construction extends to the conformal compactification S^4 of $\boldsymbol{R}^4$. If S^4 is viewed as the quaternionic projective line $\mathbf{HP}^1$ and we identify H^2 with C^4 in the standard way, we get a natural map $\pi : \mathbf{CP}^3 \to \mathbf{HP}^1$. Then $\mathbf{CP}^3$ is the twistor space of S^4 and an anti-self-dual instanton on a bundle E over S^4 induces a holomorphic structure on the bundle $\pi^*(E)$ over $\mathbf{CP}^3$.

In general, the twistor space Z of an oriented Riemannian 4-manifold M is defined to be the unit sphere bundle in Λ^+_M. This has a natural almost-complex structure which is integrable if and only if the self-dual part of the Weyl curvature of M vanishes (Atiyah *et al.* 1978). The antipodal map on the 2-sphere induces an antiholomorphic involution of Z. In such a case, an anti-self-dual instanton over M lifts to a holomorphic bundle over Z. Conversely, a holomorphic bundle over Z which is holomorphically trivial over the fibers of the fibration $Z \to M$ (projective lines in Z), and which satisfies a certain reality condition with respect to the antipodal map, arises from a unitary instanton over M. This is the Ward correspondence, part of Penrose's twistor theory.

The ADHM Construction

The problem of describing all solutions to the Yang–Mills instanton equation over S^4 is thus reduced to a problem in algebraic geometry, of classifying certain holomorphic vector bundles. This was solved by Atiyah *et al.* (1978). The resulting ADHM construction reduces the problem to certain matrix equations. The equations can be reduced to the following form. For a bundle Chern class k and rank r, we require a pair of $k \times k$ matrices α_1, α_2, a $k \times r$ matrix a, and an $r \times k$ matrix b. Then the equations are

$$\begin{aligned} [\alpha_1, \alpha_2] &= ab \\ [\alpha_1^*, \alpha_1] + [\alpha_2^*, \alpha_2] &= aa^* - b^*b \end{aligned} \qquad [1]$$

We also require certain open, nondegeneracy conditions. Given such matrix data, a holomorphic

bundle over $\mathbf{CP}^3$ is constructed via a "monad": a pair of bundle maps over $\mathbf{CP}^3$

$$C^k \otimes \mathcal{O}(-1)@ > D_1 >> C^k \oplus C^k \oplus C^r@ > D_2 >> C^k \otimes \mathcal{O}(1)$$

with $D_2 D_1 = 0$. That is, the rank-r holomorphic bundle we construct is $\operatorname{Ker} D_2 / \operatorname{Im} D_1$. The bundle maps D_1, D_2 are obtained from the matrix data in a straightforward way, in suitable coordinates. It is this matrix description which was used by Boyer *et al.* to prove the Atiyah–Jones conjecture on the topology of the moduli spaces of instantons. The only other case when the twistor space of a compact 4-manifold is an algebraic variety is the complex projective plane, with the nonstandard orientation. An analog of the ADHM description in this case was given by Buchdahl (1986).

Integrable Systems

The Ward correspondence can be viewed in the general framework of integrable systems. Working with the standard complex structure on R^4, the integrability condition for the $\bar{\partial}$-operator takes the shape

$$[\nabla_1 + \mathrm{i}\nabla_2, \nabla_3 + \mathrm{i}\nabla_4] = 0$$

where ∇_i are the components of the covariant derivative in the coordinate directions. So, the instanton equation can be viewed as a family of such commutator equations parametrized by $\lambda \in S^2$. One obtains many reductions of the instanton equation by imposing suitable symmetries. Solutions invariant under translation in one variable correspond to the Bogomolny "monopole equation" (Jaffe and Taubes 1980). Solutions invariant under three translations correspond to solutions of Nahm's equations,

$$\frac{\mathrm{d}T_i}{\mathrm{d}t} = \epsilon_{ijk}[T_j, T_k]$$

for matrix-valued functions T_1, T_2, T_3 of one variable t. Nahm (1982) and Hitchin (1983) developed an analog of the ADHM construction relating these two equations. This is now seen as a part of a general "Fourier–Mukai–Nahm transform" (Donaldson and Kronheimer 1990). The instanton equations for connections invariant under two translations, Hitchin's equations (Hitchin 1983), are locally equivalent to the harmonic map equation for a surface into the symmetric space dual to the structure group. Changing the signature of the metric on R^4 to (2, 2), one gets the harmonic mapping equations into Lie groups (Hitchin 1990). More complicated reductions yield almost all the known examples of integrable PDEs as special forms of the instanton equations (Mason and Woodhouse 1996).

Moment Maps: the Kobayashi–Hitchin Conjecture

Let Σ be a compact Riemann surface. The Jacobian of Σ is the complex torus $H^1(\Sigma, \mathcal{O})/H^1(\Sigma, \mathbf{Z})$: it parametrizes holomorphic line bundles of degree 0 over Σ. The Hodge theory (which was, of course, developed long before Hodge in this case) shows that the Jacobian can also be identified with the torus $H^1(\Sigma, \boldsymbol{R})/H^1(\Sigma, \boldsymbol{Z})$ which parametrizes flat U(1)-connections. That is, any holomorphic line bundle of degree 0 admits a unique compatible flat unitary connection.

The generalization of these ideas to bundles of higher rank began with Weil. He observed that any holomorphic vector bundle of degree 0 admits a flat connection, not necessarily unitary. Narasimhan and Seshadri (1965) showed that (in the case of degree 0) the existence of a flat, irreducible, unitary connection was equivalent to an algebro-geometric condition of stability which had been introduced shortly before by Mumford, for quite different purposes. Mumford introduced the stability condition in order to construct separated moduli spaces of holomorphic bundles – generalizing the Jacobian – as part of his general geometric invariant theory. For bundles of nonzero degree, the discussion is slightly modified by the use of projectively flat unitary connections. The result of Narasimhan and Seshadri asserts that there are two different descriptions of the same moduli space $\mathcal{M}^{d,r}(\mathrm{Sigma})$: either as parametrizing certain irreducible projectively flat unitary connections (representations of $\pi_1(\Sigma)$), or parametrizing stable holomorphic bundles of degree d and rank r. While Narasimhan and Seshadri probably did not view the ideas in these terms, another formulation of their result is that a certain nonlinear PDE for a Hermitian metric on a holomorphic bundle – analogous to the Laplace equation in the abelian case – has a solution when the bundle is stable.

Atiyah and Bott (1982) cast these results in the framework of gauge theory. (The Yang–Mills equations in two dimensions essentially reduce to the condition that the connection be flat, so they are rather trivial locally but have interesting global structure.) They made the important observation that the curvature of a connection furnishes a map

$$F : \mathcal{A} \to \mathrm{Lie}(\mathcal{G})^*$$

which is an equivariant moment map for the action of the gauge group on $\mathcal{A}$. Here the symplectic form on the affine space $\mathcal{A}$ and the map from the adjoint

bundle-valued 2-forms to the dual of the Lie algebra of $\mathcal{G}$ are both given by integration of products of forms. From this point of view, the Narasimhan–Seshadri result is an infinite-dimensional example of a general principle relating symplectic and complex quotients. At about the same time, Hitchin and Kobayashi independently proposed an extension of these ideas to higher dimensions. Let E be a holomorphic bundle over a complex manifold V. Any compatible unitary connection on E has curvature F of type (1,1). Let ω be the (1,1)-form corresponding to a fixed Hermitian metric on V. The Hermitian Yang–Mills equation is the equation

$$F \cdot \omega = \mu 1_E$$

where μ is a constant (determined by the topological invariant $c_1(E)$). The Kobayashi–Hitchin conjecture is that, when ω is Kähler, this equation has an irreducible solution if and only if E is a stable bundle in the sense of Mumford. Just as in the Riemann surface case, this equation can be viewed as a nonlinear second-order PDE of Laplace type for a metric on E. The moment map picture of Atiyah and Bott also extends to this higher-dimensional version. In the case when V has complex dimension 2 (and μ is zero), the Hermitian Yang–Mills connections are exactly the anti-self-dual instantons, so the conjecture asserts that the moduli spaces of instantons can be identified with certain moduli spaces of stable holomorphic bundles.

The Kobayashi–Hitchin conjecture was proved in the most general form by Uhlenbeck and Yau (1986), and in the case of algebraic manifolds in Donaldson (1987). The proofs in Donaldson (1985, 1987) developed some extra structure surrounding these equations, connected with the moment map point of view. The equations can be obtained as the Euler–Lagrange equations for a nonlocal functional, related to the renormalized determinants of Quillen and Bismut. The results have been extended to non-Kähler manifolds and certain noncompact manifolds. There are also many extensions to equations for systems of data comprising a bundle with additional structure such as a holomorphic section or Higgs' field (Bradlow *et al.* 1995), or a parabolic structure along a divisor. Hitchin's equations (Hitchin 1987) are a particularly rich example.

Topology of Moduli Spaces

The moduli spaces $\mathcal{M}_{r,d}(\Sigma)$ of stable holomorphic bundles/projectively flat unitary connections over Riemann surfaces Σ have been studied intensively from many points of view. They have natural Kähler structures: the complex structure being visible in the holomorphic bundles guise and the symplectic form as the "Marsden–Weinstein quotient" in the unitary connections guise. In the case when r and d are coprime, they are compact manifolds with complicated topologies. There is an important basic construction for producing cohomology classes over these (and other) moduli spaces. One takes a universal bundle U over the product $\mathcal{M} \times \Sigma$ with Chern classes

$$c_i(U) \in H^{2i}(\mathcal{M} \times \Sigma)$$

Then, for any class $\alpha \in H_p(\Sigma)$, we get a cohomology class $c_i(U)/\alpha \in H^{2i-p}(\mathcal{M})$. Thus, if R_Σ is the graded ring freely generated by such classes, we have a homomorphism $\nu : R_\Sigma \to H^*(\mathcal{M})$. The questions about the topology of the moduli spaces which have been studied include:

1. finding the Betti numbers of the moduli space $\mathcal{M}$;
2. identifying the kernel of ν;
3. giving an explicit system of generators and relations for the ring $H^*(\mathcal{M})$;
4. identifying the Pontrayagin and Chern classes of $\mathcal{M}$ within $H^*(\mathcal{M})$; and
5. evaluating the pairings

$$\int_{\mathcal{M}} \nu(W)$$

for elements W of the appropriate degree in R.

All of these questions have now been solved quite satisfactorily. In early work, Newstead (1967) found the Betti numbers in the rank-2 case. The main aim of Atiyah and Bott was to apply the ideas of Morse theory to the Yang–Mills functional over a Riemann surface and they were able to reproduce Newstead's results in this way and extend them to higher rank. They also showed that the map ν is a surjection, so the universal bundle construction gives a system of generators for the cohomology. Newstead made conjectures on the vanishing of the Pontrayagin and Chern classes above a certain range which were established by Kirwan and extended to higher rank by Earl and Kirwan (1999). Knowing that R_Σ maps on to $H^*(M)$, a full set of relations can (by Poincaré duality) be deduced in principle from a knowledge of the integral pairings in (5) above, but this is not very explicit. A solution to (5) in the case of rank 2 was found by Thaddeus (1992). He used results from the Verlinde theory (see section on 3-manifolds below) and the Riemann–Roch formula. Another point of view was developed by Witten (1991), who showed that the volume of the moduli space was related to the theory of torsion in algebraic topology and satisfied simple gluing axioms. These different

points of view are compared in Donaldson (1993). Using a nonrigorous localization principle in infinite dimensions, Witten (1992) wrote down a general formula for the pairings (5) in any rank, and this was established rigorously by Jeffrey and Kirwan, using a finite-dimensional version of the same localization method. A very simple and explicit set of generators and relations for the cohomology (in the rank-2 case) was given by King and Newstead (1998). Finally, the quantum cohomology of the moduli space, in the rank-2 case, was identified explicitly by Munoz (1999).

Hyper-Kähler Quotients

Much of this story about the structure of moduli spaces extends to higher dimensions and to the moduli spaces of connections and Higgs fields. A particularly notable extension of the ideas involves hyper-Kähler structures. Let M be a hyper-Kähler 4-manifold, so there are three covariant-constant self-dual forms $\omega_1, \omega_2, \omega_3$ on M. These correspond to three complex structures I_1, I_2, I_3 obeying the algebra of the quaternions. If we single out one structure, say I_1, the instantons on M can be viewed as holomorphic bundles with respect to I_1 satisfying the moment map condition (Hermitian Yang–Mills equation) defined by the form ω_1. Taking a different complex structure interchanges the role of the moment map and integrability conditions. This can be put in a general framework of hyper-Kähler quotients due to Hitchin *et al.* (1987). Suppose initially that M is compact (so either a K3 surface or a torus). Then the ω_i components of the curvature define three maps

$$F_i : \mathcal{A} \to \mathrm{Lie}(\mathcal{G})^*$$

The structures on M make $\mathcal{A}$ into a flat hyper-Kähler manifold and the three maps F_i are the moment maps for the gauge group action with respect to the three symplectic forms on $\mathcal{A}$. In this situation, it is a general fact that the hyper-Kähler quotient – the quotient by $\mathcal{G}$ of the common zero set of the three moment maps – has a natural hyper-Kähler structure. This hyper-Kähler quotient is just the moduli space of instantons over M. In the case when M is the noncompact manifold $\boldsymbol{R}^4$, the same ideas apply except that one has to work with the based gauge group $\mathcal{G}_0$. The conclusion is that the framed moduli spaces $\tilde{\mathcal{M}}$ of instantons over $\boldsymbol{R}^4$ are naturally hyper-Kähler manifolds. One can also see this hyper-Kähler structure through the ADHM matrix description. A variant of these matrix equations was used by Kronheimer to construct "gravitational instantons." The same ideas also apply to the moduli spaces of monopoles, where the hyper-Kähler metric, in the simplest case, was studied by Atiyah and Hitchin (1989).

Low-Dimensional Topology

Instantons and 4-Manifolds

Gauge theory has had unexpected applications in low-dimensional topology, particularly the topology of smooth 4-manifolds. The first work in this direction, in the early 1980s, involved the Yang–Mills instantons. The main issue in 4-manifold theory at that time was the correspondence between the diffeomorphism classification of simply connected 4-manifolds and the classification up to homotopy. The latter is determined by the intersection form, a unimodular quadratic form on the second integral homology group (i.e., a symmetric matrix with integral entries and determinant ± 1, determined up to integral change of basis). The only known restriction was that Rohlin's theorem, which asserts that if the form is even the signature must be divisible by 16. The achievement of the first phase of the theory was to show that

1. There are unimodular forms which satisfy the hypotheses of Rohlin's theorem but which do not appear as the intersection forms of smooth 4-manifolds. In fact, no nonstandard definite form, such as a sum of copies of the E_8 matrix, can arise in this way.
2. There are simply connected smooth 4-manifolds which have isomorphic intersection forms, and hence are homotopy equivalent, but which are not diffeomorphic.

These results stand in contrast to the homeomorphism classification which was obtained by Freedman shortly before and which is almost the same as the homotopy classification.

The original proof of item (1) above argued with the moduli space $\mathcal{M}$ of anti-self-dual instantons SU(2) instantons on a bundle with $c_2 = 1$ over a simply connected Riemannian 4-manifold M with a negative-definite intersection form (Donaldson 1983). In the model case when M is the 4-sphere the moduli space $\mathcal{M}$ can be identified explicitly with the open 5-ball. Thus the 4-sphere arises as the natural boundary of the moduli space. A sequence of points in the moduli space converging to a boundary point corresponds to a sequence of connections with curvature densities converging to a δ-function, as described earlier. One shows that in the general case (under our hypotheses on the 4-manifold M) the moduli space $\mathcal{M}$ has a similar behavior, it contains a collar $M \times (0, \delta)$

formed by instantons made using Taubes' gluing construction, described previously. The complement of this collar is compact. In the interior of the moduli space, there are a finite number of special points corresponding to U(1)-reductions of the bundle P. This is the way in which the moduli space "sees" the integral structure of the intersection form since such reductions correspond to integral homology classes with self-intersection -1. Neighborhoods of these special points are modeled on quotients $\mathbf{C}^3/\mathrm{U}(1)$; that is, cones on copies of $\mathbf{CP}^2$. The upshot is that (for generic Riemannian metrics on M) the moduli space gives a cobordism from the manifold M to a set of copies of $\mathbf{CP}^2$ which can be counted in terms of the intersection form, and the result follows easily from standard topology. More sophisticated versions of the argument extended the results to rule out some indefinite intersection forms.

On the other hand, the original proofs of item (2) used "invariants" defined by instanton moduli spaces (Donaldson 1990). The general scheme exploits the same construction outlined in the previous section. We suppose that M is a simply connected 4-manifold with $b^+(M) = 1 + 2p$, where $p > 0$ is an integer. (Here $b^+(M)$ is, as usual, the number of positive eigenvalues of the intersection matrix.) Ignoring some technical restrictions, there is a map

$$\nu : R_M \to H^*(\mathcal{M}_k)$$

where R_M is a graded ring freely generated by the homology (below the top dimension) of the 4-manifold M and $\mathcal{M}_k$ is the moduli space of anti-self-dual SU(2)-instantons on a bundle with $c_2 = k > 0$. For an element W in R_M of the appropriate degree, one obtains a number by evaluating, or integrating, $\nu(W)$ on $\mathcal{M}_k$. The main technical difficulty here is that the moduli space $\mathcal{M}_k$ is rarely compact, so one needs to make sense of this "evaluation." With all the appropriate technicalities in place, these invariants could be shown to distinguish various homotopy-equivalent, homeomorphic 4-manifolds. All these early developments are described in detail in the book by Donaldson and Kronheimer (1990).

Basic Classes

Until the early 1990s, these instanton invariants could only be calculated in isolated favorable cases (although the calculations which were made, through the work of many mathematicians, led to a large number of further results about 4-manifold topology). Deeper understanding of their structure came with the work of Kronheimer and Mrowka. This work was, in large part, motivated by a natural question in geometric topology. Any homology class $\alpha \in H_2(M; \mathbf{Z})$ can be represented by an embedded, connected, smooth surface. One can define an integer $\underline{g}(\alpha)$ to be the minimal genus of such a representative. The problem is to find $\underline{g}(\alpha)$, or at least bounds on it. A well-known conjecture, ascribed to Thom, was that when M is the complex projective plane the minimal genus is realized by a complex curve; that is,

$$\underline{g}(\pm dH) = \tfrac{1}{2}(d-1)(d-2)$$

where H is the standard generator of $H_2(\mathbf{CP}^2)$ and $d \geq 1$.

The new geometrical idea introduced by Kronheimer and Mrowka was to study instantons over a 4-manifold M with singularities along a surface $\Sigma \subset M$. For such connections, there is a real parameter: the limit of the trace of the holonomy around small circles linking the surface. By varying this parameter, they were able to interpolate between moduli spaces of nonsingular instantons on different bundles over M and obtain relations between the different invariants. They also found that if the genus of Σ is suitably small then some of the invariants are forced to vanish, thus, conversely, getting information about $\underline{g}$ for 4-manifolds with nontrivial invariants. For example, they showed that if M is a K3 surface then $\underline{g}(\alpha) = (1/2)(\alpha \cdot \alpha + 2)$.

The structural results of Kronheimer and Mrowka (1995) introduced the notion of a 4-manifold of "simple type." Write the invariant defined above by the moduli space $\mathcal{M}_k$ as $I_k : R_M \to \mathbf{Q}$. Then I_k vanishes except on terms of degree $2d(k)$, where $d(k) = 4k - 3(1+p)$. We can put all these together to define $\underline{I} = \sum I_k : R_M \to \mathbf{Q}$. The ring R_M is a polynomial ring generated by classes $\alpha \in H_2(M)$, which have degree 2 in R_M, and a class X of degree 4 in R_M, corresponding to the generator of $H_0(M)$. The 4-manifold is of simple type if

$$\underline{I}(X^2 W) = 4\underline{I}(W)$$

for all $W \in R_M$. Under this condition, Kronheimer and Mrowka showed that all the invariants are determined by a finite set of "basic" classes $K_1, \ldots, K_s \in H_2(M)$ and rational numbers $\beta_1, \ldots, \beta_s$. To express the relation, they form a generating function

$$\mathcal{D}_M(\alpha) = \underline{I}(\mathrm{e}^{\alpha}) + \underline{I}\left(\frac{X}{2}\mathrm{e}^{\alpha}\right)$$

This is *a priori* a formal power series in $H^2(M)$ but *a posteriori* the series converges and can be regarded

as a function on $H_2(M)$. Kronheimer and Mrowka's result is that

$$\mathcal{D}_M(\alpha) = \exp\left(\frac{\alpha \cdot \alpha}{2}\right) \sum_{r=1}^{s} \beta_r e^{K_r \cdot \alpha}$$

It is not known whether all simply connected 4-manifolds are of simple type, but Kronheimer and Mrowka were able to show that this is the case for a multitude of examples. They also introduced a weaker notion of "finite type," and this condition was shown to hold in general by Munoz and Froyshov. The overall result of this work of Kronheimer and Mrowka was to make the calculation of the instanton invariants for many familiar 4-manifolds a comparatively straightforward matter.

3-Manifolds: Casson's Invariant

Gauge theory has also entered into 3-manifold topology. In 1985, Casson introduced a new integer-valued invariant of oriented homology 3-spheres which "counts" the set Z of equivalence classes of irreducible flat SU(2)-connections, or equivalently irreducible representations $\pi_1(Y) \to$ SU(2). Casson's approach (Akbulut and McCarthy 1990) was to use a Heegard splitting of a 3-manifold Y into two handle bodies Y^+, Y^- with a surface Σ as common boundary. Then $\pi_1(\Sigma)$ maps onto $\pi_1(Y)$ and a flat SU(2) connection on Y is determined by its restriction to Σ. Let M_Σ be the moduli space of irreducible flat connections over Σ (as discussed in the last section) and let $L^\pm \subset M_\Sigma$ be the subsets which extend over $Y^\pm$. Then $L^\pm$ are submanifolds of half the dimension of $\mathcal{M}_\Sigma$ and the set Z can be identified with the intersection $L^+ \cap L^-$. The Casson invariant is one-half the algebraic intersection number of L^+ and L^-. Casson showed that this is independent of the Heegard splitting (and is also, in fact, an integer, although this is not obvious). He showed that when Y is changed by Dehn surgery along a knot, the invariant changes by a term computed from the Alexander polynomial of the knot. This makes the Casson invariant computable in examples. (For a discussion of Casson's formula see Donaldson (1999).) Taubes showed that the Casson invariant could also be obtained in a more differential-geometric fashion, analogous to the instanton invariants of 4-manifolds (Taubes 1990).

3-Manifolds: Floer Theory

Independently, at about the same time, Floer (1989) introduced more sophisticated invariants – the Floer homology groups – of homology 3-spheres, using gauge theory. This development ran parallel to his introduction of similar ideas in symplectic geometry. Suppose, for simplicity, that the set Z of equivalence classes of irreducible flat connections is finite. For pairs ρ_-, ρ_+ in Z, Floer considered the instantons on the tube $Y \times R$ asymptotic to $\rho^\pm$ at $\pm\infty$. There is an infinite set of moduli spaces of such instantons, labeled by a relative Chern class, but the dimensions of these moduli spaces agree modulo 8. This gives a relative index $\delta(\rho_-, \rho_+) \in \mathbf{Z}/8$. If $\delta(\rho_-, \rho_+) = 1$ there is a moduli space of dimension 1 (possibly empty), but the translations of the tube act on this moduli space and, dividing by translations, we get a finite set. The number of points in this set, counted with suitable signs, gives an integer $n(\rho_-, \rho_+)$. Then, Floer considers the free abelian groups

$$C_* = \bigoplus_{\rho \in Z} \mathbf{Z}\langle\rho\rangle$$

generated by the set Z and a map $\partial : C_* \to C_*$ defined by

$$\partial(\langle\rho_-\rangle) = \sum n(\rho_-, \rho_+)\langle\rho_+\rangle$$

Here the sum runs over the ρ_+ with $\delta(\rho_-, \rho_+) = 1$. Floer showed that $\partial^2 = 0$ and the homology $HF_*(Y) = \ker\partial/\mathrm{Im}\,\partial$ is independent of the metric on Y (and various other choices made in implementing the construction in detail). The chain complex C_* and hence the Floer homology can be graded by $\mathbf{Z}/8$, using the relative index, so the upshot is to define 8 abelian groups $HF_i(Y)$: invariants of the 3-manifold Y. The Casson invariant appears now as the Euler characteristic of the Floer homology. There has been extensive work on extending these ideas to other 3-manifolds (not homology spheres) and gauge groups, but this line of research does not yet seem to have reached a clear-cut conclusion.

Part of the motivation for Floer's work came from Morse theory, and particularly the approach to this theory expounded by Witten (1982). The Chern–Simons functional is a map

$$CS : \mathcal{A}/\mathcal{G} \to R/Z$$

from the space of SU(2)-connections over Y. Explicitly, in a trivialization of the bundle

$$\mathrm{CS}(A) = \int_Y A \wedge \mathrm{d}A + \tfrac{3}{2} A \wedge A \wedge A$$

It appears as a boundary term in the Chern–Weil theory for the second Chern class, in a similar way as holonomy appears as a boundary term in the Gauss–Bonnet theorem. The set Z can be identified with the critical points of CS and the instantons on the tube as integral curves of the gradient vector

field of CS. Floer's definition mimics the definition of homology in ordinary Morse theory, taking Witten's point of view. It can be regarded formally as the "middle-dimensional" homology of the infinite-dimensional space $\mathcal{A}/\mathcal{G}$. See Atiyah (1988) and Cohen *et al.* (1995) for discussions of these ideas.

The Floer theory interacts with 4-manifold invariants, making up a structure approximating to a $(3+1)$-dimensional topological field theory (Atiyah 1988). Roughly, the numerical invariants of closed 4-manifolds generalize to invariants for a 4-manifold M with boundary Y taking values in the Floer homology of Y. If two such manifolds are glued along a common boundary, the invariants of the result are obtained by a pairing in the Floer groups. There are, however, at the moment, some substantial technical restrictions on this picture. This theory, as well as Floer's original construction, is developed in detail by Donaldson (2002). At the time of writing, the Floer homology groups are still difficult to compute in examples. One important tool is a surgery-exact sequence found by Floer (Braam and Donaldson 1995), related to Casson's surgery formula.

3-Manifolds: Jones–Witten Theory

There is another, quite different, way in which ideas from gauge theory have entered 3-manifold topology. This is the Jones–Witten theory of knot and 3-manifold invariants. This theory falls outside the main line of this article, but we will say a little about it since it draws on many of the ideas we have discussed. The goal of the theory is to construct a family of $(2+1)$-dimensional topological field theories indexed by an integer k, assigning complex vector space $H_k(\Sigma)$ to a surface Σ and an invariant in $H_k(\partial Y)$ to a 3-manifold-with-boundary Y. If ∂Y is empty, the vector space $H_k(\partial Y)$ is taken to be $\mathbf{C}$, so one seeks numerical invariants of closed 3-manifolds. Witten's (1989) idea is that these invariants of closed 3-manifolds are Feynmann integrals

$$\int_{\mathcal{A}/\mathcal{G}} e^{i2\pi k CS(A)} \mathcal{D}A$$

This functional integral is probably a schematic rather than a rigorous notion. The data associated with surfaces can, however, be defined rigorously. If we fix a complex structure I on Σ, we can define a vector space $H_k(\Sigma, I)$ to be

$$H_k(\Sigma, I) = H^0(\mathcal{M}(\Sigma); L^k)$$

where $\mathcal{M}(\Sigma)$ is the moduli space of stable holomorphic bundles/flat unitary connections over Σ and L is a certain holomorphic line bundle over $\mathcal{M}(\Sigma)$. These are the spaces of "conformal blocks" whose dimension is given by the Verlinde formulas. Recall that $\mathcal{M}(\Sigma)$, as a symplectic manifold, is canonically associated with the surface Σ, without any choice of complex structure. The Hilbert spaces $H_k(\Sigma, I)$ can be regarded as the quantization of this symplectic manifold, in the general framework of geometric quantization: the inverse of k plays the role of Planck's constant. What is not obvious is that this quantization is independent of the complex structure chosen on the Riemann surface: that is, that there is a natural identification of the vector spaces (or at least the associated projective spaces) formed by using different complex structures. This was established rigorously by Hitchin (1990) and Axelrod *et al.* (1991), who constructed a projectively flat connection on the bundle of spaces $H_k(\Sigma, I)$ over the space of complex structures I on Σ. At a formal level, these constructions are derived from the construction of the metaplectic representation of a linear symplectic group, since the $\mathcal{M}_\Sigma$ are symplectic quotients of an affine symplectic space.

The Jones–Witten invariants have been rigorously established by indirect means, but it seems that there is still work to be done in developing Witten's point of view. If Y^+ is a 3-manifold with boundary, one would like to have a geometric definition of a vector in $H_k(\partial Y^+)$. This should be the quantized version of the submanifold L^+ (which is Lagrangian in $\mathcal{M}_\Sigma$) entering into the Casson theory.

Seiberg–Witten Invariants

The instanton invariants of a 4-manifold can be regarded as the integrals of certain natural differential forms over the moduli spaces of instantons. Witten (1988) showed that these invariants could be obtained as functional integrals, involving a variant of the Feynman integral, over the space of connections and certain auxiliary fields (insofar as this latter integral is defined at all). A geometric explanation of Witten's construction was given by Atiyah and Jeffrey (1990). Developing this point of view, Witten made a series of predictions about the instanton invariants, many of which were subsequently verified by other means. This line of work culminated in 1994 where, applying developments in supersymmetric Yang–Mills QFT, Seiberg and Witten introduced a new system of invariants and a precise prediction as to how these should be related to the earlier ones.

The Seiberg–Witten invariants (Witten 1994) are associated with a Spin^c structure on a 4-manifold M. If M is simply connected this is specified by a class $K \in H^2(M; \mathbf{Z})$ lifting $w_2(M)$. One has spin bundles

$S^+, S^- \to M$ with $c_1(S^\pm) = K$. The Seiberg–Witten equation is for a spinor field ϕ – a section of S^+ and a connection A on the complex line bundle $\Lambda^2 S^+$. This gives a connection on S^+ and hence a Dirac operator

$$D_A : \Gamma(S^+) \to \Gamma(S^-)$$

The Seiberg–Witten equations are

$$D_A\phi = 0, \qquad F_A^+ = \sigma(\phi)$$

where $\sigma : S^+ \to \Lambda^+$ is a certain natural quadratic map. The crucial differential-geometric feature of these equations arises from the Weitzenbock formula

$$D_A^* D_A \phi = \nabla_A^* \nabla_A \phi + \frac{R}{4}\phi + \rho(F^+)\phi$$

where R is the scalar curvature and ρ is a natural map from Λ^+ to the endomorphisms of S^+. Then ρ is adjoint to σ and

$$\langle \rho(\sigma(\phi))\phi, \phi \rangle = |\phi|^4$$

It follows easily from this that the moduli space of solutions to the Seiberg–Witten equation is compact. The most important invariants arise when K is chosen so that

$$K \cdot K = 2\chi(M) + 3\,\mathrm{sign}(M)$$

where $\chi(M)$ is the Euler characteristic and sign(M) is the signature. (This is just the condition for K to correspond to an almost-complex structure on M.) In this case, the moduli space of solutions is zero dimensional (after generic perturbation) and the Seiberg–Witten invariant SW(K) is the number of points in the moduli space, counted with suitable signs.

Witten's conjecture relating the invariants, in its simplest form, is that when M has simple type the classes K for which SW(K) is nonzero are exactly the basic classes K_r of Kronheimer and Mrowka and that

$$\beta_r = 2^{C(M)}\mathrm{SW}(K_r)$$

where $C(M) = 2 + (1/4)(7\chi(M) + 11\,\mathrm{sign}(M))$. This asserts that the two sets of invariants contain exactly the same information about the 4-manifold.

The evidence for this conjecture, via calculations of examples, is very strong. A somewhat weaker statement has been proved rigorously by Feehan and Leness (2003). They use an approach suggested by Pidstragatch and Tyurin, studying moduli spaces of solutions to a nonabelian version of the Seiberg–Witten equations. These contain both the instanton and abelian Seiberg–Witten moduli spaces, and the strategy is to relate the topology of these two sets by standard localization arguments. (This approach is related to ideas introduced by Thaddeus (1994) in the case of bundles over Riemann surfaces.) The serious technical difficulty in this approach stems from the lack of compactness of the nonabelian moduli spaces. The more general versions of Witten's conjecture (Moore and Witten 1997) (e.g., when $b^+(M) = 1$) contain very complicated formulas, involving modular forms, which presumably arise as contributions from the compactification of the moduli spaces.

Applications

Regardless of the connection with the instanton theory, one can go ahead directly to apply the Seiberg–Witten invariants to 4-manifold topology, and this has been the main direction of research since the 1990s. The features of the Seiberg–Witten theory which have led to the most prominent developments are the following.

1. The reduction of the equations to two dimensions is very easy to understand. This has led to proofs of the Thom conjecture and wide-ranging generalizations (Ozsvath and Szabo 2000).
2. The Weitzenboch formula implies that, if M has positive scalar curvature, then solutions to the Seiberg–Witten equations must have $\phi = 0$. This has led to important interactions with four-dimensional Riemannian geometry (Lebrun 1996).
3. In the case when M is a symplectic manifold, there is a natural deformation of the Seiberg–Witten equations, discovered by Taubes (1996), who used it to show that the Seiberg–Witten invariants of M are nontrivial. More generally, Taubes showed that for large values of the deformation parameter the solutions of the deformed equation localize around surfaces in the 4-manifold and used this to relate the Seiberg–Witten invariants to the Gromov theory of pseudoholomorphic curves. These results of Taubes have completely transformed the subject of four-dimensional symplectic geometry.

Bauer and Furuta (2004) have combined the Seiberg–Witten theory with more sophisticated algebraic topology to obtain further results about 4-manifolds. They consider the map from the space of connections and spinor fields defined by the formulas on the left-hand side of the equations. The general idea is to obtain invariants from the homotopy class of this map, under a suitable notion of homotopy. A technical complication arises from the gauge group action, but this can be reduced to the action of a single U(1). Ignoring this issue, Bauer and Furuta have obtained invariants in the stable homotopy groups $\lim_{N\to\infty} \pi_{N+r}(S^N)$, which reduce to the ordinary numerical invariants when $r = 1$. Using these invariants, they obtain results about connected sums of

4-manifolds, for which the ordinary invariants are trivial. Using refined cobordism invariants ideas, Furuta made great progress towards resolving the question of which intersection forms arise from smooth, simply connected 4-manifolds. A well-known conjecture is that, if such a manifold is spin, then the second Betti number satisfies

$$b_2(M) \geq \tfrac{11}{8}|\text{sign}(M)|$$

Furuta (2001) proved that $b_2(M) \geq (10/8)|\text{sign}(M)| + 2$.

An important and very recent achievement, bringing together many different lines of work, is the proof of "Property P" in 3-manifold topology by Kronheimer and Mrowka (2004). This asserts that one cannot obtain a homotopy sphere (counter-example to the Poincaré conjecture) by +1-surgery along a nontrivial knot in S^3. The proof uses work of Gabai and Eliashberg to show that the manifold obtained by 0-framed surgery is embedded in a symplectic 4-manifold; Taubes' results to show that the Seiberg–Witten invariants of this 4-manifold are nontrivial; Feehan and Leness' partial proof of Witten's conjecture to show that the same is true for the instanton invariants; and the gluing rule and Floer's exact sequence to show that the Floer homology of the +1-surgered manifold is nontrivial. It follows then from the definition of Floer homology that the fundamental group of this manifold is not trivial; in fact, it must have an irreducible representation in SU(2).

See also: Cotangent Bundle Reduction; Floer Homology; Gauge Theories from Strings; Gauge Theoretic Invariants of 4-Manifolds; Instantons: Topological Aspects; Knot Homologies; Moduli Spaces: An Introduction; Nonperturbative and Topological Aspects of Gauge Theory; Seiberg–Witten Theory; Topological Quantum Field Theory: Overview; Variational Techniques for Ginzburg–Landau Energies.

Further Reading

Akbulut S and McCarthy J (1990) *Casson's Invariant for Homology 3-Spheres*. Princeton, NJ: Princeton University Press.

Atiyah (1988) New invariants for 3 and 4 dimensional manifolds. In: *The Mathematical Heritage of Hermann Weyl*, Proceedings of Symposia in Pure Mathematics, vol. 48, pp. 285–299. American Mathematical Society.

Atiyah MF (1988) *Topological Quantum Field Theories*, vol. 68, pp. 135–186. Math. Publ. IHES.

Atiyah MF and Bott R (1982) The Yang–Mills equations over Riemann surfaces. *Philosophical Transactions of the Royal Society of London, Series A* 308: 523–615.

Atiyah MF, Drinfeld V, Hitchin NJ, and Manin YuI (1978) Construction of instantons. *Physics Letters A* 65: 185–187.

Atiyah MF and Hitchin NJ (1989) *The Geometry and Dynamics of Magnetic Monopoles*. Princeton, NJ: Princeton University Press.

Atiyah MF, Hitchin NJ, and Singer IM (1978) Self-duality in four-dimensional Riemannian geometry. *Proceedings of the Royal Society of London, Series A* 362: 425–461.

Atiyah MF and Jeffrey L (1990) Topological Lagrangians and cohomology. *Journal of Geometry and Physics* 7: 119–136.

Atiyah MF and Jones JDS (1978) Topological aspects of Yang–Mills theory. *Communications in Mathematical Physics* 61: 97–118.

Axelrod S, Della Pietra S, and Witten E (1991) Geometric quantisation of Chern–Simons gauge theories. *Journal of Differential Geometry* 33: 787–902.

Bauer S and Furuta M, A stable cohomotopy refinement of the Seiberg–Witten invariants, I. *Inventiones Mathematicae* 155: 1–19.

Bourguignon J-P and Lawson HB (1981) Stability and isolation phenomena for Yang–Mills fields. *Communications in Mathematical Physics* 79: 189–230.

Boyer C, Mann B, Hurtubise J, and Milgram R (1993) The topology of instanton moduli spaces. I: the Atiyah–Jones conjecture. *Annals of Mathematics* 137: 561–609.

Braam PJ and Donaldson SK (1995) Floer's work on instanton homology, knots and surgery. In: Hofer *et al.* (eds.) *The Floer Memorial Volume*, vol. 133, Progress in Mathematics. Basle: Birkhäuser.

Bradlow S, Daskalopoulos G, Garcia-Prada O, and Wentworth R (1995) Stable augmented bundles over Riemann surfaces. In: Hitchin *et al.* (eds.) *Vector Bundles in Algebraic Geometry*, pp. 15–67. Cambridge: Cambridge University Press.

Buchdahl NP (1986) Instantons on $\mathbf{CP}^2$. *Journal of Differential Geometry* 24: 19–52.

Cohen RL, Jones JDS, and Segal GB (1995) Floer's infinite-dimensional Morse theory and homotopy theory. In: Hofer *et al.* (eds.) *The Floer Memorial Volume*, pp. 297–325. Basle: Birkhäuser.

Donaldson SK (1983) An application of gauge theory to four-dimensional topology. *Journal of Differential Geometry* 18: 279–315.

Donaldson SK (1984) Nahm's equations and the classification of monopoles. *Communications in Mathematical Physics* 96: 387–407.

Donaldson SK (1985) Anti-self-dual Yang–Mills connections on complex algrebraic surfaces and stable vector bundles. *Proceedings of the London Mathematical Society* 3: 1–26.

Donaldson SK (1987) Infinite determinants, stable bundles and curvature. *Duke Mathematical Journal* 54: 231–247.

Donaldson SK (1990) Polynomial invariants of smooth four-manifolds. *Topology* 29: 257–315.

Donaldson SK (1993) Gluing techniques in the cohomology of moduli spaces. In: Goldberg and Phillips (eds.) *Topological Methods in Modern Mathematics*, pp. 137–170. Houston: Publish or Perish.

Donaldson SK (1999) *Topological Field Theories and Formulae of Casson and Meng-Taubes, Geometry and Topology Monographs*, vol. 2, pp. 87–102.

Donaldson SK (2002) *Floer Homology Groups in Yang–Mills Theory*. Cambridge: Cambridge University Press.

Donaldson SK and Kronheimer PB (1990) *The geometry of four-manifolds*. Oxford: Oxford University Press.

Donaldson SK and Thomas R (1998) Gauge theory in higher dimensions. In: Hugget *et al.* (eds.) *The Geometric Universe*, pp. 31–47. Oxford: Oxford University Press.

Earl R and Kirwan FC (1999) Pontrayagin rings of moduli spaces of arbitrary rank bundles over Riemann surfaces. *Journal of London Mathematical Society* 60: 835–846.

Feehan PMN and Leness TG (2003) A general SO(3)-monopole cobordism formula relating Donaldson and Seiberg–Witten invariants, arXiv:math.DG/0203047.

Floer A (1989) An instanton invariant for 3-manifolds. *Communications in Mathematical Physics* 118: 215–240.

Furuta M (2001) Monopole equation and the 11/8 conjecture. *Mathematical Research Letters* 8: 279–291.

Hitchin NJ (1983) On the construction of monopoles. *Communications in Mathematical Physics* 89: 145–190.

Hitchin NJ (1987) The self-duality equations over Riemann surfaces. *Proceedings of the London Mathematical Society* 55: 59–126.

Hitchin NJ (1990) Flat connections and geometric quantisation. *Communications in Mathematical Physics* 131: 347–380.

Hitchin NJ (1990) Harmonic maps from the 2-torus to the 3-sphere. *Journal of Differential Geometry* 31: 627–710.

Hitchin NJ, Karhlede A, Lindström U, and Roceck M (1987) Hyper-Kähler metrics and supersymmetry. *Communications in Mathematical Physics* 108: 535–589.

Jaffe A and Taubes CH (1980) *Vortices and Monopoles*, Progress in Physics, vol. 2. Basle: Birkhäuser.

Jeffrey L and Kirwan FC (1998) Intersection theory on moduili spaces of holomorphic bundles of arbitrary rank on a Riemann surface. *Annals of Mathematics* 148: 109–196.

Joyce D (1996) Compact Riemannian 7-manifolds with holonomy G_2, I. *Journal of Differential Geometry* 43: 291–328.

King AD and Newstead PE (1998) On the cohomology ring of the moduli space of stable rank 2 bundles over a curve. *Topology* 37: 407–418.

Kirwan FC (1994) Geometric invariant theory and the Atiyah–Jones conjecture. In: Laudel and Jahren (eds.) *The Sophus Lie Memorial Volume Pages*, pp. 161–186. Oslo: Scandinavian University Press.

Kronheimer PB and Mrowka TS (1995) embedded surfaces and the structure of Donaldson's polynomial invariants. *Journal of Differential Geometry* 41: 573–734.

Kronheimer PB and Mrowka TS, Witten's conjecture and property P. *Geometry and Topology* 8: 295–300.

Lebrun C (1996) Four-manifolds without Einstein metrics. *Mathematical Research Letters* 3: 133–147.

Mason L and Woodhouse N (1996) *Integrability, Self-duality and Twistor Theory. Advances in Theoretical and Mathematical Physics*. Oxford: Oxford University Press.

Moore G and Witten E (1997) Integration over the u-plane in Donaldson Theory. *Adv. Theor. Mathematical Phys* 1: 298–387.

Munoz V (1999) Quantum cohomology of the moduli space of stable bundles over a Riemann surface. *Duke Mathematics Journal* 98: 525–540.

Nahm W (1982) The construction of all self-dual monopoles by the ADHM method. In: *Monopoles in Quantum Field Theory*, pp. 87–94. Singapore: World Scientific.

Nakajima H (1988) Compactness of the moduli space of Yang–Mills connections in higher dimensions. *Journal of the Mathematical Society of Japan* 40: 383–392.

Narasimhan MS and Seshadri CS (1965) Stable and unitary vector bundles on compact Riemann surfaces. *Annals of Mathematics* 65: 540–567.

Newstead PE (1967) Topology of some spaces of stable bundles. *Topology* 6: 241–262.

Ozsvath P and Szabo Z (2000) The sympletic Thom conjecture. *Annals of Mathematics* 151: 93–124.

Price P (1983) A monotonicity formula for Yang–Mills fields. *Manuscripta Mathematica* 43: 131–166.

Quillen D (1985) Determinants of Cauchy–Riemann operators over a Riemann surface. *Functional Analysis and its Applications* 14: 31–34.

Sadun L and Segert J (1992) Non self-dual Yang–Mills connections with quadropole symmetry. *Communications in Mathematical Physics* 145: 363–391.

Sibner L, Sibner R, and Uhlenbeck K (1989) Solutions to the Yang–Mills equations that are not self-dual. *Proceedings of the National Academy of Sciences of the USA* 86: 8610–8613.

Taubes CH (1982) Self-dual Yang–Mills connections over non self-dual 4-manifolds. *Journal of Differential Geometry* 17: 139–170.

Taubes CH (1983) Stability in Yang–Mills theories. *Communications in Mathematical Physics* 91: 235–263.

Taubes CH (1990) Casson's invariant and gauge theory. *Journal of Differential Geometry* 31: 363–430.

Taubes CH (1998) A framework for Morse Theory for the Yang–Mills functional. *Inventiones Mathematical* 94: 327–402.

Taubes CH (1996) SW ⇒ Gr: from the Seiberg–Witten equations to pseudo-holomorphic curves. *Journal of American Mathematical Society* 9: 819–918.

Thaddeus M (1992) Conformal Field Theory and the cohomology of moduli spaces of stable bundles. *Journal of Differential Geometry* 35: 131–149.

Thaddeues M (1994) Stable pairs, linear systems and the Verlinde formulae. *Inventiones Mathematicae* 117: 317–353.

Tian G (2000) gauge theory and calibrated geometry, I. *Annals of Mathematics* 151: 193–268.

Uhlenbeck KK (1982) Connections with L^p bounds on curvature. *Communications in Mathematical Physics* 83: 11–29.

Uhlenbeck KK and Yau S-T (1986) On the existence of hermitian Yang–Mills connections on stable bundles over compact hyper-Kähler manifolds. *Communications on Pure and Applied Mathematics* 39: 257–293.

Ward RS (1877) On self-dual gauge fields. *Physics Letters A* 61: 81–82.

Witten E (1982) Supersymmetry and Morse Theory. *Journal of Differential Geometry* 16: 661–692.

Witten E (1988) Topological quantum field theory. *Communications in Mathematical Physics* 117: 353–386.

Witten E (1989) Quantum field theory and the Jones polynomial. *Communications in Mathematical Physics* 121: 351–399.

Witten E (1991) On quantum gauge theory in two dimensions. *Communications in Mathematical Physics* 141: 153–209.

Witten E (1992) Two dimensional quantum gauge theory revisited. *Journal of Geometry and Physics* 9: 303–368.

Witten E (1994) Monopoles and four-manifolds. *Mathematical Research Letters* 1: 769–796.

Instantons: Topological Aspects

M Jardim, IMECC–UNICAMP, Campinas, Brazil

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Let X be a closed (connected, compact without boundary) smooth manifold of dimension 4, provided with a Riemannian metric denoted by g. Let Ω^p_X denote space of smooth p-forms on X, that is, the sections of $\wedge^p TX$. The Hodge operator acting on p-forms,

$$* : \Omega^p_X \to \Omega^{4-p}_X$$

satisfies $*^2 = (-1)^p$. In particular, $*$ splits Ω^2_X into two subspaces $\Omega^{2,\pm}_X$ with eigenvalues ± 1:

$$\Omega^2_X = \Omega^{2,+}_X \oplus \Omega^{2,-}_X \qquad [1]$$

Note also that this decomposition is an orthogonal one, with respect to the inner product:

$$\langle \omega_1, \omega_2 \rangle = \int_X \omega_1 \wedge *\omega_2$$

A 2-form ω is said to be self-dual if $*\omega = \omega$ and it is said to be anti-self-dual if $*\omega = -\omega$. Any 2-form ω can be written as the sum

$$\omega = \omega^+ + \omega^-$$

of its self-dual ω^+ and anti-self-dual ω^- components.

Now let E be a complex vector bundle over X as above, provided with a connection ∇, regarded as a $\mathbb{C}$-linear operator

$$\nabla : \Gamma(E) \to \Gamma(E) \otimes \Omega^1_X$$

satisfying the Leibnitz rule:

$$\nabla(f\sigma) = f\nabla\sigma + \sigma \otimes \mathrm{d}f$$

for all $f \in \mathrm{C}^\infty(X)$ and $\sigma \in \Gamma(E)$. Its curvature $F_\nabla = \nabla \circ \nabla$ is a 2-form with values in $\mathrm{End}(E)$, that is, $F_\nabla \in \Gamma(\mathrm{End}(E)) \otimes \Omega^2_X$, satisfying the Bianchi identity $\nabla F_\nabla = 0$.

The Yang–Mills equation is

$$\nabla * F_\nabla = 0 \qquad [2]$$

It is a second-order nonlinear equation on the connection ∇. It amounts to a nonabelian generalization of Maxwell equations, to which it reduces when E is a line bundle; the four components of ∇ are interpreted as the electric and magnetic potentials.

An instanton on E is a smooth connection ∇ whose curvature F_∇ is anti-self-dual as a 2-form, that is, it satisfies:

$$F^+_\nabla = 0, \quad \text{that is,} \quad *F_\nabla = -F_\nabla \qquad [3]$$

The instanton equation is still nonlinear (it is linear only if E is a line bundle), but it is only first-order on the connection.

Notes:

Note that if F_∇ is either self-dual or anti-self-dual as a 2-form, then the Yang–Mills equation is automatically satisfied:

$$*F_\nabla = \pm F_\nabla \Rightarrow \nabla * F_\nabla = \pm \nabla F_\nabla = 0$$

by the Bianchi identity. In other words, instantons are particular solutions of the Yang–Mills equation. Furthermore, while the Yang–Mills equation [2] makes sense over any Riemannian manifold, the instanton equation [3] is well defined only in dimension 4.

A gauge transformation is a bundle automorphism $g: E \to E$ covering the identity. The set of all gauge transformations of a given bundle $E \to X$ forms a group through composition, called the gauge group and denoted by $\mathcal{G}(E)$. The gauge group acts on the set of all smooth connections on E by conjugation:

$$g \cdot \nabla = g^{-1} \nabla g$$

It is then easy to see that [3] is a gauge-invariant condition, since $F_{g\cdot\nabla} = g^{-1} F_\nabla g$. The anti-self-duality equation [3] is also conformally invariant: a conformal change in the metric does not change the decomposition [1], so it preserves self-dual and anti-self-dual 2-forms.

The topological charge k of the instanton ∇ is defined by the integral

$$k = -\frac{1}{8\pi^2} \int_X \mathrm{tr}(F_\nabla \wedge F_\nabla)$$
$$= c_2(E) - \frac{1}{2} c_1(E)^2 \qquad [4]$$

where the second equality follows from Chern–Weil theory.

If X is a smooth, noncompact, complete Riemannian manifold, an instanton on X is an anti-self-dual connection for which the integral [4] converges. Note that, in this case, k as above need not be an integer; however, it is always expected to be quantized, that is, always a multiple of some fixed (rational) number which depends only on the base manifold X.

Summary This note is organized as follows. After revisiting the variational approach to the anti-self-duality equation [3], we study instantons over the simplest possible Riemannian 4-manifold, $\mathbb{R}^4$ with the flat Euclidean metric. In the subsequent sections, we present 't Hooft's explicit solutions, the ADHM construction, and its dimensional reductions to $\mathbb{R}^3, \mathbb{R}^2$ and $\mathbb{R}$. We conclude by explaining the construction of the central object of study in gauge theory, the instanton moduli spaces.

Variational Aspects of Yang–Mills Equation

Given a fixed smooth vector bundle $E \to X$, let $\mathcal{A}(E)$ be the set of all (smooth) connections on E. The Yang–Mills functional is defined by

$$\mathrm{YM} : \mathcal{A}(E) \to \mathbb{R}$$
$$\mathrm{YM}(\nabla) = \|F_\nabla\|_{L^2}^2 = \int_M \mathrm{tr}(F_\nabla \wedge *F_\nabla) \qquad [5]$$

The Euler–Lagrange equation for this functional is exactly the Yang–Mills equation [2]. In particular, self-dual and anti-self-dual connections yield critical points of the Yang–Mills functional.

Splitting the curvature into its self-dual and anti-self-dual parts, we have

$$\mathrm{YM}(\nabla) = \|F_\nabla^+\|_{L^2}^2 + \|F_\nabla^-\|_{L^2}^2$$

It is then easy to see that every anti-self-dual connection ∇ is an absolute minimum for the Yang–Mills functional, and that $\mathrm{YM}(\nabla)$ coincides with the topological charge [4] of the instanton ∇ times $8\pi^2$.

One can construct, for various 4-manifolds but most interestingly for $X = S^4$, solutions of the Yang–Mills equations which are neither self-dual nor anti-self-dual. Such solutions do not minimize [5]. Indeed, at least for gauge group SU(2) or SU(3), it can be shown that there are no other local minima: any critical point which is neither self-dual nor anti-self-dual is unstable and must be a "saddle point" (Bourguignon and Lawson Jr. 1981).

Instantons on Euclidean Space

Let $X = \mathbb{R}^4$ with the flat Euclidean metric, and consider a Hermitian vector bundle $E \to \mathbb{R}^4$. Any connection ∇ on E is of the form $d + A$, where d denotes the usual de Rham operator and $A \in \Gamma(\mathrm{End}(E)) \otimes \Omega^1_{\mathbb{R}^4}$ is a 1-form with values in the endomorphisms of E; this can be written as follows:

$$A = \sum_{k=1}^{4} A_k \, dx^k, \quad A_k : \mathbb{R}^4 \to \mathfrak{u}(r)$$

In the Euclidean coordinates x_1, x_2, x_3, x_4, the anti-self-duality equation [3] is given by

$$F_{12} = F_{34}, \quad F_{13} = -F_{24}, \quad F_{14} = F_{23}$$

where

$$F_{ij} = \frac{\partial A_j}{\partial x_i} - \frac{\partial A_i}{\partial x_j} + [A_i, A_j]$$

The simplest explicit solution is the charge-1 SU(2) instanton on $\mathbb{R}^4$. The connection 1-form is given by

$$A_0 = \frac{1}{1+|x|^2} \cdot \mathrm{Im}(q\mathrm{d}\bar{q}) \qquad [6]$$

where q is the quaternion $q = x_1 + x_2\boldsymbol{i} + x_3\boldsymbol{j} + x_4\boldsymbol{k}$, while Im denotes the imaginary part of the product quaternion; we are regarding $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ as a basis of the Lie algebra $\mathfrak{su}(2)$; from this, one can compute the curvature:

$$F_{A_0} = \left(\frac{1}{1+|x|^2}\right)^2 \cdot \mathrm{Im}(\mathrm{d}q \wedge \mathrm{d}\bar{q}) \qquad [7]$$

We see that the action density function

$$|F_{A_0}|^2 = \left(\frac{1}{1+|x|^2}\right)^2$$

has a bell-shaped profile centered at the origin and decays like r^{-4}.

Let $t_{\lambda,y}: \mathbb{R}^4 \to \mathbb{R}^4$ be the isometry given by the composition of a translation by $y \in \mathbb{R}^4$ with a homothety by $\lambda \in \mathbb{R}^+$. The pullback connection $t^*_{\lambda,y}A_0$ is still anti-self-dual; more explicitly,

$$A_{\lambda,y} = t^*_{\lambda,y}A_0 = \frac{\lambda^2}{\lambda^2+|x-y|^2} \cdot \mathrm{Im}(q\mathrm{d}\bar{q})$$

$$F_{A_{\lambda,y}} = \left(\frac{\lambda^2}{\lambda^2+|x-y|^2}\right)^2 \cdot \mathrm{Im}(\mathrm{d}q \wedge \mathrm{d}\bar{q})$$

Note that the action density function $|F_A|^2$ has again a bell-shaped profile centered at y and decays like r^{-4}; the parameter λ measures the concentration of the energy density function, and can be interpreted as the "size" of the instanton $A_{\lambda,y}$.

Instantons of topological charge k can be obtained by "superimposing" k basic instantons, via the so-called 't Hooft ansatz. Consider the function $\rho: \mathbb{R}^4 \to \mathbb{R}$ given by

$$\rho(x) = 1 + \sum_{j=1}^{k} \frac{\lambda_j^2}{(x-y_j)^2}$$

where $\lambda_j \in \mathbb{R}$ and $y_j \in \mathbb{R}^4$. Then the connection 1-form $A = A_\mu \mathrm{d}x_\mu$ with coefficients

$$A_\mu = i\sum_{\nu=1}^{4} \bar{\sigma}_{\mu\nu} \frac{\partial}{\partial x_\nu} \ln(\rho(x)) \qquad [8]$$

is anti-self dual; here, $\bar{\sigma}_{\mu\nu}$ are the matrices given by $(\mu, \nu = 1, 2, 3)$:

$$\sigma_{\mu\nu} = \frac{1}{4\mathrm{i}}[\sigma_\mu, \sigma_\nu] \qquad \bar{\sigma}_{\mu 4} = \frac{1}{2}\sigma_\mu$$

where σ_μ are the Pauli matrices.

The connection [8] correspond to k instantons centered at points y_i with size λ_i. The basic instanton [6] is exactly (modulo gauge transformation) what one obtains from [8] for the case $k = 1$. The 't Hooft instantons form a $5k$-parameter family of anti-self-dual connections.

SU(2) instantons are also the building blocks for instantons with general structure group (Bernard *et al.* 1977). Let G be a compact semisimple Lie group, with Lie algebra $\mathfrak{g}$. Let $\phi: \mathfrak{su}(2) \to \mathfrak{g}$ be any injective Lie algebra homomorphism. If A is an anti-self-dual SU(2) connection 1-form, then it is easy to see that $\phi(A)$ is an anti-self-dual G-connection 1-form. Using [8] as an example, we have that

$$A = \mathrm{i}\sum_{\mu,\nu} \phi(\bar{\sigma}_{\mu\nu}) \frac{\partial}{\partial x_\nu} \ln(\rho(x))\mathrm{d}x_\mu \qquad [9]$$

is a G-instanton on $\mathbb{R}^4$.

While this guarantees the existence of G-instantons on $\mathbb{R}^4$, note that the instanton [9] might be reducible (e.g., ϕ can simply be the obvious inclusion of $\mathfrak{su}(2)$ into $\mathfrak{su}(n)$ for any n) and that its charge depends on the choice of representation ϕ. Furthermore, it is not clear whether every G-instanton can be obtained in this way, as the inclusion of a SU(2) instanton through some representation $\phi: \mathfrak{su}(2) \to \mathfrak{g}$.

The ADHM Construction

All SU(r) instantons on $\mathbb{R}^4$ can be obtained through a remarkable construction due to Atiyah, Drinfeld, Hitchin, and Manin. It starts by considering Hermitian vector spaces V and W of dimension c and r, respectively, and the following data (the so-called ADHM data):

$$B_1, B_2 \in \mathrm{End}(V), \quad i \in \mathrm{Hom}(W, V)$$
$$j \in \mathrm{Hom}(V, W)$$

Assume, moreover, that (B_1, B_2, i, j) satisfy the ADHM equations:

$$[B_1, B_2] + ij = 0 \qquad [10]$$

$$[B_1, B_1^\dagger] + [B_2, B_2^\dagger] + ii^\dagger - j^\dagger j = 0 \qquad [11]$$

Now consider the following maps

$$\alpha: V \times \mathbb{R}^4 \to (V \oplus V \oplus W) \times \mathbb{R}^4$$
$$\beta: (V \oplus V \oplus W) \times \mathbb{R}^4 \to V \times \mathbb{R}^4$$

given as follows (**1** denotes the appropriate identity matrix):

$$\alpha(z_1, z_2) = \begin{pmatrix} B_1 + z_1\mathbf{1} \\ B_2 + z_2\mathbf{1} \\ j \end{pmatrix} \qquad [12]$$

$$\beta(z_1, z_2) = (-B_2 - z_2\mathbf{1} \quad B_1 + z_1\mathbf{1} \quad \mathrm{i}) \qquad [13]$$

where $z_1 = x_1 + \mathrm{i}x_2$ and $z_2 = x_3 + \mathrm{i}x_4$ are complex coordinates on $\mathbb{R}^4$. The maps [12] and [13] should be understood as a family of linear maps parametrized by points in $\mathbb{R}^4$.

A straightforward calculation shows that the ADHM equation [10] implies that $\beta\alpha = 0$ for every $(z_1, z_2) \in \mathbb{R}^4$. Therefore, the quotient $E = \ker\beta/\mathrm{im}\ \alpha = \ker\beta \cap \ker\alpha^\dagger$ forms a complex vector bundle over $\mathbb{R}^4$ or rank r whenever (B_1, B_2, i, j) is such that α is injective and β is surjective for every $(z_1, z_2) \in \mathbb{R}^4$.

To define a connection on E, note that E can be regarded as a sub-bundle of the trivial bundle $(V \oplus V \oplus W) \times \mathbb{R}^4$. So let $\iota : E \to (V \oplus V \oplus W) \times \mathbb{R}^4$ be the inclusion, and let $P : (V \oplus V \oplus W) \times \mathbb{R}^4 \to E$ be the orthogonal projection onto E. We can then define a connection ∇ on E through the projection formula

$$\nabla s = P\underline{d}\iota(s)$$

where $\underline{d}$ denotes the trivial connection on the trivial bundle $(V \oplus V \oplus W) \times \mathbb{R}^4$.

To see that this connection is anti-self-dual, note that projection P can be written as follows:

$$P = \mathbf{1} - \mathcal{D}^\dagger\Xi^{-1}\mathcal{D}$$

where

$$\mathcal{D} : (V \oplus V \oplus W) \times \mathbb{R}^4 \to (V \oplus V) \times \mathbb{R}^4$$

$$\mathcal{D} = \begin{pmatrix} \beta \\ \alpha^\dagger \end{pmatrix}$$

and $\Xi = \mathcal{D}\mathcal{D}^\dagger$. Note that $\mathcal{D}$ is surjective, so that Ξ is indeed invertible. Moreover, it also follows from [11] that $\beta\beta^\dagger = \alpha^\dagger\alpha$, so that $\Xi^{-1} = (\beta\beta^\dagger)^{-1}\mathbf{1}$.

The curvature F_∇ is given by

$$\begin{aligned} F_\nabla &= P\left(\underline{d}(\mathbf{1} - \mathcal{D}_1^\dagger\Xi^{-1}\mathcal{D})\underline{d}\right) = P\left(\underline{d}\mathcal{D}^\dagger\Xi^{-1}(\underline{d}\mathcal{D})\right) \\ &= P\left((\underline{d}\mathcal{D}^\dagger)\Xi^{-1}(\underline{d}\mathcal{D}) + \mathcal{D}^\dagger\underline{d}(\Xi^{-1}(\underline{d}\mathcal{D}))\right) \\ &= (\underline{d}\mathcal{D}^\dagger)\Xi^{-1}(\underline{d}\mathcal{D}) \end{aligned}$$

for $P(\mathcal{D}^\dagger\underline{d}(\Xi^{-1}(\underline{d}\mathcal{D}))) = 0$ on $E = \ker\mathcal{D}$. Since Ξ^{-1} is diagonal, we conclude that F_∇ is proportional to $d\mathcal{D}^\dagger \wedge d\mathcal{D}$, as a 2-form.

It is then a straightforward calculation to show that each entry of $d\mathcal{D}^\dagger \wedge d\mathcal{D}$ belongs to $\Omega^{2,-}$.

The extraordinary accomplishment of Atiyah, Drinfeld, Hitchin, and Manin was to show that every instanton, up to gauge equivalence, can be obtained in this way (see, e.g., Donaldson and Kronheimer 1990). For instance, the basic SU(2) instanton [6] is associated with the following data ($c = 1, r = 2$):

$$B_1, B_2 = 0, \quad i = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad j = (0 \quad 1)$$

Remark The ADHM data (B_1, B_2, i, j) are said to be stable if β is surjective for every $(z_1, z_2) \in \mathbb{R}^4$, and it is said to be costable if α is injective for every $(z_1, z_2) \in \mathbb{R}^4$. (B_1, B_2, i, j) is regular if it is both stable and costable. The quotient:

$$\{\text{regular solutions of (10) and (11)}\}/U(V)$$

coincides with the moduli space of instantons of rank $r = \dim W$ and charge $c = \dim V$ on $\mathbb{R}^4$ (see below). It is also an example of a quiver variety (*see* Finite Dimensional Algebras and Quivers), associated to the quiver consisting of two vertices V and W, two loop-edges on the vertex V and two edges linking V to W, one in each direction.

Dimensional Reductions of the Anti-Self-Dual Yang–Mills Equation

As pointed out above, a connection on a Hermitian vector bundle $E \to \mathbb{R}^4$ of rank r can be regarded as 1-form

$$A = \sum_{k=1}^{4} A_k(x_1, \ldots, x_4)\mathrm{d}x^k, \quad A_k : \mathbb{R}^4 \to \mathfrak{u}(r)$$

Assuming that the connection components A_k are invariant under translation in one direction, say x_4, we can think of

$$\underline{A} = \sum_{k=1}^{3} A_k(x_1, x_2, x_3)\mathrm{d}x^k$$

as a connection on a Hermitian vector bundle over $\mathbb{R}^3$, with the fourth component $\phi = A_4$ being regarded as a bundle endomorphism $\phi : E \to E$, called a Higgs field. In this way, the anti-self-duality equation [3] reduces to the so-called Bogomolny (or monopole) equation:

$$F_{\underline{A}} = *\mathrm{d}\phi \qquad [14]$$

where $*$ is the Euclidean Hodge star in dimension 3.

Now assume that the connection components A_k are invariant under translation in two directions, say x_3 and x_4. Consider

$$\underline{A} = \sum_{k=1}^{2} A_k(x_1, x_2)\mathrm{d}x^k$$

as a connection on a Hermitian vector bundle over $\mathbb{R}^2$, with the third and fourth components combined into a complex bundle endomorphism:

$$\Phi = (A_3 + \mathrm{i} \cdot A_4)(\mathrm{d}x_1 - \mathrm{i} \cdot \mathrm{d}x_2)$$

taking values on 1-forms. The anti-self-duality equation [3] is then reduced to the so-called Hitchin's equations:

$$F_{\underline{A}} = [\Phi, \Phi^*], \qquad \bar{\partial}_{\underline{A}} \Phi = 0 \qquad [15]$$

Conformal invariance of the anti-self-duality equation means that Hitchin's equations are well defined over any Riemann surface.

Finally, assume that the connection components A_k are invariant under translation in three directions, say x_2, x_3, and x_4. After gauging away the first component A_1, the anti-self-duality equations [3] reduce to the so-called Nahm's equations:

$$\frac{\mathrm{d}T_k}{\mathrm{d}x_1} + \frac{1}{2} \sum_{j,l} \epsilon_{kjl} [T_j, T_l] = 0, \quad j,k,l = \{2,3,4\} \qquad [16]$$

where each T_k is regarded as a map $\mathbb{R} \to \mathfrak{u}(r)$.

Readers who are interested in monopoles and Nahm's equations are referred to the survey by Murray (2002) and references therein. The best source for Hitchin's equations still are Hitchin's (1987a, b) original papers. A beautiful duality, known as Nahm transform, relates the various reductions of the anti-self-duality equation to periodic instantons; see the survey article by Jardim (2004).

It is also worth mentioning the book by Mason and Woodhouse (1996), where other interesting dimensional reductions of the anti-self-duality equation are discussed, providing a deep relation between instantons and the general theory of integrable systems.

The Instanton Moduli Space

Now fix a rank-r complex vector bundle E over a four-dimensional Riemannian manifold X. Observe that the difference between any two connections is a linear operator:

$$(\nabla - \nabla')(f\sigma) = f\nabla\sigma + \sigma \cdot \mathrm{d}f - f\nabla'\sigma - \sigma \cdot \mathrm{d}f$$

$$= f(\nabla - \nabla')\sigma$$

In other words, any two connections on E differ by an endomorphism-valued 1-form. Therefore, the set of all smooth connections on E, denoted by $\mathcal{A}(E)$, has the structure of an affine space over $\Gamma(\mathrm{End}(E)) \otimes \Omega^1_M$.

The gauge group $\mathcal{G}(E)$ acts on $\mathcal{A}(E)$ via conjugation:

$$g \cdot \nabla := g^{-1} \nabla g$$

We can form the quotient set $\mathcal{B}(E) = \mathcal{A}(E)/\mathcal{G}(E)$, which is the set of gauge equivalence classes of connections on E.

The set of gauge equivalence classes of anti-self-dual connections on E is a subset of $\mathcal{B}(E)$, and it is called the moduli space of instantons on $E \to X$. The subset of $\mathcal{M}_X(E)$ consisting of irreducible anti-self-dual connections is denoted $\mathcal{M}^*_X(E)$.

Since the choice of a particular vector bundle within its topological class is immaterial, these sets are usually labeled by the topological invariants (Chern or Pontrjagyn classes) of the bundle E. For instance, $\mathcal{M}(r,k)$ denotes the moduli space of instantons on a rank-r complex vector bundle $E \to X$ with $c_1(E) = 0$ and $c_2(E) = k > 0$. It turns out that $\mathcal{M}_X(E)$ can be given the structure of a Hausdorff topological space. In general, $\mathcal{M}_X(E)$ will be singular as a differentiable manifold, but $\mathcal{M}^*_X(E)$ can always be given the structure of a smooth Riemannian manifold.

We start by explaining the notion of a L^2_p vector bundle. Recall that $L^2_p(\mathbb{R}^n)$ denotes the completion of the space of smooth functions $f : \mathbb{R}^n \to \mathbb{C}$ with respect to the norm:

$$\|f\|^2_{L^2_p} = \int_X (|f|^2 + |\mathrm{d}f|^2 + \cdots + |\mathrm{d}^{(p)} f|^2)$$

In dimension $n = 4$ and for $p > 2$, by virtue of the Sobolev embedding theorem, L^2_p consists of continuous functions, i.e., $L^2_p(\mathbb{R}^n) \subset C^0(\mathbb{R}^n)$. So we define the notion of a L^2_p vector bundle as a topological vector bundle whose transition functions are in L^2_p, where $p > 2$.

Now for a fixed L^2_p vector bundle E over X, we can consider the metric space $\mathcal{A}_p(E)$ of all connections on E which can be represented locally on an open subset $U \subset X$ as a $L^2_p(U)$ 1-form. In this topology, the subset of irreducible connections $\mathcal{A}^*_p(E)$ becomes an open dense subset of $\mathcal{A}_p(E)$. Since any topological vector bundle admits a compatible smooth structure, we may regard L^2_p connections as those that differ from a smooth connection by a L^2_p 1-form. In other words, $\mathcal{A}_p(E)$ becomes an affine space modeled over the Hilbert space of L^2_p 1-forms with values in the endomorphisms of E. The curvature of a connection in $\mathcal{A}_p(E)$ then becomes a L^2_{p-1} 2-form with values in the endomorphism bundle $\mathrm{End}(E)$.

Moreover, let $\mathcal{G}_{p+1}(E)$ be defined as the topological group of all L^2_{p+1} bundle automorphisms. By virtue of the Sobolev multiplication theorem, $\mathcal{G}_{p+1}(E)$ has the structure of an infinite-dimensional

Lie group modeled on a Hilbert space; its Lie algebra is the space of L^2_{p+1} sections of $\mathrm{End}(E)$.

The Sobolev multiplication theorem is once again invoked to guarantee that the action $\mathcal{G}_{p+1}(E) \times \mathcal{A}_p(E) \to \mathcal{A}_p(E)$ is a smooth map of Hilbert manifolds. The quotient space $\mathcal{B}_p(E) = \mathcal{A}_p(E)/\mathcal{G}_{p+1}(E)$ inherits a topological structure; it is a metric (hence Hausdorff) topological space. Therefore, the subspace $\mathcal{M}_X(E)$ of $\mathcal{B}_p(E)$ is also a Hausdorff topological space; moreover, one can show that the topology of $\mathcal{M}_X(E)$ does not depend on p.

The quotient space $\mathcal{B}_p(E)$ fails to be a Hilbert manifold because in general the action of $\mathcal{G}_{p+1}(E)$ on $\mathcal{A}_p(E)$ is not free. Indeed, if A is a connection on a rank-r complex vector bundle E over a connected base manifold X, which is associated with a principal G-bundle. Then the isotropy group of A within the gauge group

$$\Gamma_A = \{g \in \mathcal{G}_{p+1}(E) | g(A) = A\}$$

is isomorphic to the centralizer of the holonomy group of A within G.

This means that the subspace of irreducible connections $\mathcal{A}^*_p(E)$ can be equivalently defined as the open dense subset of $\mathcal{A}_p(E)$ consisting of those connections whose isotropy group is minimal, that is,

$$\mathcal{A}^*_p(E) = \{A \in \mathcal{A}_p(E) | \Gamma_A = \mathrm{center}(G)\}$$

Now $\mathcal{G}_{p+1}(E)$ acts with constant isotropy on $\mathcal{A}^*_p(E)$; hence, the quotient $\mathcal{B}^*_p(E) = \mathcal{A}^*_p(E)/\mathcal{G}_{p+1}(E)$ acquires the structure of a smooth Hilbert manifold.

Remark The analysis of neighborhoods of points in $\mathcal{B}_p(E) \backslash \mathcal{B}^*_p(E)$ is very relevant for applications of the instanton moduli spaces to differential topology. The simplest situation occurs when A is an SU(2) connection on a rank-2 complex vector bundle E which reduces to a pair of U(1) and such $[A]$ occurs as an isolated point in $\mathcal{B}_p(E) \backslash \mathcal{B}^*_p(E)$. Then a neighborhood of $[A]$ in $\mathcal{B}_p(E)$ looks like a cone on an infinite-dimensional complex projective space.

Alternatively, the instanton moduli space $\mathcal{M}_X(E)$ can also be described by first taking the subset of all anti-self-dual connections and then taking the quotient under the action of the gauge group. More precisely, consider the map

$$\begin{aligned} \rho &: \mathcal{A}_p(E) \to L^2_p(\mathrm{End}(E) \otimes \Omega^{2,+}_X) \\ \rho(A) &= F^+_A \end{aligned} \qquad [17]$$

Thus, $\rho^{-1}(0)$ is exactly the set of all anti-self-dual connections. It is $\mathcal{G}_{p+1}(E)$-invariant, so we can take the quotient to get

$$\mathcal{M}_X(E) = \rho^{-1}(0)/\mathcal{G}_{p+1}(E)$$

It follows that the subspace $\mathcal{M}^*_X(E) = \mathcal{B}^*_p(E) \cap \mathcal{M}_X(E)$ has the structure of a smooth Hilbert manifold. Index theory comes into play to show that $\mathcal{M}^*_X(E)$ is finite dimensional. Recall that if D is an elliptic operator on a vector bundle over a compact manifold, then D is Fredholm (i.e., $\ker D$ and $\mathrm{coker}\ D$ are finite dimensional) and its index

$$\mathrm{ind}\ D = \dim \ker D - \dim \mathrm{coker}\ D$$

can be computed in terms of topological invariants, as prescribed by the Atiyah–Singer index theorem. The goal here is to identify the tangent space of $\mathcal{M}^*_X(E)$ with the kernel of an elliptic operator.

It is clear that, for each $A \in \mathcal{A}_p(E)$, the tangent space $T_A\mathcal{A}_p(E)$ is just $L^2_p(\mathrm{End}(E) \otimes \Omega^1_X)$. We define the pairing

$$\langle a, b \rangle = \int_X a \wedge *b \qquad [18]$$

and it is easy to see that this pairing defines a Riemannian metric (the so-called L^2-metric) on $\mathcal{A}_p(E)$.

The derivative of the map ρ in [17] at the point A is given by

$$\begin{aligned} d^+_A &: L^2_p(\mathrm{End}(E) \otimes \Omega^1_X) \to L^2_{p-1}(\mathrm{End}(E) \otimes \Omega^2_X) \\ a &\mapsto (d_A a)^+ \end{aligned}$$

so that for each $A \in \rho^{-1}(0)$ we have

$$T_A\rho^{-1}(0) = \left\{a \in L^2_p(\mathrm{End}(E)) \otimes \Omega^1_X \mid d^+_A a = 0\right\}$$

Now for a gauge equivalence class $[A] \in \mathcal{B}^*_p(E)$, the tangent space $T_{[A]}\mathcal{B}^*_p(E)$ consists of those 1-forms which are orthogonal to the fibers of the principal $\mathcal{G}_{p+1}(E)$ bundle $\mathcal{A}^*_p(E) \to \mathcal{B}^*_p(E)$. At a point $A \in \mathcal{A}_p(E)$, the derivative of the action by some $g \in \mathcal{G}_{p+1}(E)$ is

$$-d_A\ :\ L^2_{p+1}(\mathrm{End}(E)) \to L^2_p(\mathrm{End}(E) \otimes \Omega^1_X)$$

Usual Hodge decomposition gives us that there is an orthogonal decomposition:

$$L^2_p(\mathrm{End}(E) \otimes \Omega^1_X) = \mathrm{im}\, d_A \oplus \ker d^*_A$$

which means that:

$$T_{[A]}\mathcal{B}^*_p(E) = \left\{a \in L^2_p(\mathrm{End}(E) \otimes \Omega^1_X) \mid d^*_A a = 0\right\}$$

Thus, the pairing [18] also defines a Riemannian metric on $\mathcal{B}^*_p(E)$. Putting these together, we conclude that the space $T_{[A]}\mathcal{M}^*_X$ tangent to $\mathcal{M}^*_X(E)$ at an equivalence class $[A]$ of anti-self-dual connections can be described as follows:

$$\begin{aligned} &T_{[A]}\mathcal{M}^*_X(E) \\ &\quad = \left\{a \in L^2_p(\mathrm{End}(E) \otimes \Omega^1_X) \mid d^*_A a = d^+_A a = 0\right\} \end{aligned} \qquad [19]$$

It turns out that the so-called deformation operator $\delta_A = d_A^* \oplus d_A$:

$$\delta_A : L_p^2(\mathrm{End}(E) \otimes \Omega_X^1) \rightarrow L_{p+1}^2(\mathrm{End}(E)) \oplus L_{p-1}^2(\mathrm{End}(E) \otimes \Omega_X^2)$$

is elliptic. Moreover, if A is anti-self-dual then coker δ_A is empty, so that $T_{[A]}\mathcal{M}_X^*(E) = \ker \delta_A$. The dimension of the tangent space $T_{[A]}\mathcal{M}_X^*(E)$ is then simply given by the index of the deformation operator δ_A. Using the Atiyah–Singer index theorem, we have for SU(r) bundles with $c_2(E) = k$:

$$\dim \mathcal{M}_X^*(E) = 4rk - (r^2 - 1)(1 - b_1(X) + b_+(X))$$

The dimension formula for arbitrary gauge group G can be found in Atiyah *et al.* (1978).

For example, the moduli space of SU(2) instantons on $\mathbb{R}^4$ of charge k is a smooth Riemannian manifold of dimension $8k - 3$. These parameters are interpreted as the $5k$ parameters describing the positions and sizes of k separate instantons, plus $3(k-1)$ parameters describing their relative SU(2) phases.

The detailed construction of the instanton moduli spaces can be found in Donaldson and Kronheimer (1990). An alternative source is Morgan's lecture notes (Friedman and Morgan 1998). It is interesting to note that $\mathcal{M}_X^*(E)$ inherits many of the geometrical properties of the original manifold X. Most notably, if X is a Kähler manifold, then $\mathcal{M}_X^*(E)$ is also Kähler; if X is a hyper-Kähler manifold, then $\mathcal{M}_X^*(E)$ is also hyper-Kähler. One expects that other geometric structures on X can also be transferred to the instanton moduli spaces $\mathcal{M}_X^*(E)$.

See also: Characteristic Classes; Finite-Dimensional Algebras and Quivers; Gauge Theoretic Invariants of 4-Manifolds; Gauge Theory: Mathematical Applications; Integrable Systems: Overview; Index Theorems; Moduli Spaces: An Introduction; Solitons and Other Extended Field Configurations; Twistor Theory: Some Applications [in Integrable Systems, Complex Geometry and String Theory].

Further Reading

Atiyah MF, Hitchin NJ, and Singer IM (1978) Self-duality in four-dimensional Riemannian geometry. *Proceedings of the Royal Society of London* 362: 425–461.

Bernard CW, Christ NH, Guth AH, and Weinberg EJ (1977) Pseudoparticle parameters for arbitrary gauge groups. *Physical Review D* 16: 2967–2977.

Bourguignon JP and Lawson HB Jr. (1981) Stability and isolation phenomena for Yang–Mills fields. *Communications in Mathematical Physics* 79: 189–230.

Donaldson SK and Kronheimer PB (1990) *Geometry of Four-Manifolds*. Oxford: Clarendon.

Friedman R and Morgan JW (eds.) (1998) *Gauge Theory and the Topology of Four-Manifolds*. Providence, RI: American Mathematical Society.

Hitchin N (1987a) The self-duality equations on a Riemann surface. *Proceedings of the London Mathematical Society* 55: 59–126.

Hitchin N (1987b) Stable bundles and integrable systems. *Duke Mathematical Journal* 54: 91–114.

Jardim M (2004) A survey on Nahm transform. *Journal of Geometry and Physics* 52: 313–327.

Mason LJ and Woodhouse NMJ (1996) *Integrability, Self-duality, and Twistor Theory*. New York, NY: Clarendon.

Murray M (2002) Monopoles. In: Bouwknegt P and Wu S (eds.) *Geometric Analysis and Applications to Quantum Field Theory*, Progr. Math. vol. 205, pp. 119–135. Boston, MA: Birkhauser.

Large-N and Topological Strings

R Gopakumar, Harish-Chandra Research Institute, Allahabad, India

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Topological strings have been well studied since they were introduced in the early 1990s. Essentially, they are simplified string theories that capture the information about a sector of the full (or "physical") string theory. Thus, while sharing many of the structural features of usual string theory, they hold out the possibility of being amenable to explicit calculations. This is especially true with regard to stringy quantum corrections (the higher genus contributions from the point of view of the string world sheet), which are normally rather intractable in the full physical string theory. This has allowed them to play a useful role in enhancing the understanding of string theory and many of its mysterious quantum properties, such as the various dualities.

In particular, in the last several years, topological strings have served as an important laboratory for testing and understanding the connection between the large-N expansion of gauge theories and closed-string theories. In this article we will sketch how this connection is illustrated in a duality between large-N Chern–Simons gauge theory and closed topological string theories. We will survey the origin and current status of these developments and indicated some of its remarkable mathematical ramifications.

Background

In order to appreciate the conjecture relating the Chern–Simons theory and topological string theories, we need to go back to the seminal work of 't Hooft, who pointed to the connection between the large-N expansion of gauge field theories and string theories.

The starting point is a gauge field theory (with, say, gauge group U(N)), where we take the limit of the rank N of the gauge group to infinity (see Brezin

Notes:

and Wadia (1993) for a collection of papers on the topic). The idea is then to make an expansion in inverse powers of N for various observables such as the free energy and correlation functions. For definiteness, let us take a gauge theory containing only gauge fields A in the adjoint representation of $U(N)$. The quantum theory is (schematically) defined by the path integral

$$Z = \int [\mathcal{D}A] e^{iS(A)} \qquad [1]$$

For now, the action $S(A)$ for the gauge fields is left unspecified. It could be either the usual Yang–Mills functional or of the Chern–Simons form which we describe below. $S(A)$ is normalized in such a way that the gauge coupling constant, denoted by κ, only appears via an overall multiplicative factor of $1/\kappa$.

Then the expression, for instance, for the free energy $F = \ln Z$ has an expansion in a power series in κ, whose individual terms are given by the usual Feynman diagrammatic rules. Namely, we have is a sum over connected vacuum diagrams (those without any external legs) formed from the vertices determined by the action $S(A)$. Even without going into the details of the action, we can write down the dependence on N and κ coming from a diagram with h faces, V vertices, and E edges. Every edge is associated with a propagator (arising from the inverse of the quadratic term in $S(A)$) and thus comes with a weight of κ. Every vertex, coming from the cubic and higher-order terms in $S(A)$, comes with a factor of κ^{-1}. There is a factor of N coming from summing over the color indices that circulate in every loop (face). We thus get a weight of $N^h \kappa^{E-V}$ and so the total contribution to the free energy can be organized as

$$F = \sum_{g=0,h=1}^{\infty} C_{g,h} N^h \kappa^{2g-2+h}$$

$$= \sum_{g=0,h=1}^{\infty} C_{g,h} N^{2-2g} \lambda^{2g-2+h} \qquad [2]$$

Here we have defined $\lambda \equiv \kappa N$, the 't Hooft coupling, as the combination that will be kept fixed when taking the limit of large N. We have also used the fact that $V - E + h = 2 - 2g$, where g is the number of handles of the closed two-dimensional surface one can associate with the Feynman diagram. (It is best to visualize the Feynman diagram as a "fatgraph" which forms the skeleton of a closed Riemann surface.) The coefficients $C_{g,h}$ represent the sum of the contributions from all genus g diagrams with h boundaries and depend on the details of the theory.

We note that the reorganization of the contributions to the free energy is reminiscent of the genus expansion in a string theory. In fact, eqn [2] as it stands looks like an open-string expansion on world sheets with g handles and h boundaries. Indeed, in many cases the gauge theory arises as a limit of an open-string theory. (Recall that a massless nonabelian gauge boson is one of the low-lying excitations of an open-string theory.) So the double expansion in terms of g and h is not too surprising.

However, the interesting conjecture of 't Hooft is in the relation to closed-string theory. Note that the expansion in inverse powers of N depends only on the number of handles g. In fact, $1/N$ seems to play the role of closed-string coupling in that it suppresses higher genus diagrams. The total contribution to a given genus g comes from summing over all the holes h in eqn [2], for example,

$$F = \sum_{g=0}^{\infty} N^{2-2g} F_g(\lambda) \qquad [3]$$

The conjecture is to identify this with a closed-string expansion in which $F_g(\lambda)$ is a closed-string amplitude on a genus g Riemann surface. (In carrying out the sum over the holes, we have assumed the existence of a radius of convergence. This is plausible since the number of planar diagrams ($g = 0$), for instance, grows only exponentially with the number of holes.) The question, since 't Hooft, has been: what is this closed-string theory? In other words, what is the background on which the closed string propagates?

A breakthrough came from Maldacena's identification of the background for the particular case of $U(N)$ $\mathcal{N} = 4$ supersymmetric Yang–Mills theory. His conjecture was that this theory is dual to type IIB closed-string theory on $AdS_5 \times S^5$ with a curvature scale set by λ and with closed-string coupling $\propto \lambda/N$. This proposal passed a number of nontrivial checks and is widely held to be true. It also stimulated the search for closed-string duals to other large-N gauge theories.

In what follows, we explain how the conjecture of 't Hooft has a nice realization in the case of three-dimensional $U(N)$ Chern–Simons gauge theory on S^3. The dual closed-string theory, obtained by summing over the holes, turns out to be the A-model topological string on the (six-dimensional) resolved conifold background. The parameter λ maps into a Kahler parameter in the closed-string

geometry and once again the closed-string coupling is $\propto \lambda/N$.

The Large-N Expansion of Chern–Simons Theory

Nonabelian Chern–Simons theory is based on the following action functional for the U(N) gauge connection A:

$$S_{CS}(A) = \frac{k}{4\pi}\int_M \mathrm{tr}\left(A \wedge \mathrm{d}A + \tfrac{2}{3} A \wedge A \wedge A\right) \quad [4]$$

Here M is a three-dimensional manifold. k is called the level and is integer quantized for the path-integral equation [1] to be single valued. Note that, classically, κ as defined earlier is proportional to $1/k$. One of the nice properties of $S_{CS}(A)$ is that it is independent of the metric on M, unlike the Yang–Mills functional. Thus, it is a prototype of a topological field theory. In fact, the observables in this theory capture topological information about the 3-manifold M.

Witten succeeded in quantizing the Chern–Simons theory by relating its Hilbert space to the space of conformal blocks in the two-dimensional U(N) WZW theory. (for more details on the quantization, *see* Chern-Simons Models: Rigorous Results). Here, merely the answers for various observables in the theory will be quoted. In particular, the free energy for the theory on S^3 can be written in a completely explicit form:

$$Z(S^3, N, k) = \exp F(S^3, N, k)$$

$$= \frac{1}{(N+k)^{N/2}} \prod_{j=1}^{N-1} \left(2 \sin \frac{j\pi}{N+k}\right)^{N-j} \quad [5]$$

One of the features one observes in the quantization is the shift ("finite renormalization") of the effective level from k to $k+N$. This can also be seen in perturbation theory. Consequently, while taking the large-N limit, the natural quantity to be held fixed as the 't Hooft coupling is $\lambda = 2\pi N/(k+N)$.

We can then carry out the 't Hooft expansion in powers of λ and $1/N$, of expressions, for example, for the free energy in eqn [5]:

$$F = \frac{N^2}{2}\left(\log \lambda - \frac{3}{2}\right) - \frac{1}{12}\log N + \zeta'(-1)$$
$$+ \sum_{g=2}^{\infty} \frac{1}{N^{2g-2}} \frac{B_{2g}}{2g(2g-2)}$$
$$+ \sum_{g=0}^{\infty}\sum_{h=2}^{\infty} F_{g,h} \lambda^{2g-2+h} \quad [6]$$

The coefficents $F_{g,h}$ are nonzero only for even h and are given by

$$F_{0,h} = -\frac{2\zeta(h-2)}{(2\pi)^{h-2}(h-2)h(h-1)}$$
$$F_{1,h} = \frac{\zeta(h)}{6(2\pi)^h h}$$
$$F_{g,h} = \frac{2\zeta(2g-2+h)}{(2\pi)^{2g-2+h}} \binom{2g-3+h}{h} \times \frac{B_{2g}}{2g(2g-2)} \quad [7]$$

where the last line is for $g > 1$. B_{2g} are the Bernoulli numbers. The first few terms in eqn [6] are nonperturbative contributions which do not have a Feynman-diagram interpretation. The power series in λ is, on the other hand, of the same form as eqn [2]. In fact, there is an open-string interpretation for these terms which will be considered later.

Given the explicit form of the answer, we can carry out the summation over the holes h. Using some resummation techniques, we find

$$F = \sum_{g=0}^{\infty} \left(-\mathrm{i}\frac{t}{N}\right)^{2g-2} F_g(t) \quad [8]$$

with $t \equiv \mathrm{i}\lambda$ and

$$F_g(t) = \frac{(-1)^g |B_{2g} B_{2g-2}|}{2g(2g-2)(2g-2)!} + \frac{|B_{2g}|}{2g(2g-2)!} \sum_{n=1}^{\infty} n^{2g-3} \mathrm{e}^{-nt} \quad [9]$$

(This expression is for $g > 1$. There are very similar expressions for genus 0 and 1 as well.) With the identification of the string coupling $g_s = -\mathrm{i}t/N$, the $F_g(t)$ actually turn out to be the genus g amplitudes of a closed topological string, in line with the general expectation of the previous section. This is explained in the following.

Topological Strings

Physical strings are defined in terms of a two-dimensional sigma model (the theory on the world sheet) made reparametrization invariant by coupling to two-dimensional gravity. Topological strings are simpler versions of this, where the world-sheet theory is a two-dimensional topological sigma model. The latter is defined in terms of a sigma model (usually with $N = 2$ superconformal symmetry) with an additional twist which drastically cuts down the physical states to a subset of the low-lying modes. There are actually two inequivalent twists

denoted by A and B, respectively, but we will restrict to the A twist in this article. One of the simplifications of the A twisted sigma model is that the path integral localizes to contributions from only holomorphic maps from the world sheet to the target space (which will be taken to be a Calabi–Yau 3-fold). Also, all the observables in the theory depend only on the Kahler parameters of the target space and not the complex structure parameters (*see* Topological Sigma Models as well as the book by Hori *et al.* (2003) for more details).

The topological string theory is defined by an appropriate integration of the observables of the topological sigma model over the moduli space of the world-sheet Riemann surface. For instance, the free energy of the string theory at genus g is given by

$$F_g^{\text{top}}(t) = \int_{\mathcal{M}_g} < \prod_{i=1}^{6g-6} (b, \mu_i) >_X \qquad [10]$$

Here b is one of the reparametrization ghost fields on the world sheet and μ_i are Beltrami differentials. The averaging is with respect to the world-sheet sigma model for the Calabi–Yau target X, as the subscript indicates. We have also shown the dependence of F_g on the Kahler parameters of X, collectively denoted by t. The localization to the holomorphic maps in the path integral implies that $F_g^{\text{top}}(t)$ takes the generic form

$$F_g^{\text{top}}(t) = \sum_{\beta} N_{g,\beta} q^{\beta} \qquad q^{\beta} \equiv \prod_i q_i^{n_i} \qquad [11]$$

Here $q_i = e^{-t_i}$ and n_i are the integer coefficents labeling the element $\beta \in H^2(X)$. This is in the same basis of two cycles of $H^2(X)$ in terms of which the complex Kahler parameters t_i are expressed. (Recall that in string theory the Kahler parameters are complexified because of the presence of an additional 2-form field.) The $N_{g,\beta}$ are the Gromov–Witten invariants for X and are in general rational numbers. For nonzero β, the corresponding terms are often called world-sheet instanton contributions since they correspond to topologically nontrivial maps from the world sheet to 2-cycles in the target space. The all-genus free energy of the topological string is also defined to be

$$F^{\text{top}}(t, g_s) = \sum_{g=0}^{\infty} g_s^{2g-2} F_g^{\text{top}}(t) \qquad [12]$$

with g_s being the string coupling.

Since topological strings are related to physical strings by a twist on the world sheet, it is natural that topological string computations are related to computations in the physical string theory. In fact, as shown by Antoniadis, Gava, Narain, and Taylor as well as Bershadsky, Cecotti, Ooguri, and Vafa, observables such as $F_g^{\text{top}}(t)$ are related to special superpotential terms in the type II string compactification on the Calabi–Yau X. Using duality to M-theory, these answers were reinterpreted by Gopakumar and Vafa in terms of contributions coming from BPS states of wrapped D-branes. This gives a completely different perspective on topological strings. For instance, the all-genus free energy can naturally be reorganized as

$$F^{\text{top}}(t, g_s) = \sum_{g=0}^{\infty} \sum_{\beta} \sum_{d=1}^{\infty} n_{\beta}^{g} \frac{1}{d} \left(2 \sin \frac{d g_s}{2}\right)^{2g-2} q^{d\beta} \qquad [13]$$

where the n_{β}^{g} are integer invariants (Gopakumar–Vafa) since they count the number of BPS states. This will prove to be useful in extracting all-genus answers for topological string amplitudes, which is normally quite difficult using the perturbative definition given earlier.

The Large-N Dual to Chern–Simons Theory

We are now in a position to state the duality (Gopakumar and Vafa 1999) between large-N Chern–Simons theory and topological strings in a precise way. The conjecture is that the closed topological string theory on the S^2 resolved conifold geometry is exactly dual to the U(N) Chern–Simons theory on S^3. The resolved conifold geometry is a noncompact Calabi–Yau 3-fold described by the equation

$$xy - zw = 0 \qquad [14]$$

where the singularity is resolved by a 2-sphere $x = \rho z$, $w = \rho y$. The resulting space can thus be characterized as an $O(-1) + O(-1)$ bundle over P^1. It has a single Kahler parameter t for the nontrivial 2-cycle of the S^2. In addition, the string theory is characterized by the string coupling g_s. These parameters map on the gauge theory side to the 't Hooft parameter λ and N via the dictionary

$$t = i\lambda, \qquad g_s = \frac{\lambda}{N} \qquad [15]$$

This conjecture can be checked by comparing various exact calculations in the Chern–Simons theory with corresponding calculations in the topological string on this conifold background. The use of the duality to M-theory enables us to make exact computations on this side as well. One of the

nontrivial checks of this duality comes from a comparison of the free energies. In eqns [8] and [9], we already have carried out the sum over the holes in the Chern–Simons theory and organized it as a closed-string genus expansion. Note that these expressions are already of the form [11] expected of a closed topological string. One simply has to check that it is indeed that on the S^2 resolved conifold.

In the language of the integer invariants n_β^g, the S^2 resolved conifold is particularly simple. The only nonzero invariant is $n_1^0 = 1$. Physically, this corresponds to a single brane wrapped on the genus-zero S^2. Putting this into eqn [13], and making the expansion in powers of g_s, we find exactly eqn [9] for the genus-g contribution to the free energy. This is quite a remarkable agreement and represents a triumph for the ideas of large-N duality.

Geometric Transitions and Large-N Duality

To understand the reason for this duality a bit better, we utilize an old observation of Witten that Chern–Simons theory is an open topological string theory. As mentioned earlier, the expansion [2] (or [6]) is suggestive of an open-string expansion in terms of handles and holes. Witten observed that open topological strings on the noncompact 3-fold T^*M (with Dirichlet boundary conditions on M for the end points of the string) is Chern–Simons theory on M. In fact, in the modern language of D-branes, we would say that $U(N)$ Chern–Simons theory is the world-volume theory of N D-branes wrapped on M, for the topological A-model on T^*M.

In particular, Chern–Simons theory on S^3 is the theory of branes wrapped on S^3 inside T^*S^3. The latter is the conifold geometry but now deformed by a nonzero size S^3. It is described by the equation

$$xy - zw = \mu \qquad [16]$$

where μ is the deformation which parametrizes the size of the S^3.

The above large-N duality can be considered as an open–closed string duality. Namely, that the theory of open A-model topological strings on the S^3 resolved conifold (with N D-branes) is dual to closed A-model topological strings on the S^2 resolved conifold. Cast in this way, we see that the duality involves a transition in the background geometry in going from the open-string to the closed-string description. The sum over the holes changes the background. The S^3, as it were, shrinks to zero size and a transverse S^2 opens up. This geometric transition makes the connection between the Chern–Simons theory and the closed topological string somewhat less mysterious. Maldacena's conjecture for super Yang–Mills involves a similar passage from D-branes in flat space to a closed-string theory on anti-de Sitter space. In fact, it appears as if the best way to understand 't Hooft's idea in generality is to think of it as an open–closed string duality.

Further Checks and Consequences

The free energy is not the only gauge-invariant observable in Chern–Simons theory. One important class of observables, which played an important role in the connection with knot invariants, are the Wilson loop expectation values. Given a knot $\mathcal{K}$ in S^3, we can define, in terms of an arbitrary representation R of $U(N)$, the trace of the holonomy around the knot averaged with respect to the Chern–Simons path-integral measure:

$$W_R(\mathcal{K}) = < \mathrm{tr}_R \left(P \exp i \oint_{\mathcal{K}} A \right) > \qquad [17]$$

P denotes path ordering. Similarly, we can also define the expectation values of links: products of traces of holonomies around various interlinked paths. The nonperturbative solution of Chern–Simons theory gives exact answers for the expectation values of these Wilson loops. The discussion below is, however, confined to knots.

Since the trace of holonomies is being considered in different representations, it makes sense to study the generating functional

$$Z(U,V) = \sum_R \mathrm{tr}_R(U)\mathrm{tr}_R(V)$$

$$= \exp\left[\sum_{n=1}^{\infty} \frac{1}{n} \mathrm{tr}\ U^n \mathrm{tr} V^n \right] \qquad [18]$$

The source V here is a $U(M)$ matrix, unrelated to the $U(N)$ holonomy U around $\mathcal{K}$. The second equality in [18] follows from use of the Frobenius formula. It was shown by Ooguri and Vafa that this generating functional is the natural object from the point of view of the open–closed string duality.

We have already mentioned that the $U(N)$ Chern–Simons theory can be thought of as the theory of N topological D-branes wrapped on the Lagrangian S^3 cycle inside T^*S^3. For a knot $\mathcal{K}$ in the S^3, we consider another Lagrangian 3-cycle $\hat{C}_{\mathcal{K}}$ in T^*S^3 which intersects the S^3 exactly in $\mathcal{K}$. A canonical construction for $\hat{C}_{\mathcal{K}}$ is

$$\hat{C}_{\mathcal{K}} = \left\{ (q(s), p) \in T^*S^3 \,|\, \sum_i p_i \dot{q}_i = 0 \right\} \qquad [19]$$

where the knot $\mathcal{K}$ is parametrized by the closed curve $q(s)$. By construction, $\hat{C}_{\mathcal{K}}$ intersects the S^3 in $\mathcal{K}$. Now consider M D-branes wrapped on $\hat{C}_{\mathcal{K}}$. One now has to consider the fields coming from the strings stretching between the two sets of branes. One can show that integrating out these fields (which are in the bifundamental of the product group $U(N) \times U(M)$) modifies the original Chern–Simons action to

$$S_{\text{eff}}(A) = S_{\text{CS}}(A) + \sum_{n=1}^{\infty} \frac{1}{n} \text{tr} U^n \text{tr} V^n \qquad [20]$$

Here V is the holonomy around $\mathcal{K}$ of the $U(M)$ gauge field $\tilde{A}$. Thus, this configuration of M probe branes gives rise exactly to the generating function eqn [18] for Wilson loops of $\mathcal{K}$.

The geometric transition which relates the Chern–Simons theory to the closed-string theory now suggests what one needs to do to compute this generating function on the closed-string side. We have to follow the configuration of the M probe branes on $\hat{C}_{\mathcal{K}}$ through the conifold transition in which the S^3 shrinks and one blows up the S^2. It is not easy in general to figure out the Lagrangian cycle $C_{\mathcal{K}}$ which results from following $\hat{C}_{\mathcal{K}}$ through the transition. It has only been done in a class of knots including the simple unknot. But assuming we know $C_{\mathcal{K}}$, the generating function for Wilson loops is given by the free energy on the S^2 resolved conifold in the presence of M probe branes on $C_{\mathcal{K}}$. This requires one to know more than the closed-string partition function computed earlier. We now also need to compute amplitudes for world sheets with boundary on $C_{\mathcal{K}}$. These are called open-string Gromov–Witten invariants and the study of this subject is in its infancy. For simple knots such as the unknot, for which $C_{\mathcal{K}}$ is known, these can be computed. One finds again a remarkable agreement with the nonperturbative answers of Chern–Simons theory. Thus, the computation of knot invariants gets related to open-string Gromov–Witten invariants. There have been a number of other tests involving more general knots and links. One also has to be careful of subtleties such as in the choice of framing. The reader is referred to the articles by Marino (2002, 2004) for these topics.

Conclusions

The large-N duality of 't Hooft is realized in Chern–Simons theory in a very explicit way. Thanks to the analytic control we have over both Chern–Simons theory as well as closed topological strings, the conjecture passes very nontrivial checks that extend to all-genus case. This is more than we can do in the AdS/CFT conjecture where most computations are at tree level in the supergravity limit. In contrast, here we see the essential stringiness of the closed-string dual to Chern–Simons theory.

Also, by viewing it as an open–closed string duality, many aspects of the correspondence were clarified. It, therefore, provides a useful toy model for a general understanding of open–closed string duality. Indeed, a proof of this duality using world sheet techniques has been proposed by Ooguri and Vafa. One would like to carry over some of the intuition that operates in this duality to the case of other physically interesting gauge theories.

From the mathematical point of view, as already indicated, this duality leads to previously unsuspected relations between Gromov–Witten invariants and invariants of 3-manifolds, including those of knots. In fact, by considering more general geometric transitions and using this duality locally, one can learn about all-genus topological string amplitudes for a wide class of noncompact toric geometries. This line of development culminated in the formulation of the topological vertex by Aganagic, Klemm, Marino, and Vafa, which captures the essence of the topological closed-string amplitudes for noncompact toric geometries. As in the case of the general correspondence between the gauge theory and gravity, this duality sheds new light on both sides of the equation. We learn to see new integrality properties in knot and 3-manifold invariants which have an interpretation in terms of enumerative problems in 3-folds. The surprises that such a deep connection presages have not yet been exhausted.

See also: AdS/CFT Correspondence; Chern–Simons Models: Rigorous Results; Duality in Topological Quantum Field Theory; Free Probability Theory; The Jones Polynomial; Knot Theory and Physics; Large-N Dualities; Quantum 3-Manifold Invariants; Schwarz-Type Topological Quantum Field Theory; String Field Theory; Topological Gravity, Two-Dimensional; Topological Quantum Field Theory: Overview.

Further Reading

Brezin E and Wadia SR (eds.) (1993) *The Large N Expansion in Quantum Field Theory and Statistical Physics*. Singapore: World Scientific.

Gopakumar R and Vafa C (1999) On the gauge theory/geometry correspondence. *Advances in Theoretical and Mathematical Physics* 3: 1415 (arXiv:hep-th/9811131).

Hori K, Katz S, Klemm A, Pandharipande R, and Thomas R (2003) *Mirror Symmetry*. New York: AMS Publishing.

Marino M (2002) Enumerative geometry and knot invariants (arXiv:hep-th/0210145).

Marino M (2004) Chern–Simons theory and topological strings (arXiv:hep-th/0406005).

Lattice Gauge Theory

A Di Giacomo, Università di Pisa, Pisa, Italy

© 2006 Elsevier Ltd. All rights reserved.

Introduction

As a prototype of lattice gauge theory, quantum chromodynamics (QCD) will be considered in this article. All statements about QCD can easily be extended to other theories, with different gauge group and different content of particles.

QCD is a gauge theory with gauge group SU(3) (color group), coupled to spin-1/2 particles (quarks) belonging to the fundamental representation of the color group. There exist in Nature six different species (flavors) of quarks, with masses ranging from $m_{\text{up}} \sim 5\,\text{MeV}$ to $m_{\text{top}} \sim 180\,\text{GeV}$: the values of these masses are determined by other interactions and can be treated as input parameters of the theory as well as the number of quark flavors. In standard notation, the Lagrangian reads

$$L = -\frac{1}{2}\text{tr}(G_{\mu\nu}G_{\mu\nu}) + \sum_f \bar{\psi}_f(\text{i}\not{D} - m_f)\psi_f \qquad [1]$$

The sum runs over the six quark flavors f. $G_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\nu + \text{ig}[A_\mu, A_\nu]$ is the field strength tensor, $A_\mu = \sum T^a A^a_\mu$ the (gluon) gauge field, $T^a (a = 1, \ldots, 8)$ are the eight generators of the gauge group in the fundamental representation, normalized as $\text{tr}(T^a T^b) = (1/2)\delta^{ab}$. ψ_f is a color triplet of fields. Under a gauge transformation $U(x)$,

$$\psi_f(x) \to \psi'_f(x) = U(x)\psi_f(x) \qquad [2]$$

$$A_\mu(x) \to A'_\mu(x) = U(x)A_\mu U^\dagger(x) + \text{i}U(x)\partial_\mu U^\dagger(x) \qquad [3]$$

D_μ is the covariant derivative of ψ

$$D_\mu\psi_f = (\partial_\mu - \text{ig}A_\mu)\psi_f \qquad [4]$$

and transforms like ψ_f by construction.

L is invariant under the gauge transformation equations [2] and [3]. As a consequence of gauge invariance, the theory has one single coupling constant g.

To make connection with the observations, one has to solve the theory, that is, one has to construct a Hilbert space on which the fields act as operators obeying the equations of motion and the canonical commutation relations. In textbook field theory, this is done by splitting the Lagrangian L into two parts:

$$L = L_0 + L_{\text{I}} \qquad [5]$$

with L_0 the part of L which is bilinear in the fields and L_{I} the rest. L_0 can be solved exactly since it describes free particles and the corresponding equations of motion are linear. The resulting Hilbert space is the Fock space of free particles. L_{I} is treated as a perturbation producing scattering between the fundamental particles. This approach works well in quantum electrodynamics, where the observed particles (electrons and photons) coincide with the excitations of the fundamental fields of the Lagrangian.

In QCD, the fundamental excitations (the quarks and the gluons) are observed as particles neither in Nature nor as a product of high-energy collisions between elementary particles. This feature is known as confinement of color. The conjecture is that excitations with nontrivial color are forbidden to propagate as free particles. However, if hadrons are probed at short distances by photons or by leptons, everything works as if they were composite states of quarks. The accepted explanation relies on asymptotic freedom: the effective coupling constant becomes small at short distances (high momentum transfers) and the constituents behave as free particles.

At large distances, the fundamental excitations are not observed, the interaction is strong and the perturbative picture describing scattering between quarks and gluons is not adequate for the real world.

An alternative quantization procedure is needed which does not rely on perturbation theory. A formally exact quantization procedure is the Feynman path integral. The solution of the theory is given in terms of a functional integral $Z[J]$, which generates the correlators of the fields in the ground state (vacuum). Indicating symbolically the Lagrangian coordinates, namely the fields, by a single symbol Φ, one has

$$Z[J] = \int \prod_x \text{d}\Phi(x) \exp\left[-S[\Phi] - \int J(x)\Phi(x)\text{d}x\right] \qquad [6]$$

The connected Euclidean vacuum correlators are given in terms of functional derivatives of $Z[J]$

$$< 0|T(\Phi(x_1)\Phi(x_2)\cdots\Phi(x_n))|0 >_{\text{conn}} = \frac{1}{Z[0]}\frac{\delta^n Z[J]}{\delta J(x_1)\delta J(x_2)\cdots\delta J(x_n)}\bigg|_{J(x)=0} \qquad [7]$$

"Euclidean" means that they are analytic continuations to imaginary times. Going to Euclidean system is necessary to isolate the vacuum state. The amplitudes can be analytically continued back to Minkowski space. The Hilbert space and all the physical observables can be constructed in terms of the correlators, a property known as reconstruction theorem. Formally (i.e., assuming that everything makes sense only if the functional integral exists),

$$< 0|T(\Phi(x_1)\Phi(x_2)\cdots\Phi(x_n))|0 >_{\text{conn}} = \frac{1}{Z}\int \prod_x d\Phi(x) \times \exp(-S[\Phi])\Phi(x_1)\Phi(x_2)\cdots\Phi(x_n) \quad [8]$$

The continuation to imaginary time changes sign to the kinetic energy, and Z formally becomes the partition function of a four-dimensional statistical model with Hamiltonian $S_E[\Phi]$, a general fact in Feynman integrals.

By definition of functional integral, Z is defined by discretizing a finite volume V of spacetime to a finite set of points and then sending their number to infinity, making a set dense in V. If the limit exists, a Z_V is obtained. The volume V is then sent to infinity, to cover the whole spacetime (thermodynamical limit) and Z_V eventually converges to Z. A rigorous proof of the existence of these limits does not exist for QCD, but there are qualitative arguments that this is the case, which will be presented below.

In the lattice formulation of field theory, a regular lattice, usually cubic, is taken as a discretization of spacetime.

From the very definition of Feynman integral, it follows that the formulation of field theory on the lattice is nothing but an approximation to the limit which defines Z. It will provide a good approximation if the lattice spacing is small enough with respect to the physical lengths involved and if the lattice is large compared to them.

Perturbation theory amounts to split the action into a bilinear term S_0 and an interaction term S_I containing the higher powers of the fields. The Z integral is then computed by expanding the weight in a power series of S_I:

$$\int \prod_x d\Phi(x) \exp(-S_0 - S_I) = \int \prod_x d\Phi(x) \exp(-S_0) \sum_n \frac{(-S_I)^n}{n!} \quad [9]$$

The Feynman integral thus becomes Gaussian, can be computed, and gives the usual perturbative expansion. The two limits (integral and series expansion) do not commute in general. For QCD, there are indeed arguments that the renormalized perturbative expansion does not converge and is plagued by singularities known as renormalons.

Wilson's Formulation

For field theories of scalar particles, the lattice discretization is performed by assigning a value of the field to each site of the lattice. The Wilson formulation for gauge theories is not made in terms of the fields A_μ, which are defined in the Lie algebra of the gauge group, but in terms of parallel transports, which are elements of the group itself. The building blocks are parallel transports along links parallel to spacetime axes connecting neighboring sites

$$U_\mu(x) \equiv P \exp\left(ig \int_x^{x+\hat{\mu}} A_\rho dx^\rho\right) \approx \exp(igaA_\mu(x)) \quad [10]$$

where $\hat{\mu}$ indicates the vector of length a in the μ direction and P the ordered product. The last approximate equality is valid in the limit of small lattice spacing a. g is the coupling constant.

Under a gauge transformation $V(x)$,

$$U_\mu(x) \to V(x)U_\mu(x)V^\dagger(x+\hat{\mu}) \quad [11]$$

It follows from eqn [11] that the parallel transport along a closed path is gauge invariant. The density of action can be written in terms of the parallel transport along the elementary square of links in the hyperplanes $\mu\nu$ $\Pi_{\mu\nu}$, known as plaquette:

$$\prod_{\mu\nu} = \text{tr}[U_\mu(x)U_\nu(x+\hat{\mu})U_\mu^\dagger(x+\hat{\nu})U_\nu^\dagger(x)] \quad [12]$$

By expanding in powers of a, one easily finds

$$\prod_{\mu\nu} = N_c - \frac{1}{2}a^4 \text{tr}[G_{\mu\nu}G_{\mu\nu}] + O(a^6) \quad [13]$$

with N_c the number of colors, 3 for QCD. The lattice action can be defined as

$$S = \sum_{x\mu\nu} \beta\left(1 - \frac{1}{N_c}\Pi_{\mu\nu}\right) \quad [14]$$

with $\beta = 2N_c/g^2$, and tends to the continuum action as $a \to 0$, $O(a^2)$. An infinite number of higher-order terms in a exist, which come from the expansion of the links, but they are expected to be irrelevant in the continuum limit $a \to 0$.

The measure of the Feynman integral is assumed to be the Haar measure of the gauge group for each link, which again can be shown to tend to the continuum measure in the continuum limit.

Everything is gauge invariant, contrary to the perturbative formulation, where a gauge fixing is required to define the vector meson propagator.

By Weierstrass theorem, the integral is finite for any finite number of links, the gauge group being compact.

Any other choice of the lattice action differing from the Wilson action of eqn [14] by terms of higher order in a will have the same continuum limit: there is significant freedom in the choice of the action.

In the language of statistical mechanics, the Euclidean lattice formulation is a spin model. Different choices of the action correspond to different spin models. In the vicinity of a second-order phase transition, however, the correlation length becomes large with respect to the lattice spacing and all the irrelevant terms become negligible. All the spin models at the critical point belong to the same universality class and define the same field theory.

This is what happens for QCD because of asymptotic freedom. By renormalization group arguments, the lattice spacing behaves as

$$a(\beta) \approx \frac{1}{\Lambda}\exp(-b_0\beta) \qquad [15]$$

at sufficiently large β, where $-b_0$ is the coefficient of lowest-order term of the β-function, b_0 is positive and Λ is a physical scale. As $\beta \to \infty$, a tends exponentially to zero in physical units and the coarse structure of the lattice becomes unimportant, indicating that the short-distance limit in the definition of the Feynman integral exists. The theory also develops a mass scale Λ which insures the existence of a finite correlation length and hence of the thermodynamical limit. In practice, when β is increased, the lattice space becomes exponentially small in physical units. As a consequence, however, the physical scale becomes exponentially large in lattice units, and an exponentially large lattice is needed to insure the large-distance convergence. This makes life difficult if the Feynman integral has to be computed numerically.

Quarks

Fermion fields are defined on lattice sites. The naive lattice transcription of the fermion term in eqn [1] consists in replacing the covariant derivatives by finite differences with parallel transports to make the result gauge covariant. In principle, $D_\mu\psi(x) = U^\dagger(x)\psi(x+\hat{\mu}) - \psi(x)$ is a correct definition. In practice, a more symmetric difference is used which is correct $O(a^2)$, namely

$$D^{\mathrm{L}}_\mu\psi(x) = \frac{1}{2}[U(x)\psi(x+\hat{\mu}) - U^\dagger(x-\hat{\mu})\psi(x-\hat{\mu})] \qquad [16]$$

The fermionic Lagrangian then reads

$$\sum_x \bar{\psi}(x)[\mathrm{i}\not{D}^{\mathrm{L}} - m]\psi(x) \equiv \sum_{x x' \alpha\beta} \bar{\psi}_\alpha(x) M^{-1}_{\alpha\beta(x,x')}\psi_\beta(x') \qquad [17]$$

It is convenient to indicate this expression in the form $S_f = \bar{\psi}M^{-1}\psi$, where ψ is a large column whose elements are labeled by the site x and by the component α. The functional integral over ψ can explicitly be done by using the standard rules of integration on Grassman variables, since the action is bilinear,

$$Z = \int \prod \mathrm{d}U_\mu(x)\mathrm{d}\psi(x)\mathrm{d}\bar{\psi}(x) \times \exp(-S_{\mathrm{E}}[U] - \bar{\psi}M\psi) \qquad [18]$$

The result is

$$Z = \int \prod \mathrm{d}U_\mu(x) \exp(-S_{\mathrm{E}}[U]) \det M \qquad [19]$$

The effect of fermions is to multiply the weight by a functional determinant which depends on the gauge field configuration.

A problem exists, however, in this procedure already at the level of free fermions, that is, putting $U=1$ in the action and in the determinant of eqn [18]. The equation of motion reads, in Fourier transform,

$$\sum_\mu \left[\gamma_\mu \sin\left(2\pi\frac{k_\mu}{L}\right) - m\right]\tilde{\psi}(k) = 0 \qquad [20]$$

With respect to the continuum, the momentum $p_\mu = 2\pi k_\mu/L$ has been replaced by its sinus. At small values of p_μ, eqn [20] coincides with the Dirac equation. However, an alternative solution exists at $p_\mu \approx \pi$, for each μ independently. The new equation differs from the other by a change of sign of γ_μ. Changing sign of one of the gammas means changing sign to $\gamma^5 \equiv \gamma^1\gamma^2\gamma^3\gamma^4$, which is the chirality of the fermion. Instead of one fermion, we then have $2^4 = 16$ fermion species, organized in pairs with opposite chiralities. It is impossible to have a single fermion with a given chirality. A number of recipes have been proposed to circumvent this artifact of the lattice regulation, for example, introduce by hand a term in the action which removes the spurious particles in the limit of zero lattice spacing (Wilson's fermions); double the lattice spacing by constructing two sublattices on even and odd sites, respectively, which propagate fermions of opposite chirality (staggered fermions),

so that the argument of the sinus in the derivative is doubled. More recently, an idea which goes back to Ginsparg and Wilson has been implemented, which consists in replacing a strictly local equation of motion like eqn [20] by an equation with the same continuum limit which is nonlocal, but with a nonlocality falling off exponentially at large distances, a recipe which makes propagation of chiral fermions possible. This is an important improvement, even if very demanding in computer power.

Numerical Simulations

Solving analytically the lattice version of QCD would allow one to follow constructively all the steps which bring to the definition of Z, that is, the ultraviolet and the infrared limit, as explained earlier. Presently that is out of reach. Also an attempt by Wilson to solve the lattice renormalization group equations by techniques of decimation is not conclusive.

The problem can be attacked numerically. One way would be to compute the integral numerically. That is, however, prohibitive: it would be like solving exactly the equations of motion for the molecules of a gas. The lattice theory is in fact a four-dimensional statistical mechanics with the Boltzmann factor $\beta = 2N_c/g^2$ and Hamiltonian equal to the Euclidean action. As in statistical mechanics the way out is to create a significant sample of configurations with weight $\exp(-\beta S_E)$ and to determine the field correlators which describe physics by an average on this ensemble. This is done by Monte Carlo techniques.

The basic principle is to start from an arbitrary field configuration and make a sequence of random changes, normally on a single link at a time, with uniform probability in the group measure so as to converge toward the equilibrium distribution $\exp(-\beta S_E)$. For that purpose, the probability $P_{C'C}$ to change from a configuration C to another C' is constrained to obey the detailed balance relation

$$P_{C'C} \exp(-\beta S[C]) = P_{CC'} \exp(-\beta S[C']) \qquad [21]$$

A common algorithm is known as metropolis. The way to implement the condition (eqn [21]) is to accept the new trial configuration C' if $S[C'] \leq S[C]$, and to accept it with probability $\exp(-\beta[S(C') - S(C)])$ if $S[C'] \geq S[C]$. An alternative method is known as "heat-bath". If the probability of the configuration for one link at a fixed value of the other variables is explicitly known, the change can be accepted with that probability.

In the presence of dynamical quarks, the integral eqn [18] is converted into an integral on bosonic variables by inverting the matrix M:

$$Z = \int \prod \mathrm{d}U_\mu(x)\, \mathrm{d}\phi(x)\, \mathrm{d}\phi(x)^\dagger \times \exp(-S_E[U] - \phi^\dagger [M^\dagger M]^{-1} \phi) \qquad [22]$$

The property has been used such that $\int \prod \mathrm{d}\phi(x)\, \mathrm{d}\phi^\dagger(x)\, \exp(-\phi^\dagger [M^\dagger M]^{-1} \phi) = |\det M|$. A metropolis updating is then performed on the combined U_μ and ϕ variables. To have a choice of the trial uniform in the measure, an algorithm is commonly used which is based on ergodicity, known as hybrid molecular dynamics. A fictitious conjugate momentum is associated with all variables, and a fictitious Hamiltonian is defined by adding to the action, considered as a potential energy, the sum of the squares of the conjugate momenta. A classical evolution is then performed in time by small steps which should displace the state in phase space ergodically: the evolution is called a trajectory. After a number of steps, a metropolis test is made as explained above.

Typically, the computer time needed to produce a significant configuration is proportional to the volume V of the lattice for pure gauge systems, to $V^{5/4}$ in the hybrid algorithm for full QCD.

As explained before, in order to have a good approximation to the Feynman integral the lattice spacing has to be small compared to the physical scales, for example, with respect to the Compton wavelength of the heaviest quark. On the other hand, to control volume effects it has to be large compared to the biggest physical length, for example, with respect to the Compton wavelength of the lightest quark. Since there is a factor $m_{\mathrm{top}}/m_{\mathrm{up}} \approx 3 \times 10^3$ between these two lengths, the lattice size needed would be prohibitive from numerical point of view. In practice, lattices of size L^4 are affordable with $L \leq 64 - 128$. For this reason, only the light quarks u, d, s are kept, which have mass smaller than the typical scale of the theory, which can be identified as the square root of the string tension. In the limit in which light quark masses are small compared to QCD scale, the Lagrangian is invariant under any unitary mixing of them. A global SU(3) invariance exists, which is known as flavor symmetry, and is broken by the difference of quark masses. Heavier quarks can be described by an effective theory, since they have negligible dynamical effects at low energies.

A Selection of Physics Results

String Tension

A big excitement followed the first numerical calculations by M Creutz at the beginning of the 1980s in which the static potential $V(r)$ between a quark and an antiquark was computed in pure-gauge theory on the lattice. One way to measure it is to measure the correlator of two Polyakov lines at a distance r on a significant ensemble of field configurations. The Polyakov line is the parallel transport in the fundamental representation along the time axis across the lattice: with periodic boundary conditions it is a closed loop, and hence it is gauge invariant. It can be proved that the log of this correlator is equal to $-V(r)aL_t$ with L_t the extension of the lattice in the time direction. It was found that

$$V(r) = \sigma r \qquad [23]$$

The parameter σ is known as string tension. A potential of the form eqn [23] means confinement: an infinite amount of energy is required to pull apart the particles at infinite distance. The parameter σ can be determined phenomenologically from the mass spectrum of the mesons and $\sigma 2\pi \approx 1\,\text{GeV}$. What is measured on the lattice is

$$\sigma a(\beta)^2 n^2 \qquad [24]$$

where n is the distance of the two Polyakov lines in lattice spacings and $a(\beta)$ the lattice spacing in physical units. In fact, the computer only produces pure numbers. If the lattice QCD belongs to the same universality class of QCD at the critical point, that is, if the lattice really defines QCD, the dependence of $a(\beta)$ on β is dictated by the β-function of the renormalization group. At sufficiently large $\beta = 6/g^2$,

$$a(\beta) \approx \frac{1}{\Lambda_{\text{latt}}} \exp(-b_0\beta) \qquad [25]$$

with $b_0 = (11/3)N_c/16\pi^2$. Λ_{latt} is the energy scale of the theory. The measurement of the potential gives indeed a dependence of the lattice spacing on β consistent with eqn [25] and allows one to determine $\sigma/\Lambda^2_{\text{latt}}$. The absolute value of the lattice spacing can be determined by comparison with the physical value of the string tension. The theory is able to produce a physical scale. The correlation length is finite and as a consequence the infrared limit of the Feynman integral exists.

Mass Spectrum

Any operator with the quantum numbers of a particle can be used as interpolating field for it. The correlator of the operator at large distances behaves like a sum of exponentials $\exp(-mr)$ with m the masses of the particles with the same quantum numbers. At large distances the lightest particle dominates, especially if the operator has a good overlap, that is, if its matrix element between vacuum and the state of the particle is the biggest. From the correlators mr can be determined. On the lattice $r = na(\beta)$ so that, by eqn [25] what is really determined is the ratio m/Λ_{latt}. If Λ_{latt} has been determined, for example, from the string tension, the mass of the particle results in physical units. Alternatively, the ratios of any two masses can be determined and the scale fixed by the value of one of them. A good agreement is obtained already in pure gauge (quenched approximation) indicating that the quark loops are relevant at the level of 10% typically. This fact supports the idea that the large N_c-limit is a good approximation to reality, quark loops being nonleading in that limit. The light particle masses are more difficult to compute, being sensitive to the masses of light quarks which cannot be taken at realistic values due to computational difficulties: large lattices are required and big fluctuations are present near the chiral point. The spectrum of particles made of heavy quarks can be computed using effective theories, and nicely fits experiment. A byproduct is a precise determination of the gauge coupling constant, competitive with phenomenological determinations from short distance perturbative QCD.

Weak Interaction Matrix Elements

There exist matrix elements of currents (or products thereof) entering in weak amplitudes which involve large distances and are not computable in perturbation theory. Lattice can be used to evaluate them. Renormalization problems can appear in this approach when the cutoff is removed, which, however, are not difficulties of principle but only of technical nature. This activity is of fundamental importance to have precise predictions in order to understand the limits of the standard model.

Finite-Temperature QCD and the Deconfinement Transition

The static thermodynamics of a system of fields is described by the partition function

$$Z_T = \text{tr}[\exp(-H/T)] \qquad [26]$$

It is easy to show that Z_T is equal to the Euclidean Feynman integral on the imaginary time interval $(0, 1/T)$ with boundary conditions in time periodic for bosons and antiperiodic for fermions. Indeed, the

Boltzmann factor is formally an imaginary time evolution by $1/T$. A lattice of extension $L_t L_S^3$ with $L_s \gg L_t$ provides the partition function at a temperature $T = 1/aL_t$, if a is the lattice spacing in physical units.

Finite-temperature simulations are important to investigate the transition from the phase in which color is confined to a phase in which quarks and gluons can propagate as free particles. This phase is called deconfined phase or quark gluon plasma.

Big experiments at Brookhaven and at CERN are looking for this phase transition in high-energy collisions between heavy nuclei, but no definite evidence has yet been produced for it. Lattice simulations instead definitely prove that such a transition exists. For pure SU(3) gauge theory (quenched) at $T \approx 270\,\text{MeV}$, a first-order phase transition is observed, at which the string tension vanishes. In a more realistic theory with dynamical quarks, a transition is also observed at $T \approx 160\,\text{MeV}$, where chiral symmetry, which is spontaneously broken at zero temperature, is restored. This transition is also associated to deconfinement even if, in the presence of light quarks, the string tension does not exist. Indeed, when pulling apart a quark and an antiquark, an instability for production of quark–antiquark pairs sets in when the potential energy becomes large enough, which physically manifests itself as a production of light mesons. An alternative order parameter is needed. The possibility of defining alternative order parameters is discussed in next section.

The equation of state can also be studied relating internal energy to pressure, which is useful to understand heavy ion collisions.

From the features of the deconfinement transition, information can be extracted on the mechanisms by which QCD confines color.

A connected issue is the behavior of QCD at nonzero baryon density or chemical potential. The corresponding thermodynamics is described by a grand canonical ensemble

$$Z_\mu = \text{tr}[\exp[-(H + \mu N)/T]] \qquad [27]$$

where $N = \int \mathrm{d}^3 x \psi^\dagger \psi$ is the baryon number operator and μ the chemical potential. In the process of converting the partition function Z_μ into a Feynman integral, the term H at the exponent of eqn [27] generates the Euclidean action, which is real. The term proportional to N becomes imaginary. The integral is well defined, but the analogy with a four-dimensional statistical mechanics is broken, the effective Hamiltonian being non-Hermitian and no sampling can be made. Approximate methods have been developed, but the problem is open. Exploring numerically the region of phase space with $\mu \neq 0$ would be interesting, since a rich structure is expected, which could be relevant to dense systems such as neutron stars.

Mechanisms of Color Confinement

Understanding how QCD manages to confine color is one of the most fascinating problems in field theory.

To prove confinement, one should, in principle, prove that, at zero temperature, no gauge-invariant quasilocal operator exists, carrying nontrivial color and obeying cluster property at large distances. This proof is not known. There exists evidence form lattice simulations that a string tension exists, as discussed before. In any case, a guess can be made of the physical mechanism of confinement. If confinement is an absolute property reflecting a symmetry property of the vacuum, an order parameter should exist which discriminates between confined and deconfined phase, and the transition between the two phases has to be a true transition. Observing a crossover in some part of the boundary between the two phases would disprove this view. A lattice determination of the order of the deconfining transition is therefore of fundamental importance.

A possible mechanism of confinement proposed by G't Hooft is dual superconductivity of the vacuum: dual means interchange of electric with magnetic with respect to ordinary superconductors. In the same way as the magnetic field is constrained into Abrikosov flux tubes in an ordinary superconductor, the chromoelectric field acting between a quark and an antiquark would be constrained into flux tubes by a dual Meissner effect producing an energy proportional to the distance, or a string tension.

This mechanism can be investigated by lattice simulations, by checking if any magnetically charged operator exists whose vacuum expectation value is nonzero in the confined phase signaling condensation of magnetic charges and zero in the deconfined phase. Progress has been made in this direction which, however, is not yet conclusive. Chromoelectric flux tubes between q–$\bar{q}$ pairs are observed in lattice field configurations.

Topology

Euclidean QCD admits classical solutions with finite action and with a nontrivial topology which makes them stable. These solutions, known as instantons or multi-instantons, realize a mapping of the three-dimensional sphere at infinity on the gauge group, and the topological charge is the winding number of this mapping. The Jacobian of this mapping is the Chern

current K_μ and its divergence $\partial_\mu K_\mu(x) \equiv Q(x)$ is the density of topological charge. $Q = \int d^4x\, Q(x)$ is the topological charge which has integer values. Explicitly,

$$Q(x) = \frac{g^2}{16\pi^2}\mathrm{tr}[G_{\mu\nu}G^*_{\mu\nu}] \qquad [28]$$

with $G^*_{\mu\nu} = (1/2)\epsilon_{\mu\nu\rho\sigma}G_{\rho\sigma}$ the dual field strength tensor.

$Q(x)$ plays an important role in hadron physics, being related to the anomaly of the flavor singlet axial current $J^5_\mu = \sum_f \bar{\psi}\gamma^5\gamma_\mu\psi_f$. J^5_μ is conserved at the classical level in the chiral limit $m_f = 0$, but this symmetry does not survive quantization. In fact,

$$\partial_\mu J^5_\mu = 2N_f Q(x) \qquad [29]$$

A consequence of eqn [29] is the high mass $m_{\eta'} \approx 1\,\mathrm{GeV}$ of the flavor singlet partner η' of the pseudoscalar flavor octet. An $N_c \to \infty$ argument by Witten and Veneziano relates $m_{\eta'}$ to the response of the quenched (no quark) vacuum to topological excitation, the topological susceptibility $\chi \equiv \int d^4x < 0|TQ(x)Q(0)|0 >$. The relation is

$$\frac{2N_f}{f_\pi^2}\chi = [m^2_{\eta'} + m^2_\eta - 2m^2_K][1 + O(1/N_c)] \qquad [30]$$

This approximate relation has been checked on the lattice. χ has been determined by different methods which agree in confirming it. This is an important verification of QCD.

Instantons are stable solutions in the continuum, approximately stable in the lattice discretized version. A cooling procedure which locally freezes short-distance quantum fluctuations would leave the instantons untouched if they were stable. On the lattice the instanton is stable anyhow if the distance in correlation reached by the local cooling procedure is small compared to the size of the instanton: cooling is indeed a diffusion process and the distance involved grows as the square root of the number of cooling iterations. Instanton configurations can nicely be exposed by cooling.

See also: Anomalies; Quantum Chromodynamics; Renormalization: General Theory; Spin Foams; Symmetry Breaking in Field Theory.

Further Reading

Creutz M (1983) *Quarks, Gluons and Lattices.* Cambridge: Cambridge University Press.

Feynman RP (1948) Space-time approach to nonrelativistic quantum mechanics. *Reviews of Modern Physics* 20: 367.

Feynman RP and Hibbs AR (1965) *Quantum Mechanics and Path Integral.* New York: McGraw-Hill.

Ginsparg PH and Wilson KG (1982) A remnant of chiral symmetry on the lattice. *Physical Review D* 25: 2649.

Gottlieb S, Liu W, Toussaint D, Renken RL, and Sugar RL (1987) Hybrid-molecular-dynamics algorithms for the numerical simulation of QCD. *Physical Review D* 35: 2531.

Kogut J and Susskind L (1975) Hamiltonian formulation of Wilson's lattice gauge theories. *Physical Review D* 11: 395.

't Hooft G (1981) Topology of the gauge condition and new confinement phases in non-abelian gauge theories. *Nuclear Physics B* 190: 455.

Veneziano G (1979) U(1) without instantons. *Nuclear Physics B* 159: 213.

Wilson KG (1974) Confinement of quarks. *Physical Review D* 10: 2445–2459.

Wilson KG (1983) The renormalization group and critical phenomena (Nobel Lecture). *Reviews of Modern Physics* 55: 583.

Witten E (1979) Current algebra theorems for the U(1) goldstone boson. *Nuclear Physics B* 156: 269.

Measure on Loop Spaces

H Airault, Université de Picardie, Amiens, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Loop spaces have been considered for their geometric interest (Freed Daniel 1988) where the space of based loops on a compact Lie group is endowed with a Kählerian structure; see also the survey by L Gross (1988). The harmonic analysis on loop groups, developed by Pressley and Segal, is reviewed by Hsu (1997). Loop groups have also an impact in string theory (Bowick and Rajeev 1987). They are related to Yang–Mills theory (Levy 2003). A presentation of the history of measure on infinite-dimensional spaces has been given by P Malliavin (see Malliavin (1992) and references therein). The main problem is the construction of measures on the loop space which have quasi-invariance property. This has implications in representation theory (Neretin 1994, Jones 1995). Here we mainly concentrate on the nonlinear stochastic point of view and its interference with geometry. The geometrical study of the space of closed curves over a compact Riemannian manifold M, that is, the loop space over M, was initiated by Marston Morse in 1932. The loop space is itself a manifold where one can define a Laplace–Beltrami operator. A diffusion process can be considered on this manifold. Wiener defined the Brownian loop by the Fourier series

$$u(\tau) = \sum_{k \geq 1} \frac{\sin k\pi\tau}{k} G_k \qquad [1]$$

where the G_k are independent normal variables. The time evolution of the Wiener loop and the extension of the theory to the case of a compact Riemannian manifold of finite dimension has been considered by Airault and Malliavin (1996, and references therein). The Brownian loop evolutes in the time parameter t as a Brownian sheet where the independent random variables G_k are function of t.

Starting from the zero loop, one obtains at time t, a random loop, and the law of this loop gives a measure on the loop space. A construction of this measure with functional analysis on infinite-dimensional manifold was done by Gaveau and Mazet (1979). The tools of stochastic analysis are important to the subject. The loop space of continuous maps from the circle to the multiplicative group of complex numbers has a group structure, hence the term "loop group." On the loop group, we consider the multiplicative Brownian motion starting at one point of the circle and conditioned to come back at this point at time s. It defines a probability measure on the loop group. One can also consider the set of continuous maps from the circle to the set of complex numbers of modulus equal to 1. The loop group is the space of continuous closed paths on a Lie group. More generally, on a Riemannian manifold M, the Brownian motion on M defines a Wiener measure on the loops over M. To go from the path space to the loop space, an important tool is the quasisure analysis in infinite dimension. The quasisure analysis was developed by Airault and Malliavin (1996, and references therein) to obtain disintegrations of the Wiener measure and they have used this tool in 1992 to construct measures on the loop group. The main problems are:

1. The construction of heat kernel measures and the existence of a Brownian motion on the loop space, the existence of pinned Wiener measures obtained as the law of Brownian motions conditioned on the loops.
2. The quasi-invariance of these transition probability measures under translation, or multiplication if we have a multiplicative structure, or under the infinitesimal action of suitable vector fields. For the path space over the n-dimensional Euclidean space R^n, the Cameron–Martin theorem (1944) ensures the existence of a density which shows the quasi-invariance of the Wiener measure under translations. For the quasi-invariance, an important fact is the choice of the metric on the Cameron–Martin space. In the case of the Wiener measure, one considers the paths of finite energy, $\int_0^1 |h'(s)|^2 ds < +\infty$. This corresponds to the metric "1." P Malliavin (1989, and references therein) discussed the case of metrics α with $1/2 < \alpha < 1$.
3. To define the "good" Cameron subspace, that is, find the vector fields that yield integration-by-parts formulas. The question occurs whether the Cameron–Martin space depends on time. For the loop space, it has been proved by Driver (2003) that it is not the case. A time evolution of the tangent Cameron–Martin space could appear eventually.
4. The determination of the support of the measures (e.g., the Wiener measure) is carried by the set of Hölder functions of order $1/2 - \epsilon$.
5. The absolute continuity of the measures with respect to each other.

The Construction of Heat Measures on the Loop Space and Their Quasi-Invariance

The construction of measures giving a solution to the infinite-dimensional heat equation as well as the study of the quasi-invariance of the Wiener measure on the path space was started extensively in the work by Bismut, followed by Gross (1998), then by Aida and Elworthy (1995) where the loop group is a suitable manifold to extend to infinite-dimensional manifolds the log-Sobolev inequalities, by Malliavin and Malliavin (1992, and references therein) where the measures on the path space and the path group have been studied. Consider a compact Lie group G with unit e and let $\mathcal{G}$ be its Lie algebra. From the $\mathcal{G}$-valued Brownian motion, one can construct a family of measures $(\mu_t^e)_{t\geq 0}$ on the path space. These measures μ_t^e are the images of the Wiener measure on $\mathcal{G}$ through the Ito map

$$dg_x(\tau) = \sqrt{t}\, g_x(\tau) dx(\tau) \quad \text{with } g_x(0) = e \qquad [2]$$

The convolution of two measures μ_t^e and $\mu_{t'}^e$ is equal to $\mu_{t+t'}^e$. By choosing the initial value of the path randomly distributed according to the Haar measure on G, it defines a family of measures $(\mu_t)_{t\geq 0}$ on the path space with

$$\int f(\gamma)\, \mu_t(d\gamma) = \int dg \int f(g\gamma)\mu_t^e(d\gamma)$$

The Laplacian on the path group is defined by

$$(\Delta_P f)(g) = \lim_{\epsilon\to 0} \frac{1}{\epsilon}\left[\int f(g\gamma)\mu_\epsilon(d\gamma) - f(g)\right]$$

The heat equation is valid for the measures $(\mu_t)_{t\geq 0}$ on the paths,

$$\frac{\partial}{\partial t}\int f(g)\mu_t(dg) = \int (\Delta_P f)(g)\mu_t(dg)$$

Moreover, there is a quasi-invariance density $k_{g_0}(g)$ defined on the path group (g_0 and g are paths with values in G) such that

$$\mu_t(g_0 A) = \int_A k_{g_0}(g)\mu_t(dg)$$

where $g_0 A$ is the translated on the left of the subset A in the path space over G. This is a generalization to the path space of the classical Cameron–Martin theorem. Then, one can consider the loop space. The free loop space is the set of continuous maps g from $[0, 1]$ to G such that $g(0) = g(1)$, and the loop space with a base point is the set of maps such that $g(0) = g(1) = m$ is fixed. One can define the pinned Brownian motion on the group G to obtain the pinned Wiener measures $(\mu_t^{L_e})_{t\geq 0}$ on the loop group (Malliavin and Malliavin 1992, Driver and Srimurthy 2001). Denote by $p_t(g)$ the solution of the heat equation on the group G. Let g be a map from $[0, 1]$ to the finite-dimensional Lie group G. For $\tau_1, \tau_2, \ldots, \tau_n \in [0, 1]$, consider the evaluations of the map g, $g_{\tau_1}, g_{\tau_2}, \ldots, g_{\tau_n} \in G$, Let f be a real function defined on G and denote by dg the Haar measure on G. The measure $\mu_t^{L_e}$ on the loop group is given by

$$\int f(g_{\tau_1}, g_{\tau_2}, \ldots, g_{\tau_n})\, d\mu_t^{L_e}(g)$$
$$= \int f(g_1, g_2, \ldots, g_n) p_{t\tau_1}(g_1) p_{t(\tau_2-\tau_1)}(g_1^{-1}g_2)\cdots$$
$$\times\, p_{t(\tau_n-\tau_{n-1})}(g_{n-1}^{-1}g_n) p_{t(1-\tau_n)}(g_n)\, dg_1 \cdots dg_n$$

From $\mu_t^{L_e}$, one defines a measure μ_t^L on the free loops by taking the mean over G as

$$\int f(\gamma)\mu_t^L(d\gamma) = \int_G dg \int f(g\gamma)\mu_t^{L_e}(d\gamma)$$

The quasi-invariance property for the pinned Wiener measure was proved by Malliavin and Malliavin (1992).

When the measures $(\mu_t^L)_{t\geq 0}$ are obtained by conditioning and quasisure analysis, we have heat kernel measures. The case of heat kernel measures defined on the loop group has been studied by Airault and Malliavin by disintegrating the measures on the path space and using the quasisure analysis. The Laplacian on the loop group is defined as it has been for the Laplacian on the path space,

$$(\Delta_L f)(g) = \lim_{\epsilon\to 0} \frac{1}{\epsilon}\left[\int f(gg_1)\mu_\epsilon^L(dg_1) - f(g)\right]$$

but now the heat equation has a Kac's potential Φ_t defined on the loops. On the loop group, the heat equation is

$$\frac{\partial}{\partial t}\int f(l)\mu_t^L(dl) = \int [(\Delta_L f)(l) + \Phi_t(l) f(l)]\mu_t^L(dl) \qquad [3]$$

where

$$\Phi_t(l) = \frac{1}{t^2}\left\|\int_0^1 dl(s) l(s)^{-1}\right\|_{\mathcal{G}}^2 - 2\frac{d}{dt}\log p_t(e)$$
$$-\frac{1}{t}\dim \mathcal{G}$$

The case of the circle, $G = R/2\pi Z$, is interesting. The law of the functional

$$\int_0^1 dl(s) l(s)^{-1}$$

is given in Airault and Malliavin (1996, and references therein). Moreover, the study of the heat

measures over the loop group of $R/2\pi Z$ brings new identities on the classical Jacobi theta function

$$p_t(\theta) = 1 + 2\sum_{n\geq 1} \cos(n\theta)\, e^{-n^2 t/2} \quad \text{at } \theta = 0$$

Let

$$c_t = -2\frac{d}{dt}\log p_t(0) - \frac{1}{t}$$

The following system of differential equations is given by Airault–Malliavin (1996, and references therein):

$$c_t = -\frac{1}{t^2}\sum_{n\in Z} a_n(t)$$

$$\frac{d}{dt}a_n(t) = \frac{1}{2}c_t a_n(t) + \frac{2\pi^2 n^2}{t^2}a_n(t)$$

To pass from path space to loop space, it is convenient to use the "tubular chart" introduced by Gross and the quasisure analysis developed by Airault–Malliavin. Let $\Phi : \gamma \to \gamma(1)\gamma(0)^{-1}$ from the path space to the group G; then the free loop space over G is $\Phi^{-1}(e)$. There exists a neighborhood V of the neutral of G such that $\Phi^{-1}(V)$ is diffeomorphic to $V \times L(G)$, the product of V with the loop space over G. With this diffeomorphism, one can disintegrate the measures on the path space and obtain the measures on the loop space. The Cameron–Martin formula on the path space of the group G is obtained from the Cameron–Martin formula for the Wiener space and the Ito's map. Let γ be a differentiable path with finite energy on G, that is,

$$\int_0^1 \left\| \gamma(s)^{-1}\frac{d}{ds}\gamma(s)\right\|_{\mathcal{G}}^2 < +\infty$$

it holds

$$\int f(\gamma g)\mu_t(dg) = \int f(g)k_\gamma(g)\mu_t(dg)$$

Let us denote by $(|)_{\mathcal{G}}$ the Euclidean scalar product on the Lie algebra $\mathcal{G}$; then the density is given by

$$k_\gamma(g) = \exp\left[\frac{1}{t}\int_0^1 \left(\gamma(s)^{-1}\frac{d}{ds}\gamma(s)|dg(s)g(s)^{-1}\right)_{\mathcal{G}} - \frac{1}{2t}\int_0^1 \left\|\gamma(s)^{-1}\frac{d}{ds}\gamma(s)\right\|_{\mathcal{G}}^2 ds\right]$$

The previous approach relies on the heat equation on the loop space. Thus, the metric on the Cameron–Martin loop or path space is important. The problem of quasi-invariance for metrics α with $1/2 < \alpha < 1$ relates to the random series

$$u_\alpha(\tau) = \sum_{k\geq 1}\frac{\sin k\pi\tau}{k^\alpha}G_k \qquad [4]$$

where the G_k are independent normal variables. Driver (2003) solved the problem for $1/2 < \alpha < 1$ by Riemannian geometry in infinite dimension. The Ricci curvature appears in the integration-by-parts formulas on the loop space. The case of the metric $1/2$ is out of reach. Fang (1999) calculated the Ricci curvature of the loop manifold for metrics $\alpha > 1/2$ and showed that when $\alpha \to 1/2$, these Ricci curvatures tend to a limit. Another presentation of the problem is that of Pickrell (1987), where he obtains a family of quasi-invariant measures on Grassmannians.

Given a family of measures $(\mu_t)_{t\geq 0}$ on the path space of a Riemannian manifold, one defines a heat operator as a family $(\mathcal{L}_t)_{t\geq 0}$ of operators depending upon $t \in [0, +\infty[$ such that

$$\int \mathcal{L}_t F\, d\mu_t = \frac{d}{dt}\int F\, d\mu_t \qquad [5]$$

where F is a function defined on the path space. The heat equation with a potential as [3] gives an example of a heat operator. Heat operators have been constructed for the path space over R^n by Airault–Malliavin, obtaining, after an integration by parts on the path space, a heat operator of first order. This introduces the notion of dilatation vector fields on the path space. In the case of the flat Wiener space, to each point x in the path space is associated the dilatation vector field Y such that $(Yf)(x) = (x|(\text{grad}\, f)(x))$. This gives a rescaling of the Wiener measure under dilatations. This idea has been exploited by Mancino (1999), who extended the method to free loop groups.

Integration-by-Parts Formulas

The Cameron–Martin space plays the role of the tangent space to the Wiener space. The integration-by-parts formulas are an infinitesimal version of the Cameron–Martin quasi-invariance property. Let G be a compact Lie group or any product of R^n by a compact Lie group. For a vector field z, the differentiation on the right $\partial_z^{\text{right}}$ and differentiation on the left ∂_z^{left} are given by

$$\partial_z^{\text{left}} F(p) = \lim_{\epsilon\to 0}\frac{F(\exp(\epsilon z)p) - F(p)}{\epsilon}$$

and

$$\partial_z^{\text{right}} F(p) = \lim_{\epsilon \to 0} \frac{F(p \exp(\epsilon z)) - F(p)}{\epsilon}$$

The operator $\partial_z^{\text{right}}$ commutes with the translation on the left, for a translation τ_h^{left}, then $\partial_z^{\text{right}}(F o \tau_h^{\text{left}}) = (\partial_z^{\text{right}} F) o \tau_h^{\text{left}}$ and vice versa for ∂_z^{left}.

For the measures on the path space or loop space, the problem is to prove the integration-by-parts formulas. On the path spaces on G, let μ_{P_e} be the Wiener measure on the set of paths starting from e, there exists a density k_z such that $E[\exp(ck_z)]$ is finite and

$$\int_{P_e(G)} \partial_z^{\text{left}} F(g)\, d\mu_{P_e}(g) = \int_{P_e(G)} F(g) k_z\, d\mu_{P_e}(g)$$

The density k_z is defined on the path space by

$$k_z(g) = \int_0^1 < g(t) z'(t) g(t)^{-1}, d\omega(t) >$$

This was proved by a number of authors (see, e.g., Pickrell (1987) and, in a geometrical context, Cruzeiro and Malliavin (1996)).

The existence of a density for the differentiation on the left is valid for any Lie group. This is not true for the differentiation on the right. If G is noncompact or is not the product of R^n by a compact Lie group, the existence of k_z is not proved on the right. This comes from the fact that the map Ad defined on the path group as a parallel transport does not preserve the Cameron–Martin subspace. In the case where G is not a product of a flat space by a compact Lie group, the Cameron space, which is a kind of "tangent space" to the infinite-dimensional loop manifold, is not closed under the Lie bracket of vector fields.

The integration-by-parts formulas are obtained with the stochastic calculus of variation. On a group G, consider $Y_1, Y_2, \ldots, Y_p$, p independent left-invariant vector fields. Let $\mathcal{G}$ be the Lie algebra of G. The second-order differential operator $\triangle = \sum_{j=1}^p Y_j^2$ defines a left-invariant diffusion $g_\omega(t)$ on the group G with the stochastic equation $dg_\omega(t) = g_\omega(t)\left[\sum_k (Y_k)_e o\, d\omega^k\right]$ where (ω_k) are independent Brownian motions on the Euclidean space $\mathcal{G}$. In the work by Malliavin and Malliavin (1992, and references therein), the stochastic calculus of variation is done with the right-invariant connection on the Lie group by setting

$$\phi^{\text{right}} = \frac{d}{d\epsilon}_{|\,\epsilon=0} (g_{\omega+\epsilon h}) o g_\omega^{-1}$$

where h is a differentiable function of t with values in the Lie algebra $\mathcal{G}$, with finite energy $\int_0^1 |h'(s)|^2\, ds < +\infty$. By taking the derivative with respect to ϵ in the Stratonovitch equation

$$g^\epsilon(t)^{-1} o\, dg^\epsilon(t) = d\omega(t) + \epsilon h'(t)\, dt$$

and letting $\epsilon = 0$, it turns out that ϕ^{right} is a differentiable function of t and its derivative is given by

$$\frac{d}{dt} \phi^{\text{right}}(t) = g_\omega(t) h'(t) g_\omega(t)^{-1}$$

The situation is not the same for

$$\phi^{\text{left}} = \frac{d}{d\epsilon}_{|\epsilon=0} g_\omega^{-1} o (g_{\omega+\epsilon h})$$

where $d\phi^{\text{left}}(t)$ is a stochastic differential. This generalizes to an arbitrary Riemannian manifold using a coupling of connections (see Airault and Malliavin (1996), and references therein). The construction of the appropriate Cameron subspace, that is, the choice of the infinitesimal action of vector fields on the measure, is of importance. In the commutative case of the path space over R^n, the classical Cameron–Martin subspace of paths h such that $\int_0^1 |h'(s)|^2\, ds < +\infty$ is time invariant. To define the vector fields acting on the path (or loop) space over M, it is necessary to consider the geometry of the manifold M. The infinitesimal transformations which preserve the Riemannian metric are called Riemannian connections. In the case where M is a group, the natural connections are those defined by the parallelism on the group. For a Riemannian manifold, Driver proved the existence of integration-by-parts formulas for the measures on the path space of M when M is endowed with a torsion skew-symmetric connection. The Levi-Civita connection, since it is torsionless, is of course a Driver (2003) connection. If the connection is not skew-symmetric, then two coupled connections permit study of the ϵ-variation or "reduced variation" of a path, and one obtains a Cameron–Martin formula on the path and on the loop space of the Riemannian manifold M (Fang 1999). The method of reduced variation can be used to obtain the integration-by-parts formulas over path and loop spaces. Another approach to the quasi-invariance problem, using two-parameter processes, has been provided by Norris (1995).

The Support of the Measures and Absolute Continuity with Respect to Each Other

Given a Riemannian manifold M, let $(\mu_t)_t$ be the heat kernel measures on the path space of M and let $(\rho_t)_t$ be heat kernel measures on the loop space of M; the question arises whether ρ_t is absolutely continuous

with respect to μ_t. For a connected compact Lie group G, consider the path and loop groups on G. The pinned Wiener measure on the loop group is defined as the law of a G-valued Brownian motion starting at e and conditioned to end at e, and the heat kernel measure is the endpoint distribution of Brownian motion on the loop group.

It has been shown (Driver and Srimurthy 2001) that the heat kernel measure is absolutely continuous with respect to the pinned Wiener measure, and that the Radon–Nikodym derivative is bounded. This proof relies on the heat formula with a potential [3], which is satisfied by the heat kernel measure. They give a new proof of this heat formula. When the group G is simply connected, Aida and Driver (2000) prove that the heat kernel measure over a based loop group, constructed by using the Brownian motion is equivalent to the Brownian bridge measure over a based loop group. When G is the circle, the Radon–Nikodym derivative of the heat kernel measure with respect to the pinned Wiener measure can be calculated in terms of the Jacobi theta function (Driver and Srimurthy 2001). On the loop space of R^n, at time t, the two measures, "heat kernel" and "pinned Wiener" are the same.

See also: Abelian and Nonabelian Gauge Theories Using Differential Forms; Lie Groups: General Theory; Malliavin Calculus; Path Integrals in Noncommutative Geometry.

Further Reading

Aida S and Driver BK (2000) Equivalence of heat kernel measure and pinned Wiener measure on loop groups. *Comptes Rendus de l'Academie des Sciences, Paris, I* 331: 709–712.

Aida, Shigeki, Elworthy, and David (1995) Differential calculus on path and loop spaces. Logarithmic Sobolev inequalities on path spaces. *Comptes Rendus de l'Academie des Sciences Paris Serie I Mathematics* 321(1): 97–102.

Airault H and Malliavin P (1996) Integration by parts formulas and dilatation vector fields on elliptic probability spaces. *Probability Theory and Related Fields* 106: 447–494.

Bowick MJ and Rajeev SG (1987) The holomorphic geometry of closed bosonic string theory and DiffS^1/S^1. *Nuclear Physics B* 293: 348–384.

Cruzeiro AB and Malliavin P (1996) Renormalized differential geometry on path space: structural equation, curvature. *Journal of Functional Analysis* 139: 119–181.

Driver BK (2003) *Heat Kernels Measures and Infinite Dimensional Analysis. Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces*, Paris, Contemp. Math. vol. 338, pp. 101–141. Providence, RI: American Mathematical Society.

Driver BK and Srimurthy VK (2001) Absolute continuity of heat kernel measure with pinned Wiener measure on loop groups. *Ann. Probab.* 29(2): 691–723.

Fang S (1999) Integration by parts for heat measures over loop groups. *J. Math. Pures Appl.* 78(9): 877–894.

Freed DS (1988) The geometry of loop groups. *Journal of Differential Geometry* 28: 223–276.

Gaveau B and Mazet E (1979) Diffusions sur des variétés de chemins. *Comptes Rendus de l'Academie des Sciences, Paris* 289: 643–645.

Gross L (1998) Harmonic functions on loop groups, Séminaire Bourbaki, vol. 1997–1998, Astérisque No 252, Exp. No. 846, 5, 271–286.

Hsu EP (1997) Analysis on path and loop spaces. In: Hsu EP and Varadhan SRS (eds.) *IAS/Park City Mathematics Series 5*. Princeton: American Mathematical Society and Institute for Advanced Study.

Jones V (1995) Fusion en algèbres de von Neumann et groupes de lacets (A. Wassermann). Séminaire Bourbaki, vol. 1994/95. Astérisque No. 237 (1996), Exp. No. 800, 5, 251–273.

Levy T (2003) Yang–Mills measures on compact surfaces. *Mem. Amer Math. Soc.* 166: 790.

Malliavin P (1989) Hypoellipticity in infinite dimensions. In: Pinsky M (ed.) *Diffusion Process and Related Problems in Analysis*. Chicago: Birkhauser.

Malliavin MP and Malliavin P (1992) Integration on loop groups III. Asymptotic Peter–Weyl orthogonality. *Journal of Functional Analysis* 108: 13–46.

Mancino ME (1999) Dilatation vector fields on the loop group. *Journal of Functional Analysis* 166(1): 130–147.

Neretin Y (1994) Some remarks of quasi-invariant actions of loop groups and the group of the diffeomorphisms of the circle. *Communication in Mathematical Physics* 164: 599–626.

Norris JR (1995) Twisted sheets. *Journal of Functional Analysis* 132: 273–334.

Pickrell D (1987) Measures on infinite-dimensional Grassmann manifolds. *Journal of Functional Analysis* 70(2): 323–356.

Noncommutative Geometry and the Standard Model

T Schücker, Université de Marseille, Marseille, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The aim of this contribution is to explain how Connes derives the standard model of electromagnetic, weak, and strong forces from noncommutative geometry. The reader is supposed to be aware of two other derivations in fundamental physics: the derivation of the Balmer–Rydberg formula for the spectrum of the hydrogen atom from quantum mechanics and Einstein's derivation of gravity from Riemannian geometry.

At the end of the nineteenth century, new physics was discovered in atoms, namely their discrete spectra. Balmer and Rydberg succeeded to put order into the fast-growing set of experimental results with the help of a phenomenological ansatz for the frequencies ν of the spectral rays of, for example, the hydrogen atom,

$$\nu = g(n_2^q - n_1^q), \quad n_j \in \mathbb{N}, \ q \in \mathbb{Z}, \ g \in \mathbb{R} \qquad [1]$$

The integer variables n_1 and n_2 reflect the discreteness of the spectrum. On the other hand, the discrete parameter q and the continuous parameter g were fitted by experiment: $q = -2$ and $g = 3.289 \times 10^{15}$ Hz, the famous Rydberg constant. Later quantum mechanics was discovered and allowed to derive the Balmer–Rydberg ansatz and to constrain its parameters:

$$q = 2 \quad \text{and} \quad g = \frac{m_e}{4\pi\hbar^3}\frac{e^4}{(4\pi\epsilon_0)^2} \qquad [2]$$

in beautiful agreement with the anterior experimental fit.

The Standard Model

We propose to introduce the standard model (*see* Standard Model of Particle Physics) in analogy with the Balmer–Rydberg formula (**Table 1**).

Table 1 An analogy between atomic and particle physics elements

	Atomic physics	*Particle physics*
New physics	Discrete spectra	Forces mediated by gauge bosons
Ansatz	$\nu = g(n_2^q - n_1^q)$	Yang–Mills–Higgs models
Experimental fit	$q = -2, g = 3.289 \times 10^{15}$ Hz	Standard model
Underlying theory	Quantum mechanics	Noncommutative geometry

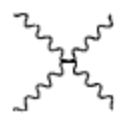
Notes:

The Yang–Mills–Higgs Ansatz

The variables of this Lagrangian ansatz are spin-1 particles A, spin-(1/2) particles decomposed into left- and right-handed components $\psi = (\psi_L, \psi_R)$ and spin-0 particles φ. There are four discrete parameters, a compact real Lie group G, the "gauge group," and three unitary representations on complex Hilbert spaces $\mathcal{H}_L, \mathcal{H}_R$, and $\mathcal{H}_S$. The spin-1 particles come in a multiplet living in the complexified of the Lie algebra of G, $A \in \mathrm{Lie}(G)^C$. The left- and right-handed spinors come in multiplets living in the Hilbert spaces, $\psi_L \in \mathcal{H}_L, \psi_R \in \mathcal{H}_R$, respectively. The (Higgs) scalar is another multiplet, $\varphi \in \mathcal{H}_S$. The Yang–Mills–Higgs Lagrangian, together with its Feynman diagrams, is spelled out in **Table 2**.

There are several continuous parameters: the gauge coupling $g \in \mathbb{R}_+$, the Higgs self-couplings $\lambda, \mu \in \mathbb{R}_+$, and several Yukawa couplings $g_Y \in \mathbb{C}$.

Table 2 The Yang–Mills–Higgs Lagrangian and its Feynman diagrams

$\mathcal{L}[A, \psi, \varphi] = \frac{1}{2}\mathrm{tr}(\partial_\mu A_\nu \partial^\mu A^\nu - \partial_\mu A_\nu \partial^\nu A^\mu)$	
$+g\ \mathrm{tr}(\partial_\mu A_\nu [A^\mu, A^\nu])$	
$+g^2\ \mathrm{tr}([A_\mu, A_\nu][A^\mu, A^\nu])$	
$+\bar{\psi}\partial\!\!\!/\psi$	
$+ig\bar{\psi}(\tilde{\rho}_L \oplus \tilde{\rho}_R)(A_\mu)\gamma^\mu\psi$	
$+\frac{1}{2}\partial_\mu\varphi^*\partial^\mu\varphi$	
$+\frac{1}{2}g\{(\tilde{\rho}_S(A_\mu)\varphi)^*\partial^\mu\varphi + \partial_\mu\varphi^*\tilde{\rho}_S(A_\mu)\varphi\}$	
$+\frac{1}{2}g^2(\tilde{\rho}_S(A_\mu)\varphi)^*\tilde{\rho}_S(A^\mu)\varphi$	
$+\lambda\varphi^*\varphi\varphi^*\varphi$	
$-\frac{1}{2}\mu^2\varphi^*\varphi$	
$+g_Y\ \bar{\psi}\varphi\psi + \bar{g}_Y\ \bar{\psi}\varphi^*\psi$	

Let us choose $G = \mathrm{U}(1) \ni e^{i\theta}$. Its irreducible unitary representations are all one-dimensional, $\mathcal{H} = \mathbb{C} \ni \psi$ characterized by the charge $q \in \mathbb{Z}$: $\rho(e^{i\theta})\psi = e^{iq\theta}\psi$. Then with $q_L = q_R$ and $\mathcal{H}_S = \{0\}$, we get Maxwell's theory with the photon (or gauge boson or 4-potential) A coupled to the Dirac theory of a massless spinor of electric charge q_L whose (relativistic) wave function is ψ. The gauge coupling is given by $g = e/\sqrt{\epsilon_0}$. Gauge invariance of the Yang–Mills–Higgs Lagrangian implies, via Noether's theorem, electric charge conservation in this case (*see* Symmetries and Conservation Laws).

Yang–Mills models are therefore simply nonabelian generalizations of electromagnetism where the abelian gauge group U(1) is replaced by any compact real Lie group. We insist on a compact group because all irreducible unitary representations of compact groups are finite dimensional. Finally, the Higgs scalar is added to give masses to spinors and gauge bosons via spontaneous symmetry breaking (*see* Symmetry Breaking in Field Theory).

We use compact groups and unitary representations as (discrete) parameters. One motivation is Noether's theorem and conserved quantities. The other comes from Wigner's theorem: the irreducible unitary representations of the Poincaré group are classified by mass and spin. Its orthonormal basis vectors are classified by energy–momentum and by the z-component of angular momentum. This theorem leads to the widely accepted definition of a particle as an orthonormal basis vector in a Hilbert space $\mathcal{H}$ carrying a unitary representation ρ of a group G.

A precious property of the Yang–Mills–Higgs ansatz is its perturbative renormalizability necessary for fine-structure calculations like the anomalous magnetic moment of the muon.

The Experimental Fit

Physicists have spent some 30 years and some 10^9 Swiss Francs to distill the fit (Particle Data Group 2004):

$$G = \mathrm{SU}(2) \times \mathrm{U}(1) \times \mathrm{SU}(3)/(\mathbb{Z}_2 \times \mathbb{Z}_3) \qquad [3]$$

$$\mathcal{H}_L = \bigoplus_1 [(2, \tfrac{1}{6}, 3) \oplus (2, -\tfrac{1}{2}, 1)] \qquad [4]$$

$$\mathcal{H}_R = \bigoplus_1^3 [(1, \tfrac{2}{3}, 3) \oplus (1, -\tfrac{1}{3}, 3) \oplus (1, -1, 1)] \qquad [5]$$

$$\mathcal{H}_S = (2, -\tfrac{1}{2}, 1) \qquad [6]$$

Here (n_2, y, n_3) denotes the tensor product of an n_2-dimensional representation of SU(2), "(weak) isospin," an n_3-dimensional representation of SU(3), "color," and the one-dimensional representation of

U(1) with "hyper charge" y. For historical reasons, the hypercharge is an integer multiple of 1/6. This is irrelevant: in the abelian case, only the product of the hypercharge with its gauge coupling is measurable, and we do not need multivalued representations, which are characterized by noninteger, rational hypercharges. In the direct sum, we recognize the three generations of fermions, the quarks, "up, down, charm, strange, top, bottom," are SU(3) triplets, the leptons, "electron, μ, τ" and their neutrinos, are color singlets. The basis of the fermion representation space is

$$\begin{pmatrix} u \\ d \end{pmatrix}_{\mathrm{L}}, \quad \begin{pmatrix} c \\ s \end{pmatrix}_{\mathrm{L}}, \quad \begin{pmatrix} t \\ b \end{pmatrix}_{\mathrm{L}}$$
$$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_{\mathrm{L}}, \quad \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_{\mathrm{L}}, \quad \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_{\mathrm{L}}$$
$$u_{\mathrm{R}}, c_{\mathrm{R}}, t_{\mathrm{R}}, e_{\mathrm{R}}, \mu_{\mathrm{R}}, \tau_{\mathrm{R}}$$
$$d_{\mathrm{R}}, s_{\mathrm{R}}, b_{\mathrm{R}},$$

The parentheses indicate isospin doublets.

The eight gauge bosons associated with su(3) are called gluons. Warning: the U(1) is not the one of electric charge; it is called hypercharge, the electric charge is a linear combination of hypercharge and weak isospin. This mixing is necessary to give electric charges to the W bosons. The W^+ and W^- are pure isospin states, while the Z^0 and the photon are (orthogonal) mixtures of the third isospin generator and hypercharge.

As the group G contains three simple factors, there are three gauge couplings,

$$\begin{aligned} g_2 &= 0.6518 \pm 0.0003 \\ g_1 &= 0.3574 \pm 0.0001 \\ g_3 &= 1.218 \pm 0.01 \end{aligned} \qquad [7]$$

The Higgs couplings are usually expressed in terms of the W and Higgs masses:

$$m_W = \tfrac{1}{2} g_2 v = 80.419 \pm 0.056\,\mathrm{GeV} \qquad [8]$$

$$m_\varphi = 2\sqrt{2}\sqrt{\lambda} v > 98\,\mathrm{GeV} \qquad [9]$$

with the vacuum expectation value $v := (1/2)\mu/\sqrt{\lambda}$. Because of the high degree of reducibility of the spin-(1/2) representations there are 27 complex Yukawa couplings. They constitute the fermionic mass matrix which contains the fermion masses and mixings:

$$\begin{aligned} m_e &= 0.510998902 \pm 0.000000021\,\mathrm{MeV} \\ m_u &= 3 \pm 2\,\mathrm{MeV}, \quad m_d = 6 \pm 3\,\mathrm{MeV} \\ m_\mu &= 0.105658357 \pm 0.000000005\,\mathrm{GeV} \\ m_c &= 1.25 \pm 0.1\,\mathrm{GeV}, \quad m_s = 0.125 \pm 0.05\,\mathrm{GeV} \\ m_\tau &= 1.77703 \pm 0.00003\,\mathrm{GeV} \\ m_t &= 174.3 \pm 5.1\,\mathrm{GeV}, \quad m_b = 4.2 \pm 0.2\,\mathrm{GeV} \end{aligned}$$

For simplicity, we have taken massless neutrinos. Then mixing only occurs for quarks and is given by a unitary matrix, the Cabibbo–Kobayashi–Maskawa matrix

$$C_{\mathrm{KM}} := \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \qquad [10]$$

whose matrix elements in terms of absolute values are:

$$\begin{pmatrix} 0.9750 \pm 0.0008 & 0.223 \pm 0.004 & 0.004 \pm 0.002 \\ 0.222 \pm 0.003 & 0.9742 \pm 0.0008 & 0.040 \pm 0.003 \\ 0.009 \pm 0.005 & 0.039 \pm 0.004 & 0.9992 \pm 0.0003 \end{pmatrix} \qquad [11]$$

Mathematically, the Cabibbo–Kobayashi–Maskawa matrix comes from a polar decomposition of the mass matrix. The physical meaning of the quark mixings is the following: when a sufficiently energetic W^+ decays into a u quark, this u quark is produced together with a $\bar{d}$ quark with probability $|V_{ud}|^2$, an $\bar{s}$ quark with probability $|V_{us}|^2$, and a $\bar{b}$ quark with probability $|V_{ub}|^2$.

The phenomenological success of the standard model is phenomenal: with only a handful of parameters, it reproduces correctly some millions of experimental numbers: cross sections, lifetimes, branching ratios.

Noncommutative Geometry

Noncommutative geometry is an analytic geometry generalizing three other geometries that also had important impact on our understanding of forces and time. Let us start by briefly recalling the three forerunners (**Table 3**). Euclidean geometry underlies Newton's mechanics as a geometry in the space of positions. Forces are described by vectors living in the same space and the Euclidean scalar product is needed to define work and potential energy. Time is not part of geometry – it is absolute. This point of view is abandoned in special relativity unifying space and time into Minkowskian geometry. This new point of view allows to derive the magnetic

Table 3 Four nested analytic geometries

Geometry	*Force*	*Time*
Euclidean	$E = \int \vec{F} \cdot \mathrm{d}\vec{x}$	Absolute
Minkowskian	$\vec{E}, \epsilon_0 \Rightarrow \vec{B}, \mu_0 = \epsilon_0^{-1} c^{-2}$	Universal
Riemannian	Coriolis $\leftrightarrow$ gravity	Proper, τ
Noncommutative	Gravity $\Rightarrow$ YMH, $\lambda = \frac{1}{3} g_2^2$	$\Delta\tau \sim 10^{-40}$ s

field from the electric field as a pseudoforce associated with a Lorentz boost. Although time becomes relative, one can still imagine a grid of synchronized clocks, that is, a universal time. The next generalization is "Riemannian geometry = curved spacetime." Here gravity can be viewed as the pseudoforce associated with a uniformly accelerated coordinate transformation. At the same time, universal time loses all meaning and we must content ourselves with proper time. With today's precision in time measurement, this complication of life becomes a bare necessity, for example, the global positioning system (GPS).

Our last generalization is "noncommutative geometry = curved space(time) with an uncertainty principle." As in quantum mechanics, this uncertainty principle is introduced via noncommutativity.

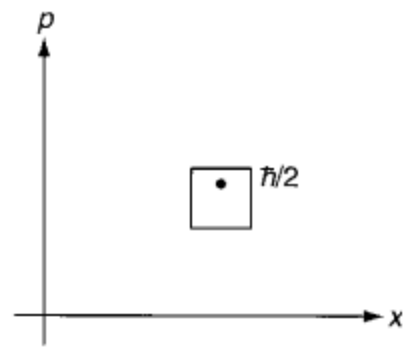


Figure 1 The first example of noncommutative geometry.

Quantum Mechanics

Consider the classical harmonic oscillator. Its phase space is $\mathbb{R}^2$ with points labeled by position x and momentum p. A classical observable is a differentiable function on phase space such as the total energy $p^2/(2m) + kx^2$. Observables can be added and multiplied, and they form the algebra $\mathcal{C}^\infty(\mathbb{R}^2)$, which is associative and commutative. To pass to quantum mechanics, this algebra is rendered noncommutative by means of a noncommutation relation for the generators x and p: $[x, p] = i\hbar 1$. Let us call $\mathcal{A}$ the resulting algebra "of quantum observables." It is still associative, and has an involution $\cdot^*$ (the adjoint or Hermitian conjugation) and a unit 1.

Of course, there is no space anymore of which $\mathcal{A}$ is the algebra of functions. Nevertheless, we talk about such a "quantum phase space" as a space that has no points or a space with an uncertainty relation. Indeed, the noncommutation relation implies Heisenberg's uncertainty relation $\Delta x \Delta p \geq \hbar/2$ and tells us that points in phase space lose all meaning; we can only resolve cells in phase space of volume $\hbar/2$, see **Figure 1**. To define the uncertainty Δa for an observable $a \in \mathcal{A}$, we need a faithful representation of the algebra on a Hilbert space, that is, an injective homomorphism ρ from $\mathcal{A}$ into the algebra of operators on $\mathcal{H}$. For the harmonic oscillator, this Hilbert space is $\mathcal{H} = \mathcal{L}^2(\mathbb{R})$. Its elements are the wave functions $\psi(x)$, square-integrable functions on configuration space. Finally, the dynamics is defined by the Hamiltonian, a self-adjoint observable $H = H^* \in \mathcal{A}$ via Schrödinger's equation $(i\hbar\partial/\partial t - \rho(H))\psi(t, x) = 0$. Here time is an external parameter; in particular, time is not an observable. This is different in the special-relativistic setting, where Schrödinger's equation is replaced by Dirac's equation $\partial\!\!\!/\psi = 0$. Now the wave function ψ is the four-component spinor consisting of left- and right-handed, particle and antiparticle wave functions. Unlike the Hamiltonian, the Dirac operator does not lie in $\mathcal{A}$, but it is still an operator on $\mathcal{H}$. In Euclidean spacetime, the Dirac operator is also self-adjoint, $\partial\!\!\!/^* = \partial\!\!\!/$.

Spectral Triples

Noncommutative geometry (Connes 1994, 1995) does to a compact Riemannian spin manifold M what quantum mechanics does to phase space. A noncommutative geometry is defined by the three purely algebraic items $(\mathcal{A}, \mathcal{H}, \partial\!\!\!/)$, called a spectral triple. $\mathcal{A}$ is a real, associative, and possibly noncommutative involution algebra with unit, faithfully represented on a complex Hilbert space $\mathcal{H}$, and $\partial\!\!\!/$ is a self-adjoint operator on $\mathcal{H}$. As the spectral triple, also the axioms linking its three items are motivated by relativistic quantum mechanics.

When $\mathcal{A} = \mathcal{C}^\infty(M)$, the functions on a Riemannian spin manifold M, represented on spinors ψ, and $\partial\!\!\!/$ is the gravitational Dirac operator, one has a spectral triple. The converse is also true when $\mathcal{A}$ is a suitable commutative algebra (Connes 1996), but the axioms make sense even when $\mathcal{A}$ is not commutative. As for quantum phase space, Connes defines a noncommutative geometry by a spectral triple whose algebra is allowed to be noncommutative and he shows how important properties like dimensions, distances, differentiation, integration, general coordinate transformations, and direct products generalize to the noncommutative setting. As a bonus, the algebraic axioms of a spectral triple, commutative or not, include discrete, that is, zero-dimensional spaces that now are naturally equipped with a differential calculus. These spaces have finite-dimensional algebras and Hilbert spaces, meaning that their algebras are just matrix algebras.

An "almost commutative geometry" is defined as a direct product of a four-dimensional commutative geometry, "ordinary spacetime," by a zero-dimensional noncommutative geometry, the "internal space." If the

latter is also commutative, for example, the ordinary two-point space, then the direct product describes a two-sheeted universe or a Kaluza–Klein space whose fifth dimension is discrete, (Madone 1995). In general, the axioms of spectral triples imply that the Dirac operator of the internal space is precisely the fermionic mass matrix.

As a generic example, here is the internal spectral triple underlying the standard model with one generation of quarks and leptons. The algebra $\mathcal{A} = \mathbb{H} \oplus \mathbb{C} \oplus M_3(\mathbb{C}) \ni (a, b, c)$ contains quaternions, that is, 2×2 matrices of the form

$$a = \begin{pmatrix} x & -\bar{y} \\ y & \bar{x} \end{pmatrix}, \quad x, y \in \mathbb{C}$$

complex numbers b and complex 3×3 matrices c. The Hilbert space is 30-dimensional, where we count particles and antiparticles $(\cdot^c)$ separately: $\mathcal{H} = \mathcal{H}_L \oplus \mathcal{H}_R \oplus \mathcal{H}_L^c \oplus \mathcal{H}_R^c = \mathbb{C}^8 \oplus \mathbb{C}^7 \oplus \mathbb{C}^8 \oplus \mathbb{C}^7$. The representation is block-diagonal, with the four blocks

$$\rho_L(a) := \begin{pmatrix} a \otimes 1_3 & 0 \\ 0 & a \end{pmatrix}$$

$$\rho_R(b) := \begin{pmatrix} b1_3 & 0 & 0 \\ 0 & \bar{b}1_3 & 0 \\ 0 & 0 & \bar{b} \end{pmatrix} \qquad [12]$$

$$\rho_L^c(b, c) := \begin{pmatrix} 1_2 \otimes c & 0 \\ 0 & \bar{b}1_2 \end{pmatrix}$$

$$\rho_R^c(b, c) := \begin{pmatrix} c & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & \bar{b} \end{pmatrix} \qquad [13]$$

The internal Dirac operator (= fermionic mass matrix) contains two quark masses m_u, m_d and one lepton mass m_e, and no mixing:

$$\mathcal{D} = \begin{pmatrix} 0 & \mathcal{M} & 0 & 0 \\ \mathcal{M}^* & 0 & 0 & 0 \\ 0 & 0 & 0 & \bar{\mathcal{M}} \\ 0 & 0 & \bar{\mathcal{M}}^* & 0 \end{pmatrix} \qquad [14]$$

$$\mathcal{M} = \begin{pmatrix} \begin{pmatrix} m_u & 0 \\ 0 & m_d \end{pmatrix} \otimes 1_3 & 0 \\ 0 & \begin{pmatrix} 0 \\ m_e \end{pmatrix} \end{pmatrix}$$

These matrices look rather *ad hoc*; they are not. They define an irreducible spectral triple and, for a given algebra, there is only a finite number of such triples.

The Spectral Action

Chamseddine and Connes (1997) generalize general relativity to noncommutative spacetimes in two strokes, kinematics and dynamics. They explicitly compute this generalization for almost commutative geometries.

Kinematics In noncommutative geometry, general coordinate transformations are algebra automorphisms lifted to the Hilbert space of spinors. For almost commutative geometries, these transformations are precisely general coordinate transformations of ordinary spacetime and gauge transformations. Now remember how Einstein uses the equivalence principle to produce "gravity = curvature" starting from the flat metric, which in Connes' language is the ordinary flat Dirac operator. When applied to an almost commutative geometry (Connes 1996), the equivalence principle produces again a curved metric via the ordinary coordinate transformations on M, while the gauge transformations applied to the fermionic mass matrix produce a new field, the Higgs scalar φ. For the example above, this field is precisely the isospin doublet, color singlet with hypercharge $-1/2$ of eqn [6]. Gauge transformations also apply to the ordinary Dirac operator, thereby producing the gauge fields A.

Dynamics The group of generalized coordinate transformations allowed us to construct the configuration space. In the almost commutative case it consists of Riemannian metrics, gauge fields, and Higgs scalars. We now want a dynamics on this configuration space. Of course, we want this dynamics to be invariant under the group of generalized coordinate transformations. Note that the spectrum of the Dirac operator is invariant under this group and Chamseddine and Connes (1997) define the spectral action as a regularized partition function of these eigenvalues.

On almost commutative geometries, the spectral action is equal to the Einstein–Hilbert action plus the Yang–Mills–Higgs ansatz (**Figure 2**). In other words, almost commutative geometry explains the forces mediated by gauge bosons and Higgs scalars as pseudoforces accompanying the gravitational force in the same way that Minkowskian geometry (i.e., special relativity) explains the magnetic force as a pseudoforce accompanying the electric force.

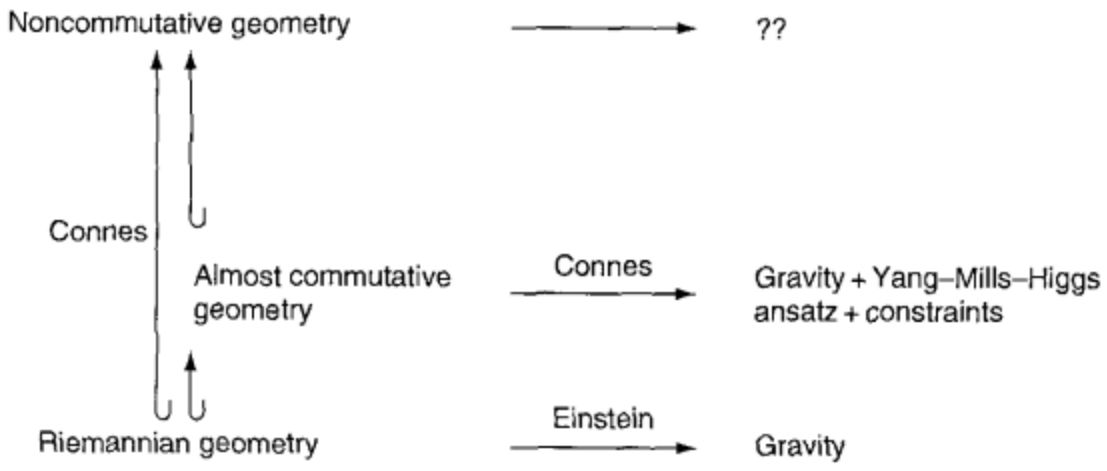


Figure 2 Deriving the Yang–Mills–Higgs ansatz from gravity.

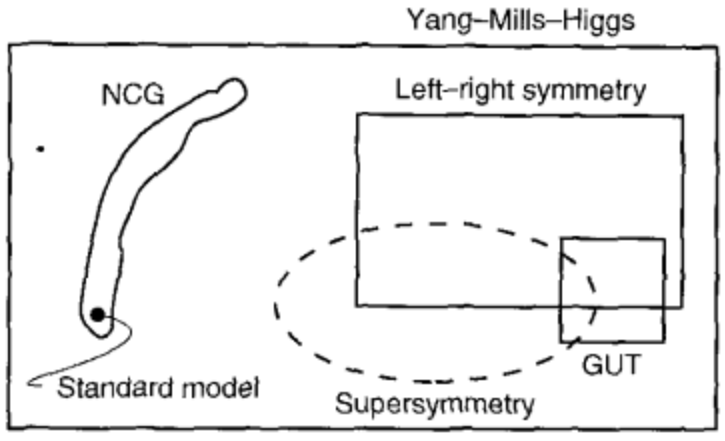


Figure 3 Constraints inside the ansatz.

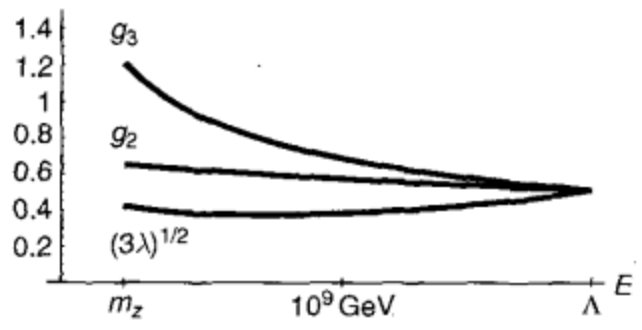


Figure 4 Running coupling constants.

There are constraints on the discrete and continuous parameters in the Yang–Mills–Higgs ansatz deriving from the spectral action **Figure 3**.

In particular, if we consider only irreducible spectral triples and among them only those which produce nondegenerate fermion masses compatible with renormalization, then we only get the standard model with one generation of quarks and leptons, with a massless neutrino and with an arbitrary number of colors, and a few submodels thereof. More than one generation and neutrino masses are possible but imply reducible triples. However, in at least one generation, the neutrino must remain purely left and massless.

For the standard model with N generations and N_c colors, we have the constraints $g_{N_c}^2 = g_2^2 = (9/N)\lambda$ on the continuous parameters. If we put $N = N_c = 3$ and if we believe in the popular "big desert" then these constraints yield a "unification scale" $\Lambda = 10^{17}$ GeV at which the uncertainty relation in spacetime should become manifest, $\Delta\tau = \hbar/\Lambda$, and a Higgs mass of $m_\varphi = 171.6 \pm 5$ GeV for $m_t = 174.3 \pm 5.1$ GeV (see **Figure 4**).

It is clear that almost commutative geometries only scratch the surface of a gold mine. May we hope that a genuinely noncommutative geometry will solve our present problems with quantum field theory and quantum gravity?

See also: Compact Groups and Their Representations; Dirac Fields in Gravitation and Nonabelian Gauge Theory; Effective Field Theories; General Relativity: Overview; Hopf Algebras and q-Deformation Quantum Groups; Positive Maps on C^*-Algebras; Quantum Hall Effect; Standard Model of Particle Physics; Symmetries and Conservation Laws; Symmetry Breaking in Field Theory; von Neumann Algebras: Introduction, Modular Theory, and Classification Theory.

Further Reading

Chamseddine A and Connes A (1997) The spectral action principle. *Communications in Mathematical Physics* 186: 731 (hep-th/9606001).

Connes A (1994) *Noncommutative Geometry*. San Diego: Academic Press.

Connes A (1995) Noncommutative geometry and reality. *Journal of Mathematical Physics* 36: 6194.

Connes A (1996) Gravity coupled with matter and the foundation of noncommutative geometry. *Communications in Mathematical Physics* 155: 109 (hep-th/9603053).

Gracia-Bondía JM, Várilly JC, and Figueroa H (2000) *Elements of Noncommutative Geometry*. Boston: Birkhäuser.

Kastler D (2000) Noncommutative geometry and fundamental physical interactions: the Lagrangian level. *Journal of Mathematical Physics* 41: 3867.

Landi G (1997) *An Introduction to Noncommutative Spaces and Their Geometry*, hep-th/9701078. Berlin: Springer.

Madore J (1995) *An Introduction to Noncommutative Differential Geometry and Its Physical Applications*. Cambridge: Cambridge University Press.

Martín CP, Gracia-Bondía JM, and Várilly JC (1998) The standard model as a noncommutative geometry: the low mass regime. *Physics Reports* 294: 363 (hep-th/9605001).

O'Raifeartaigh L (1986) *Group Structure of Gauge Theories*. Cambridge: Cambridge University Press.

Scheck F, Werner W, and Upmeier H (eds.) (2002) *Noncommutative Geometry and the Standard Model of Elementary Particle Physics*, Lecture Notes in Physics, vol. 596. Berlin: Springer.

Schücker T (2005) Forces from Connes' geometry. In: Bick E and Steffen F (eds.) *Topology and Geometry in Physics*, Lecture Notes in Physics, vol. 659, hep-th/0111236. Berlin: Springer.

The Particle Data Group, *Particle Physics Booklet* and http://pdg.lbl.gov



Nonperturbative and Topological Aspects of Gauge Theory

R W Jackiw, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Classical fields that enter a classical field theory provide a mapping from the "base" manifold on which they are defined (space or spacetime) to a "target" space over which they range. The base and target spaces, as well as the map, may possess nontrivial topological features, which affect the fixed-time description and the temporal evolution of the fields, thereby influencing the physical reality that these fields describe. Quantum fields of a quantum field theory are operator-valued distributions whose relevant topological properties are obscure. Nevertheless, topological features of the corresponding classical fields are important in the quantum theory for a variety of reasons: (1) Quantized fields can undergo local (spacetime-dependent) transformations (gauge transformations, coordinate diffeomorphisms) that involve classical functions whose topological properties determine the allowed quantum field theoretic structures. (2) One formulation of the quantum field theory uses a functional integral over classical fields, and classical topological features become relevant. (3) Semiclassical (WKB) approximations to the quantum theory rely on classical dynamics, and again classical topology plays a role in the analysis.

Topological effects of gauge fields in quantum theory were first appreciated by Dirac in his study of the quantum mechanics for (hypothetical) magnetic point monopoles. Although here one is not dealing with a field theory, the consequences of his analysis contain many features that were later encountered in field theory models.

The Lorentz equations of motion for a charged (e) massive (M) particle in a monopole magnetic field ($\boldsymbol{B}=m\boldsymbol{r}/r^3$) are unexceptional,

$$\dot{\boldsymbol{r}}=\frac{\boldsymbol{p}}{M} \qquad [1a]$$

$$\dot{\boldsymbol{p}}=\frac{e}{M}\boldsymbol{p}\times\boldsymbol{B} \quad (c=1) \qquad [1b]$$

and completely determine classical dynamics. But knowledge of the Lagrangian L and of the action I – the time integral of L: $I=\int \mathrm{d}t\, L$ – is further needed for quantum mechanics, either in its functional integral formulation or in its Hamiltonian formulation, which requires the canonical momentum $\boldsymbol{\pi}\equiv\partial L/\partial\dot{\boldsymbol{r}}$. The Lorentz-force action is expressed in terms of the vector potential $\boldsymbol{A}$, $\boldsymbol{B}=\boldsymbol{\nabla}\times\boldsymbol{A}$: $I_{\text{Lorentz}}=e\int \mathrm{d}t\,\dot{\boldsymbol{r}}\cdot\boldsymbol{A}=e\int \mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}$. The magnetic monopole vector potential is necessarily singular because $\boldsymbol{\nabla}\cdot\boldsymbol{B}=4\pi m\delta^3(\boldsymbol{r})\neq 0$. The singularity (Dirac string) can be moved, but not removed, by gauge transformations, which also are singular, and do not leave the Lorentz action invariant. Noninvariance of the action can be tolerated provided its change is an integral multiple of 2π, since the functional integrand involves $\exp(\mathrm{i}I)$ (with $\hbar=1$). The quantal requirement, which is not seen in the equations of motion, is met when

$$eg=N/2 \qquad [2]$$

The topological background to this (Dirac) quantization condition is the fact that $\Pi_1(\mathrm{U}(1))$ is the group of integers, that is, the map of the unit circle into the gauge group, here U(1), is classified by integers.

Further analysis shows that only point magnetic sources can be incorporated in particle quantum mechanics, which is governed by the particle Hamiltonian $H=\boldsymbol{p}^2/2M$ (magnetic fields do no work and are not seen in H). Quantum Lorentz equations are regained by commutation with H: $\dot{\boldsymbol{r}}=\mathrm{i}[H,\boldsymbol{r}]$, $\dot{\boldsymbol{p}}=\mathrm{i}[H,\boldsymbol{p}]$, provided

$$\mathrm{i}[r^i,r^j]=0 \qquad [3a]$$

$$\mathrm{i}[p^i,r^j]=\delta^{ij} \qquad [3b]$$

$$\mathrm{i}[p^i,p^j]=-e\varepsilon^{ijk}B^k \qquad [3c]$$

But [3c] implies that the Jacobi identity is obstructed by magnetic sources $\boldsymbol{\nabla}\cdot\boldsymbol{B}\neq 0$.

$$\tfrac{1}{2}\varepsilon^{ijk}[p^i,[p^j,p^k]]=e\,\boldsymbol{\nabla}\cdot\boldsymbol{B} \qquad [4]$$

This obstruction is better understood by examining the unitary operator $U(\boldsymbol{a})\equiv\exp(\mathrm{i}\boldsymbol{a}\cdot\boldsymbol{p})$, which according to [3b] implements finite translations of $\boldsymbol{r}$ by $\boldsymbol{a}$. The commutator algebra [3] and the failure of the Jacobi identity [4] imply that these operators do not associate. Rather one finds

$$U(\boldsymbol{a}_1)(U(\boldsymbol{a}_2)U(\boldsymbol{a}_3))=\mathrm{e}^{\mathrm{i}\Phi}(U(\boldsymbol{a}_1)U(\boldsymbol{a}_2))U(\boldsymbol{a}_3) \qquad [5]$$

where $\Phi=e\int \mathrm{d}^3x\,\boldsymbol{\nabla}\cdot\boldsymbol{B}$ is the total flux emerging from the tetrahedron formed from the three vectors $\boldsymbol{a}_i$ with vertex at $\boldsymbol{r}$ (see Figure 1). But quantum mechanics realized by linear operators acting on a Hilbert space requires that operator multiplication

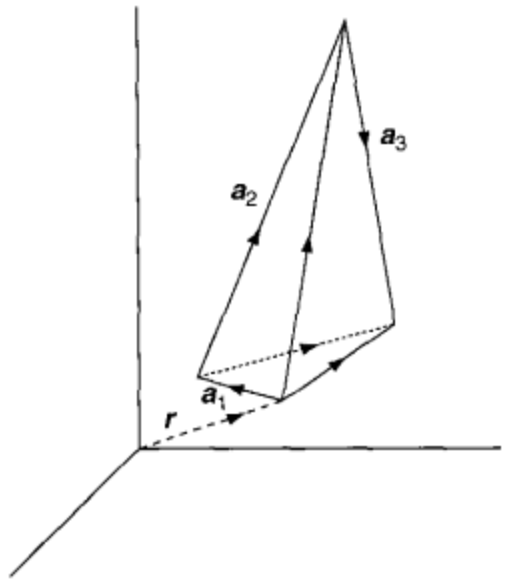


Figure 1 Tetrahedron pierced by magnetic flux that obstructs associativity.

be associative. This can be achieved, in spite of [5], provided Φ is an integral multiple of 2π, hence invisible in the exponent. This then needs that (1) $\nabla \cdot \boldsymbol{B}$ be localized at points, so that the volume integral of $\nabla \cdot \boldsymbol{B}$ retain integrality for arbitrary $\boldsymbol{a}_i$ and (2) the strengths of the localized poles obey Dirac quantization. The points at which $\nabla \cdot \boldsymbol{B}$ is localized can now be removed from the manifold and the Jacobi identity is regained. The above argument, which rederives Dirac's quantization, makes no reference to gauge variance of magnetic potentials.

In the remainder we shall discuss related phenomena for selected gauge field theories in four, three, and two dimensions that describe actual physical events occurring in nature. We shall encounter in generalized form, analogs to the above quantum mechanical system.

Some definitions and notational conventions: Nonabelian gauge potentials A_μ^a carry a spacetime index (μ) (metric tensor $g_{\mu\nu} = \text{diag}(1, -1, \dots)$) and an adjoint group index (a). When contracted with anti-Hermitian matrices T_a that represent the group's Lie algebra (structure constants $f_{ab}{}^c$)

$$[T_a, T_b] = f_{ab}{}^c T_c \qquad [6]$$

they become Lie algebra-valued.

$$A_\mu \equiv A_\mu^a T_a \qquad [7]$$

Gauge transformations transform A_μ by group elements U:

$$A_\mu \to A_\mu^U \equiv U^{-1} A_\mu U + U^{-1} \partial_\mu U \qquad [8a]$$

For infinitesimal gauge transformations, $U \approx I + \lambda$, $\lambda \equiv \lambda^a T_a$; this leads to the covariant derivative D_μ:

$$\begin{aligned} A_\mu &\to A_\mu + \partial_\mu \lambda + [A_\mu, \lambda] \equiv A_\mu + D_\mu \lambda \\ A_\mu^a &\to A_\mu^a + \partial_\mu \lambda^a + f_{bc}{}^a A_\mu^b \lambda^c \equiv A_\mu^a + (D_\mu \lambda)^a \end{aligned} \qquad [8b]$$

(In a quantum field theory, A_μ becomes an operator but the gauge transformations U, λ remain c-number functions.) The field strength $F_{\mu\nu}$ given by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu] \qquad [9a]$$

is also given by

$$[D_\mu, D_\nu] \dots = [F_{\mu\nu}, \dots] \qquad [9b]$$

(coupling strength g has been scaled to unity). The definition [9] implies the Bianchi identity

$$D_\mu F_{\nu\omega} + D_\omega F_{\mu\nu} + D_\nu F_{\omega\mu} = 0 \qquad [10]$$

Here $F_{\mu\nu}$ is gauge covariant

$$F_{\mu\nu} \to F_{\mu\nu}{}^U = U^{-1} F_{\mu\nu} U \qquad [11a]$$

or, infinitesimally,

$$F_{\mu\nu} \to F_{\mu\nu} + [F_{\mu\nu}, \lambda] \qquad [11b]$$

In the gauge invariant Yang–Mills action I_{YM}, the Yang–Mills Lagrange density $\mathcal{L}_{\text{YM}}$ is integrated over the base space,

$$\begin{aligned} \mathcal{L}_{\text{YM}} &= \tfrac{1}{2} \text{tr}\, F^{\mu\nu} F_{\mu\nu} \\ I_{\text{YM}} &= \int \mathcal{L}_{\text{YM}} = \frac{1}{2} \int \text{tr}\, F^{\mu\nu} F_{\mu\nu} \end{aligned} \qquad [12]$$

The trace is evaluated with the convention

$$\text{tr}\, T_a T_b = -\tfrac{1}{2} \delta_{ab} \qquad [13]$$

and henceforth there is no distinction between upper and lower group indices. The Euler–Lagrange condition for stationarizing I_{YM} gives the Yang–Mills equation

$$D_\mu F^{\mu\nu} = 0 \qquad [14a]$$

Should sources J^μ be present, [14a] becomes

$$D_\mu F^{\mu\nu} = J^\nu \qquad [14b]$$

and J^μ must be covariantly conserved:

$$\begin{aligned} D_\nu J^\nu &= D_\nu D_\mu F^{\mu\nu} = -\tfrac{1}{2} [D_\mu, D_\nu] F^{\mu\nu} \\ &= -\tfrac{1}{2} [F_{\mu\nu}, F^{\mu\nu}] = 0 \end{aligned} \qquad [15]$$

All this is a nonabelian generalization of familiar Maxwell electrodynamics.

Gauge Theories in Four Dimensions

Gauge theories in four-dimensional spacetime are at the heart of the standard particle physics model. Their topological features have physical consequences and merit careful study.

Yang–Mills Theory

In four dimensions, we define nonabelian electric E^a and magnetic B^a fields,

$$E^{ia} = F^{a}_{0i}, \qquad B^{ia} = -\tfrac{1}{2}\varepsilon^{ijk}F^{a}_{jk} \qquad [16]$$

Canonical analysis and quantization is carried out in the Weyl gauge ($A^a_0 = 0$), where the Lagrangian and Hamiltonian (energy) densities read

$$\mathcal{L}_{\mathrm{YM}} = \tfrac{1}{2}(\boldsymbol{E}^a \cdot \boldsymbol{E}^a - \boldsymbol{B}^a \cdot \boldsymbol{B}^a) \qquad [17]$$

$$\mathcal{H}_{\mathrm{YM}} = \tfrac{1}{2}(\boldsymbol{E}^a \cdot \boldsymbol{E}^a + \boldsymbol{B}^a \cdot \boldsymbol{B}^a) \qquad [18]$$

The first term is kinetic, with $\boldsymbol{E}^a = -\partial_t \boldsymbol{A}^a$ also functioning as the (negative) canonical momentum $\boldsymbol{\pi}^a$, conjugate to the canonical variable $\boldsymbol{A}^a$; the second magnetic term gives the potential. In the Weyl gauge, the theory remains invariant against time-independent gauge transformations. The time component of equation [14] (Gauss law) is absent (because there is no A^a_0 to vary); rather it is imposed as a fixed-time constraint on the canonical variables $\boldsymbol{E}^a$ and $\boldsymbol{A}^a$. This regains the Gauss law:

$$(\boldsymbol{\mathcal{D}} \cdot \boldsymbol{E})^a = 0 \quad \text{(in the absence of sources)} \qquad [19a]$$

In the quantum theory $\boldsymbol{\mathcal{D}} \cdot \boldsymbol{E}$ annihilates "physical" states. Explicitly, in a functional Schrödinger representation, where states are functionals of the canonical fixed-time variable $\boldsymbol{A}|\Psi\rangle \to \Psi(\boldsymbol{A})$, [19a] requires

$$\left(\boldsymbol{\mathcal{D}} \cdot \frac{\delta}{\delta \boldsymbol{A}}\right)^a \Psi(\boldsymbol{A}) = 0 \qquad [19b]$$

that is, physical states must be invariant against infinitesimal gauge transformation, or equivalently, against gauge transformations that are homotopic (continuously deformable) to the identity (the so-called "small" gauge transformations)

$$\Psi(\boldsymbol{A} + \boldsymbol{\mathcal{D}}\lambda) = \Psi(\boldsymbol{A}) \qquad [20]$$

But homotopically nontrivial gauge transformation functions that cannot be deformed to the identity (the so-called "large" gauge transformations) may be present. Their effect is not controlled by Gauss' law, and must be discussed separately.

Fixed-time gauge transformation functions depend on the spatial variable $\boldsymbol{r}: U(\boldsymbol{r})$. For a topological classification, we require that U tend to a constant at large r. Equivalently, we compactify the base space R^3 to S^3. Thus, the gauge functions provide a mapping from S^3 into the relevant gauge group G, and for nonabelian compact gauge groups such mappings fall into disjoint homotopy classes labeled by an integer winding number n: $\Pi^3(G) = Z$. Gauge functions U_n belonging to different classes cannot be deformed into each other; only those in the "zero" class are deformable to the identity. An analytic expression for the winding number $\omega(U)$ is

$$\omega(U) = \frac{1}{24\pi^2}\int \mathrm{d}^3x\, \varepsilon^{ijk}\mathrm{tr}(U^{-1}\partial_i U U^{-1}\partial_j U U^{-1}\partial_k U) \qquad [21]$$

This is a most important topological entity for gauge theories in four-dimensional spacetime, that is, in 3-space, and we shall meet it again in a description of gauge theories in three-dimensional spacetime, that is, on a plane. Various features of ω expose its topological character: (1) $\omega(U)$ does not involve a metric tensor, yet it is diffeomorphism invariant. (2) $\omega(U)$ does not change under local variations of U:

$$\begin{aligned}\delta\omega(U) &= \frac{1}{8\pi^2}\int \mathrm{d}^3x\, \partial_i \varepsilon^{ijk}\mathrm{tr}(U^{-1}\delta U U^{-1}\partial_j U U^{-1}\partial_k U) \\ &= \frac{1}{8\pi^2}\int \mathrm{d}S^i\, \varepsilon^{ijk}\mathrm{tr}(U^{-1}\delta U U^{-1}\partial_j U U^{-1}\partial_k U) \\ &= 0\end{aligned} \qquad [22]$$

The last integral is over the surface (at infinity) bounding the base space and vanishes for localized variations δU. In fact, the entire $\omega(U)$, not only its variation, can be presented as a surface integral, but this requires parametrizing the group element U on R^3. For example, for SU(2),

$$\begin{aligned}&U = \exp\lambda, \quad \lambda = \lambda^a\sigma^a/2\mathrm{i} \quad (\boldsymbol{\sigma} \equiv \text{Pauli matrices}) \\ &\omega(U) = \frac{1}{16\pi^2}\int \mathrm{d}S^i\, \varepsilon^{ijk}\varepsilon_{abc}\hat{\lambda}^a\partial_i\hat{\lambda}^b\partial_k\hat{\lambda}^c(\sin|\lambda| - |\lambda|) \\ &|\lambda| \equiv \sqrt{\lambda^a\lambda^a}, \qquad \hat{\lambda}^a \equiv \lambda^a/|\lambda|\end{aligned} \qquad [23]$$

Specifically, with $|\lambda| \underset{r\to\infty}{\longrightarrow} 2\pi n$ (so that $U \underset{r\to\infty}{\longrightarrow} \pm I$), $\omega(U) = -n$. As befits a topological entity, $\omega(U)$ is determined by global (here large distance) properties of U.

Since all gauge transformations, small and large, are symmetry operations for the theory, [20] should be generalized to

$$\Psi(\boldsymbol{A}^{U_n}) = \mathrm{e}^{\mathrm{i}n\theta}\Psi(\boldsymbol{A}) \qquad [24]$$

where θ is an universal constant. Thus, Yang–Mills quantum states behave as Bloch waves in a periodic lattice, with large gauge transformations playing the role of lattice translations and the Yang–Mills vacuum angle θ playing the role of the Bloch momentum. This is further understood by noting that the profile of the potential energy density, $\frac{1}{2}\boldsymbol{B}^a \cdot \boldsymbol{B}^a$ possesses a periodic structure symbolically depicted in Figure 2.

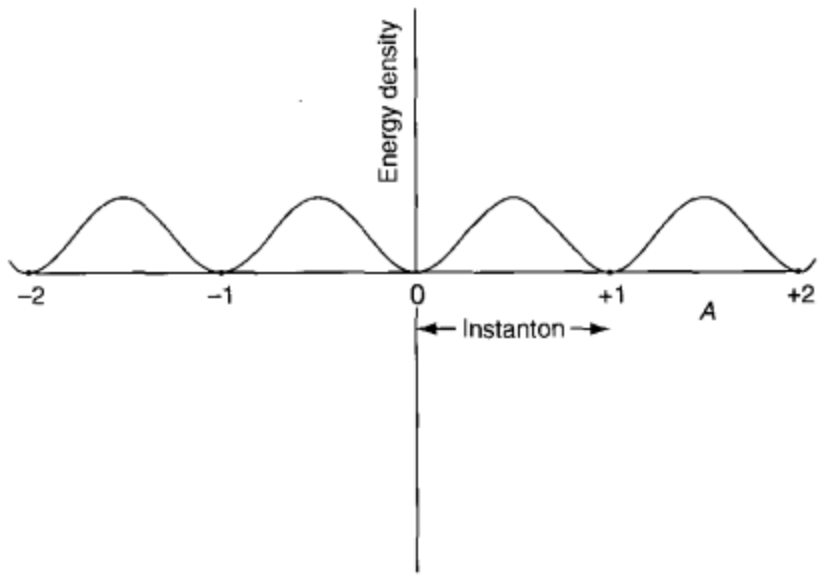


Figure 2 Schematic for energy periodicity of Yang–Mills fields.

Thanks to Gauss' law, potentials A that differ by small gauge transformations are identified, while those differing by large gauge transformations give rise to the periodicity. Zero energy troughs correspond to pure gauge vector potentials in different homotopy classes n: $A = -U_n^{-1}\nabla U_n$.

The θ angle (Bloch momentum) arises from quantum tunneling in A space. Usually, in field theory tunneling is suppressed by infinite energy barriers. (This gives rise to spontaneous symmetry breaking.) However, in Yang–Mills theory there are paths in field space that avoid such barriers. Quantum tunneling paths are exhibited in a semiclassical approximation by identifying classical motion in imaginary time (Euclidean space) that interpolates between classically degenerate vacua and possesses finite action.

In Yang–Mills theory, continuation to imaginary time, $x^0 \to ix^4$, places a factor of i on E^a. Zero (Euclidean) energy is maintained when $E^a = \pm B^a$, or with covariant notation in Euclidean space,

$$\tfrac{1}{2}\varepsilon^{\mu\nu\alpha\beta}F_{\alpha\beta} \equiv {}^*F^{\mu\nu} = \pm F^{\mu\nu} \qquad [25]$$

Euclidean finite action field configurations that satisfy [25] are called self-dual or anti-self-dual instantons. By virtue of the Bianchi identity [10], instantons also solve the field equation [14a] in Euclidean space. Since the Euclidean action may also be written as

$$\begin{aligned} I_{\mathrm{YM}} &= \frac{1}{4}\int \mathrm{d}^4x\,\mathrm{tr}(F^{\mu\nu} \pm {}^*F^{\mu\nu})(F_{\mu\nu} \pm {}^*F_{\mu\nu}) \\ &\quad \mp \frac{1}{2}\int \mathrm{d}^4x\,\mathrm{tr}\,{}^*F^{\mu\nu}F_{\mu\nu} \end{aligned} \qquad [26]$$

and the first term vanishes for instantons, we see that instantons are characterized by the last term, the Chern–Pontryagin index,

$$\begin{aligned} \mathcal{P} &\equiv -\frac{1}{16\pi^2}\int \mathrm{d}^4x\,\mathrm{tr}({}^*F^{\mu\nu}F_{\mu\nu}) \\ &= -\frac{1}{32\pi^2}\int \mathrm{d}^4x\,\varepsilon^{\mu\nu\alpha\beta}\,\mathrm{tr}(F_{\alpha\beta}F_{\mu\nu}) \end{aligned} \qquad [27]$$

This again is an important topological entity:

1. The diffeomorphism invariant $\mathcal{P}$ does not involve the metric tensor.
2. $\mathcal{P}$ is insensitive to local variations of A_μ,

$$\begin{aligned} \delta\mathcal{P} &= -\frac{1}{8\pi^2}\int \mathrm{d}^4x\,\mathrm{tr}({}^*F^{\mu\nu}\delta F_{\mu\nu}) \\ &= -\frac{1}{4\pi^2}\int \mathrm{d}^4x\,\mathrm{tr}({}^*F^{\mu\nu}D_\mu\delta A_\nu) \\ &= \frac{1}{4\pi^2}\int \mathrm{d}^4x\,\mathrm{tr}(D_\mu{}^*F^{\mu\nu}\delta A_\nu) = 0 \end{aligned} \qquad [28]$$

3. $\mathcal{P}$ may be presented as a surface integral owing to the formula

$$\tfrac{1}{4}\mathrm{tr}\,{}^*F^{\mu\nu}F_{\mu\nu} = \partial_\mu K^\mu \qquad [29]$$

$$K^\mu \equiv \varepsilon^{\mu\nu\alpha\beta}\,\mathrm{tr}\left(\tfrac{1}{2}A_\alpha\partial_\beta A_\gamma + \tfrac{1}{3}A_\alpha A_\beta A_\gamma\right) \qquad [30]$$

where K^μ is the Chern–Simons current,

$$\mathcal{P} = -\frac{1}{4\pi^2}\int \mathrm{d}S_\mu K^\mu \qquad [31]$$

The integral [31] is over the base space boundary, S^3. The Chern–Pontryagin index of any gauge field configuration with finite (Euclidean) action (not only instantons) is quantized. This is because finite action requires $F_{\mu\nu}$ to vanish at large distances; equivalently, $A_\mu \to U^{-1}\partial_\mu U$. Using this in [30] renders [31] as

$$\begin{aligned} \mathcal{P} &= \frac{1}{24\pi^2}\int \mathrm{d}S_\mu\varepsilon^{\mu\alpha\beta\gamma} \\ &\quad \times \mathrm{tr}(U^{-1}\partial_\alpha UU^{-1}\partial_\beta UU^{-1}\partial_\gamma U) \end{aligned} \qquad [32]$$

which is the same as [20] and, for the same reason, is given by an integer [$\Pi^3(G) = Z$]. Alternatively, for instantons in the (Euclidean) Weyl gauge ($A_4 = 0$), which interpolate as x^4 passes from $-\infty$ to $+\infty$ between degenerate, classical vacua $A_i = 0$ and $A_i = -U^{-1}\nabla_i U$, $\mathcal{P}$ becomes

$$\begin{aligned} \mathcal{P} &= \frac{1}{4\pi^2}\int \mathrm{d}x^4\,\mathrm{d}^3x\,(\partial_4 K^4 + \nabla\cdot K) \\ &= \frac{1}{4\pi^2}\int \mathrm{d}^3x\,K^4\,|_{x^4=\infty} \\ &= \frac{1}{24\pi^2}\int \mathrm{d}^3x\,\varepsilon^{ijk}\,\mathrm{tr}(U^{-1}\partial_i UU^{-1}\partial_j UU^{-1}\partial U) \\ &= \omega(U) \end{aligned} \qquad [33]$$

We have assumed that the potentials decrease at large arguments sufficiently rapidly so that the gradient term in the first integrand does not contribute. This rederivation of [32] relies on the "motion" of an instanton between vacuum configurations of different winding numbers.

An explicit 1-instanton SU(2) solution ($\mathcal{P}=1$) is

$$A_\mu = \frac{-2\mathrm{i}}{(x-\xi)^2+\rho^2}\sigma_{\mu\nu}x^\nu \qquad [34]$$

(Upon reinserting the coupling constant g, which has been scaled to unity, the field profiles acquire the factor g^{-1}.) In [34], $\sigma_{\mu\nu} \equiv (1/4\mathrm{i})(\sigma_\mu^\dagger\sigma_\nu - \sigma_\mu^\dagger\sigma_\mu), \sigma_\mu \equiv (-\mathrm{i}\sigma, I)$. ξ is the "location" of the instanton, ρ is its "size," and there are three more implicit parameters fixing the gauge, for a total of eight parameters that are needed to specify a single SU(2) instanton. One can show that there exist N instanton/anti-instanton solutions ($\mathcal{P}=N/-N$) and in SU(2) they depend on $8N$ parameters. From [26] we see that at fixed N, instantons minimize the (Euclidean) action. Explicit formulas exist for the most general $N=2$ solution, while for $N\geq 3$ explicit formulas exhibit only $5N+7$ parameters. But algorithms have been found that construct the most general $8N$-parameter instantons. The 1-instanton solution is unchanged by SO(5) rotations, the maximal compact subgroup of the SO(5, 1) conformal invariance group for the Euclidean 4-space Yang–Mills equation [14a].

The Chern–Pontryagin index also appears in the Yang–Mills quantum action, for the following reason. Since all physical states respond to gauge transformations U_n with the universal phase $n\theta$ [24], physical states may be presented in factorized form,

$$\Psi(A) = \mathrm{e}^{\mathrm{i}\mu W(A)\theta}\Phi(A) \qquad [35]$$

where $\Phi(A)$ is invariant against all gauge transformations, small and large, while the phase response is carried by $W(A)$,

$$W(A^{U_n}) = W(A) + n \qquad [36]$$

An explicit expression for $W(A)$ is given by $-(1/4\pi^2)\int \mathrm{d}^3x\, K^0$, where K^0 is the time (fourth) component of K^μ, with dependence on the fourth variable suppressed, that is, K^0 is defined on 3-space,

$$W(A) = -\frac{1}{4\pi^2}\int \mathrm{d}^3x\,\varepsilon^{ijk}\,\mathrm{tr}\left(\tfrac{1}{2}A_i\partial_jA_k + \tfrac{1}{3}A_iA_jA_k\right) \qquad [37]$$

The gauge transformation properties of $W(A)$ are

$$\begin{aligned} W(A^U) &= W(A) + \frac{1}{8\pi^2}\int \mathrm{d}^3x\,\varepsilon^{ijk}\partial_i\,\mathrm{tr}(\partial_jUU^{-1}A_k) \\ &\quad + \frac{1}{24\pi^2}\int \mathrm{d}^3x\,\varepsilon^{ijk}\mathrm{tr}(U^{-1}\partial_iU\;U^{-1}\partial_jU\;U^{-1}\partial_kU) \end{aligned} \qquad [38]$$

The middle surface term does not contribute for well-behaved A; the last term is again $\omega(U)$, the winding number of the gauge transformation U. Thus, [36] is verified.

The universal gauge-varying phase $\mathrm{e}^{\mathrm{i}\theta W(A)}$, which multiplies all gauge-invariant functional states, may be removed at the expense of subtracting from the action

$$\theta\int \mathrm{d}^4x\,\partial_t W(A) = -\frac{\theta}{4\pi^2}\int \mathrm{d}^4x\,\partial_tK^0 = \theta\mathcal{P}$$

(as in [33]). Thus, the Yang–Mills quantum action extends [12] to

$$I_{\mathrm{YM}}{}^{\mathrm{quantum}} = \int \mathrm{d}^4x\,\mathrm{tr}\left(\frac{1}{2}F^{\mu\nu}F_{\mu\nu} + \frac{\theta}{16\pi^2}{}^*F^{\mu\nu}F_{\mu\nu}\right) \qquad [39]$$

The additional Chern–Pontryagin term in [35] does not contribute to equations of motion, but it is needed to render all physical states invariant against all gauge transformations, large and small. With this transformation, one sees that the θ-angle is a Lorentz invariant, but CP noninvariant effect. Evidently, specifying a classical gauge theory requires fixing a group; a quantized gauge theory is specified by a group and a θ-angle, which arises from topological properties of the gauge theory. The energy eigenvalues depend on θ, and distinct θ's correspond to distinct theories.

Note that the reasoning leading to [24] and [39] relies on exact quantum-mechanical arguments, while the instanton-based tunneling discussion is semiclassical.

Adding Fermions

When fermions couple to the gauge fields, the previously described topological effects are modified by action of the chiral anomaly. Dirac fields, either noninteracting but quantized, or unquantized but interacting with a gauge potential through a covariantly conserved current J_a^μ, $\mathcal{L}_{\mathrm{I}} = -J_a^\mu A_\mu^a$, also possess a chiral current $j_5^\mu = \bar{\psi}\gamma^\mu\gamma_5\psi$, which satisfies

$$\partial_\mu j_5^\mu = 2m\,\mathrm{i}\bar{\psi}\gamma_5\psi \qquad [40]$$

Here m is the mass, if any, of the fermions. j_5^μ is conserved for massless fermions, which therefore enjoy a chiral symmetry: $\psi \to e^{i\alpha\gamma_5}\psi$. However, when the interacting fermions are quantized, there arises correction to [40]; this is the chiral anomaly:

$$\partial_\mu \langle j_5^\mu \rangle_A = 2im\langle \bar{\psi}\gamma_5\psi\rangle_A + C\,{}^*F^{\mu\nu\,a}F^a_{\mu\nu} \qquad [41]$$

C is determined by the fermion quantum numbers and coupling strengths. (For a single charged (e) fermion and a U(1) gauge potential, $C = e^2/8\pi^2$.) $\langle\,|\,\rangle_A$ signifies the fermionic vacuum matrix element in the presence of A_μ. The modified equation [41] indicates that even in the massless limit chiral symmetry remains broken due to the anomaly, which arises with quantized fermions.

$\langle j_5^\mu \rangle_A$ may also be presented as

$$\langle j_5^\mu \rangle_A = \mathrm{tr}\,\gamma_5\gamma^\mu \langle \psi\bar{\psi}\rangle_A \qquad [42]$$

In Euclidean space $\langle\psi\bar{\psi}\rangle_A$ is the coincident-point limit of the resolvent $R(x, y; \mu)$ for the Dirac equation,

$$R(x, y; \mu) = \sum_\epsilon \frac{\psi_\epsilon(x)\,\psi_\epsilon^\dagger(y)}{\epsilon + i\mu} \qquad [43]$$

Here ψ_ϵ is an eigenfunction of the massless, Euclidean Dirac operator in the presence of the gauge field A_μ,

$$i\,\gamma^\mu(\partial_\mu + A_\mu)\psi_\epsilon = \epsilon\psi_\epsilon \qquad [44]$$

The coincident-point limit is singular, so R must be regulated: $R \to R - R_{\mathrm{Reg}}$ (we do not specify the regularization procedure). It then follows that

$$\begin{aligned}\partial_\mu\langle j_5^\mu\rangle &= 2i\mu\sum_\epsilon \frac{\psi_\epsilon^\dagger(x)\gamma_5\,\psi_\epsilon(x)}{\epsilon + i\mu} - \mathrm{tr}\,\gamma_5\,\gamma^\mu\partial_\mu R_{\mathrm{Reg}} \\ &= 2i\mu\sum_\epsilon \frac{\psi_\epsilon^\dagger(x)\gamma_5\psi_\epsilon\,(x)}{\epsilon + i\mu} + C\,{}^*F^{\mu\nu\,a}F^a_{\mu\nu} \qquad [45]\end{aligned}$$

The first term on the right-hand side is the (Euclidean space) analog of the mass term in [40] or [41], while the second survives even after the regulators are removed, giving the anomaly $\mathrm{tr}\,{}^*F^{\mu\nu}F_{\mu\nu}$.

The anomaly formula [41], or more explicitly [45], is also the local form of the Atiyah–Singer index theorem, which follows after [45] is integrated over all space: The left-hand side integrates to zero. The integral of the first term on the right-hand side, $\int dx\,\psi_\epsilon^*\gamma_5\psi_\epsilon$, vanishes for $\epsilon \neq 0$ by orthogonality, because $\gamma_5\psi_\epsilon$ is an eigenfunction of [44] with eigenvalue $-\epsilon$. Only zero modes contribute to the ϵ sum since these can be chosen to be eigenfunctions of γ_5, $n_\pm$ of them satisfying $\psi_0 = \pm\gamma_5\psi_0$. For a single multiplet, the normalizations work out so that

$$n_+ - n_- = \frac{1}{16\pi^2}\int d^4x\,\mathrm{tr}\,{}^*F^{\mu\nu}F_{\mu\nu} \qquad [46]$$

The result that the (signed) number of zero modes is the Chern–Pontryagin index is an instance of the Atiyah–Singer theorem. (In specific applications, one can frequently show that n_+ or n_- vanishes.) It, therefore, follows that in the background field of instantons, the Euclidean Dirac equation possesses zero modes.

Another viewpoint on the chiral anomaly arises within the functional integral formulation, where the exponentiated action is constructed from unquantized fields, over which the functional integration is performed. Here the classical action retains chiral symmetry $\psi \to e^{i\alpha\gamma_5}\psi$, but the Grassmann fermion measure $d\psi d\bar{\psi}$, once it is properly regularized, looses chiral invariance and acquires the anomaly,

$$d\psi d\bar{\psi} \to d\psi d\bar{\psi}\,\exp iC\int d^4x\,\alpha\,\mathrm{tr}\,{}^*F^{\mu\nu}F_{\mu\nu} \qquad [47]$$

Evidently, the chiral anomaly involves the gauge-theoretic topological entity, the Chern–Pontryagin density. Not unexpectantly, the anomaly phenomenon affects significantly the topological properties of the gauge theory that are connected to $\mathcal{P}$ and were described previously.

When there is (at least) one massless fermion coupling to the Yang–Mills fields, the Yang–Mills θ-angle looses physical relevance. This is because a chiral transformation that redefines the massless Dirac field does not modify the classical action, but owing to the chiral noninvariance of the functional measure, [47], an anomaly term is induced in the (effective) quantum action. The strength of this induced term can be fixed so that it cancels the θ-term in [39]. Since field redefinition cannot affect physics, the elimination of the θ-term indicates that it had no physical relevance in the first place. In particular, energy eigenvalues no longer depend on θ.

An alternate argument for the same conclusion is based on the functional determinant that arises when the functional integral is performed over the massless Dirac field: $\det[\gamma^\mu(\partial_\mu + A_\mu)]$. The semiclassical tunneling analysis of the θ-angle is based on instantons, but in the presence of instantons the Dirac equation has a zero mode [46]. Consequently the determinant vanishes, tunneling is suppressed and so is the θ-angle.

However, in the standard model for particle physics, there are no massless fermions, so the presence of the θ-angle entails the following physical consequences. The tunneling amplitude Γ in leading semiclassical approximation is determined by the Euclidean action, namely the continuation of iI_{YM} in

[39] to imaginary time. This results in the same expression except that the topological θ-term acquires a factor of i. Only the 1-instanton and anti-instanton give the dominant contribution,

$$\Gamma \propto \cos\theta\, e^{-8\pi^2/g^2} \qquad [48]$$

where the coupling constraint g has been reinserted; the proportionality constant has not been computed, owing to infrared divergences. (Higher-instanton-number configurations contribute at an exponentially subdominant order and have thus far played no role in physics.) The tunneling leads to baryon decay, but fortunately at an exponentially small rate. More useful is the fact that instanton tunneling gives semiclassical evidence for the removal of an unwanted chiral U(1) Goldstone symmetry, which would be present in the standard model if the chiral anomaly did not interfere. Furthermore, the chiral anomaly facilitates the decay of the neutral pion to two photons; a process forbidden by other apparent chiral symmetries of the standard model, which in fact are modified by the chiral anomaly. Gauge fields in four dimensions must interact with anomaly free currents. This necessitates a precise adjustment of fermion content and charges so that the anomaly coefficients (analogs of "C" in [41]) vanish for currents coupled to gauge fields. Finally, θ provides a tantalizing source of CP violation in the strong-interaction sector of the standard model. But no experimental signal (e.g., neutron electric dipole moment) for this effect has been seen. At present, we do not know what mechanism is responsible for keeping θ vanishingly small.

These are the physical consequences of topological effects in four-dimensional gauge theories. Although they have provided experimentalists with only a few numbers to measure (e.g., $\pi^0 \to 2\gamma$ decay amplitude, prediction of anomaly-free arrangements of quarks and leptons in families), they have added enormously to our appreciation of the complexities of quantized gauge theories.

That chiral anomalies are an obstruction to consistent gauge interactions can be established within perturbation theory. A similar, but nonperturbative effect is seen in an SU(2) gauge theory with N Weyl fermion ($\gamma_5\psi = \pm\psi$) SU(2) doublets, which lead upon functional integration to $\det[\gamma^\mu(\partial_\mu + A_\mu)]^{N/2}$. But because $\Pi^4(\mathrm{SU}(2)) = Z_2$, there exists a single homotopy class of gauge transformations which are not deformable to the identity. One shows that the determinant changes sign when such a gauge transformation is performed. Thus, the theory is ill-defined for odd N. Consistent SU(2) gauge theories must possess an even number of Weyl fermion doublets, but such models have not found a place in physical theory.

Adding Bosons

Instantons are finite-action solutions to classical equations continued to imaginary time; they provide a semiclassical description of quantum-mechanical tunneling. A field theory may also possess finite-energy, time-independent (static) solutions to the real-time equations of motion. When these solutions are stable for topological reasons, they are called "solitons." Solitons give semiclassical evidence for the existence in the quantum field theory of a particle sector disjoint from the particles obtained by quantizing field fluctuations around the vacuum state. The soliton particles are heavy for weak coupling g. (Their energy is $O(1/g^2)$; the field profiles are $O(1/g)$.) They do not decay owing to the conservation of "charges" that do not arise from Noether's theorem but are topological.

Yang–Mills theory does not possess soliton solutions (except in five-dimensional spacetime, where the static solitons are just the four-dimensional instantons discussed previously). However, when a gauge theory, based on a simple group is coupled to a scalar field that undergoes symmetry breaking to U(1), soliton solutions exist. These are the 't Hooft–Polyakov magnetic monopoles, found in a SU(2) gauge theory with scalar fields in the adjoint representation, as well as various generalizations. The topological consideration that arises here concerns finite energy of the static, scalar field multiplet φ, which in the Weyl gauge is

$$E(\varphi) = \int d^3x \left(|(\boldsymbol{D}\varphi)^a \cdot (\boldsymbol{D}\varphi)^a|^2 + V(\varphi) \right) \qquad [49]$$

V is non-negative and possesses non trivial symmetry breaking zeroes. On the sphere S^2 at spatial infinity, φ must tend to such a zero. Thus, the fields belong to G/H, where G is the gauge group and H the unbroken subgroup. For the 't Hooft–Polyakov monopole these are SU(2) and U(1), respectively, and the scalar field provides a mapping of the sphere at infinity S^2 to $S^2 \approx \mathrm{SU}(2)/\mathrm{U}(1)$.

One now considers $\Pi^2(S^2) = \Pi^2(\mathrm{SU}(2)/\mathrm{U}(1)) = \Pi^1(\mathrm{U}(1)) = Z$, and one shows that the magnetic flux is determined by the winding number. Hence, the magnetic charge is quantized. Explicitly, the electromagnetic U(1) gauge field is given by

$$\begin{aligned} f_{\mu\nu} &\equiv \hat{\varphi}^a\, {}^*F^a_{\mu\nu} - \varepsilon_{abc}\hat{\varphi}^a (D_\mu\hat{\varphi})^b (D_\nu\hat{\varphi})^c \\ &= \partial_\mu a_\nu - \partial_\nu a_\mu \\ a_\mu &\equiv \hat{\varphi}^a A^a_\mu - \cos\alpha\, \partial_\mu \beta \end{aligned} \qquad [50]$$

where $\hat{\varphi}^a$ is the unit isovector, parametrized as $\hat{\varphi}^a = (\sin\alpha\cos\beta, \sin\alpha\sin\beta, \cos\alpha)$. The manifestly conserved magnetic current

$$j^\mu_m = \partial_\nu {}^*f^{\mu\nu} \qquad [51a]$$

is rearranged to read

$$j^\mu_m = -\tfrac{1}{2}\varepsilon^{\mu\alpha\beta\gamma}\varepsilon_{abc}\partial_\alpha\hat{\varphi}^a\partial_\beta\hat{\varphi}^b\partial_\gamma\hat{\varphi}^c \qquad [51b]$$

and is nonvanishing because φ^a possesses zeroes, where $\partial_\alpha\hat{\varphi}^a$ acquires localized singularities. The magnetic charge

$$m = -\frac{1}{4\pi}\int \mathrm{d}^3x\, j^0_m = \frac{1}{4\pi}\int \mathrm{d}^3x\, \nabla\cdot b \qquad [52]$$

($b^i = \mathrm{U}(1)$ magnetic field: $-\frac{1}{2}\varepsilon^{ijk}f_{jk} = {}^*f^{i0}$) is given by the topological entity (Kronecker index of the mapping)

$$\begin{aligned} m &= \frac{1}{8\pi}\int \mathrm{d}^3x\, \varepsilon^{ijk}\varepsilon_{abc}\partial_i\hat{\varphi}^a\partial_j\hat{\varphi}^b\partial_k\hat{\varphi}^c \\ &= \frac{1}{8\pi}\int \mathrm{d}S^i\, \varepsilon^{ijk}\varepsilon_{abc}\hat{\varphi}^a\partial_j\hat{\varphi}^b\partial_k\hat{\varphi}^c \\ &= -\frac{1}{4\pi}\int \mathrm{d}S^i\, \varepsilon^{ijk}\partial_j\cos\alpha\partial_k\beta \end{aligned} \qquad [53]$$

which readily evaluates the integer winding number.

The theory also supports charged magnetic monopole solutions called "dyons." Here the profiles involve time-periodic gauge potentials, where the time variation is just a gauge transformation $\partial_t A_\mu = D_\mu\lambda$. (Gauge-equivalent, static expressions have slow large-distance fall-off, which is removed by the time-dependent gauge function.) For dyons, the integer valued Chern–Pontryagin index, with the integration taken over all space and in time over the dyon period, reproduces the magnetic monopole strength.

Regrettably, these fascinating structures are not found in nature. Nor do they arise in the standard model, whose structure group is not simple, although speculative grand unified models, with simple G and $H = \mathrm{SU}(3)\times\mathrm{U}(1)$, would support magnetic monopoles and dyons. While challenged physically, the magnetic monopole phenomena have produced extensive and interesting mathematical analysis.

Gauge Theories in Two Dimensions

Two-dimensional gauge theories have only a few physical applications; edge states of the planar quantum Hall effect can be described by excitations moving on a line. However, the abelian model with fermions is useful in that it provides a very accurate reflection of topological behavior in the physically important four-dimensional theory.

Abelian Gauge Theory

Take the spatial interval to be $[-L, L]$. Homotopically nontrivial gauge transformations satisfy $\lambda(L) - \lambda(-L) = 2\pi n$ ($\Pi^1 U(1) = Z$). States $\Psi(A)$ of the free gauge theory that satisfy Gauss' law and respond with a θ-angle are

$$\begin{aligned} \Psi(A) &= \exp\frac{\mathrm{i}\theta}{2\pi}\int \mathrm{d}x\, A \\ \Psi(A + \partial\lambda) &= \mathrm{e}^{\mathrm{i}n\theta}\Psi(A) \end{aligned} \qquad [54]$$

In this model, θ has the interpretation of a constant background electric field $\mathcal{E} = -\theta/2\pi$,

$$\begin{aligned} E\Psi(A) &= \mathcal{E}\Psi(A), \quad E \equiv F_{01} \\ \mathrm{i}\frac{\delta}{\delta A}\Psi(A) &= -\frac{\theta}{2\pi}\Psi(A) \end{aligned} \qquad [55]$$

This also gives the energy eigenvalue:

$$\frac{1}{2}\int \mathrm{d}x\, E^2\Psi(A) = \frac{1}{2}\int \mathrm{d}x\, \mathcal{E}^2\Psi(A) \qquad [56]$$

The phase may be removed by adding to the Lagrangian $-(\theta/2\pi)\int \mathrm{d}x\, \partial_t A$; equivalently, the action becomes

$$I^{\text{quantum}}_{\text{EM}} = \int \mathrm{d}^2x\left(-\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \frac{\theta}{4\pi}\varepsilon^{\mu\nu}F_{\mu\nu}\right) \qquad [57a]$$

which apart from a constant is also given by a formula with the background field:

$$I^{\text{quantum}}_{\text{EM}} = \frac{1}{2}\int \mathrm{d}x(E + \mathcal{E})^2 \qquad [57b]$$

Because of gauge invariance, there is only one state, annihilated by E and carrying energy $\frac{1}{2}\int \mathrm{d}x\, \mathcal{E}^2$. Distinct θ (different $\mathcal{E}$) correspond to distinct theories.

We recognize in [57a] the two-dimensional Chern–Pontryagin density, contributing a total derivative to the action,

$$\mathcal{P} = \frac{1}{4\pi}\int \mathrm{d}^2x\, \varepsilon^{\mu\nu}F_{\mu\nu} \qquad [58]$$

the Chern–Simons current, whose divergence is $\mathcal{P}$,

$$K^\mu = \frac{1}{2\pi}\varepsilon^{\mu\nu}A_\nu \qquad [59]$$

and the Chern–Simons term, which carries the phase of Ψ

$$\int \mathrm{d}x\, K^0 = \frac{1}{2\pi}\int \mathrm{d}x\, A \qquad [60]$$

For Euclidean-space gauge potentials, which are given at large distance by the pure gauge $2\pi n \tan^{-1} y/x$, $\mathcal{P}=n$. All this is just as in the four-dimensional theory, except there are no instantons and no tunneling.

Adding Fermions

The addition of massless fermions to the U(1) gauge theory results in the Schwinger model of massless quantum electrodynamics in two-dimensional spacetime. The equation of motion becomes

$$\partial_\mu F^{\mu\nu} = J^\nu \qquad [61]$$

with the vector current constructed from the Dirac fields as $J^\mu = \bar{\psi}\gamma^\mu\psi$. This current remains conserved in the quantized version because it couples to the gauge field. But the axial vector current $j_5^\mu = \bar{\psi}\gamma^\mu\gamma_5\psi$ acquires an anomaly that involves the Chern–Pontryagin density in [58],

$$\partial_\mu j_5^\mu = \frac{1}{2\pi}\varepsilon^{\mu\nu}F_{\mu\nu} \qquad [62]$$

The model is readily solved, and shows no θ-angle (background field) dependence in physical quantities. The solution is directly obtained by combining [61] with [62] into a second-order differential equation and using the matrix identity of two-dimensional Dirac (=Pauli) matrices: $\varepsilon^{\mu\nu}\gamma_\nu\gamma_5 = \gamma^\mu$. It follows that

$$\left(\Box + \frac{1}{\pi}\right)E = 0 \qquad [63]$$

So the theory describes a free massive photon (mass squared $= 1/\pi$ in units of $\hbar$ and the coupling constant, which have been scaled to unity), with no sign of a θ-angle (background field).

However, in parallel with four-dimensional behavior, the model with massive fermions regains a θ dependence in the particles' energy spectrum; a result that is established perturbatively, because a complete solution is not available.

Note that in the Schwinger model, the gauge particle ("photon") acquires a mass, even though local gauge invariance is preserved. This happens essentially for topological/anomaly reasons. Such topological mass generation is met again in three dimensions.

Adding Bosons

Scalar electrodynamics with a negative mass squared term in $(3+1)$-dimensional spacetime leads to the Higgs mechanism and short-range interactions due to the massive photons. In $(1+1)$ spacetime dimensions, the model possesses instantons – scalar and gauge field profiles that solve the imaginary-time equations of motion – labeled by $\Pi^1(\mathrm{U}(1)) = Z$. These disorder the Higgs condensate so that the force between charged particles remains long-range, like in the positive mass-squared case. This is a vivid example of how excitations arising from nontrivial topological issues significantly effect physical content.

Gauge Theories in Three Dimensions

Gauge theories on three-dimensional spacetime, that is, evolving on a plane, have physical application to planar phenomena, like the quantum Hall effect. Also, the high-temperature limit of four-dimensional field theories is governed by the corresponding field theory in three Euclidean dimensions.

In three (more generally, odd) dimensions, there are no Chern–Pontryagin quantities, no Chern–Simon currents, no axial vector currents or anomalies (there is no γ_5 matrix). These are replaced by odd-dimensional entities that can modify Yang–Mills dynamics.

Yang–Mills and Other Gauge Theories

Using the three-index Levi-Civita tensor, one can construct a gauge-covariant, covariantly conserved vector, which can be added to the Yang–Mills equation. Thus, [14] can be modified to

$$D_\mu F^{\mu\nu} + \frac{m}{2}\varepsilon^{\mu\alpha\beta}F_{\alpha\beta} = J^\nu \qquad [64a]$$

or, equivalently, in terms of the dual-field strength ${}^*F^\mu \equiv \frac{1}{2}\varepsilon^{\mu\alpha\beta}F_{\alpha\beta}$,

$$\varepsilon^{\mu\nu\alpha}D_\mu {}^*F_\alpha + m^*F^\nu = J^\nu \qquad [64b]$$

For dimensional balance, m carries dimension of mass. Indeed, in the source-free case [64] implies

$$(D^\alpha D_\alpha + m^2){}^*F_\mu = \varepsilon_{\mu\alpha\beta}[{}^*F^\alpha, {}^*F^\beta] \qquad [65]$$

This shows that excitations are massive, even though local gauge invariance is preserved. Otherwise, as in the Dirac monopole case, the equations of motion are unexceptional.

However, for the quantum theory we need the action, whose variation produces the mass term in [64]. This is just the Chern–Simons term $W(A)$ in [37], multiplied by $-8\pi^2 m$ and now defined on $(2+1)$-dimensional spacetime:

$$I_{\mathrm{CS}} = 2m\int \mathrm{d}^3x\,\varepsilon^{\alpha\beta\gamma}\mathrm{tr}\left(\tfrac{1}{2}A_\alpha\partial_\beta A_\gamma + \tfrac{1}{3}A_\alpha A_\beta A_\gamma\right) \qquad [66]$$

Everything holds also in the abelian theory; the last term in [66] is then absent.

In this model, the mass is generated by a topological mechanism since I_{CS} possesses the usual attributes for a topological entity: it is diffeomorphisms invariant without a metric tensor; when the potentials are appropriately parametrized, it is given by a surface term. (In the abelian case, the appropriate parametrization is in terms of Clebsch decomposition, $A_\mu = \partial_\mu\theta + \alpha\partial_\mu\beta$.) Most importantly, in the nonabelian theory [66] changes by $8\pi^2 mn$ with three-dimensional gauge transformations carrying winding number n. Hence, for consistency of the nonabelian quantum theory, m must be quantized as $n/4\pi$ (in units of $\hbar$ and the coupling constant, which have been scaled to unity). All this is a clear field-theoretic analog to the quantum mechanics of the Dirac monopole, and just as for the magnetic monopole, a Hamiltonian argument for quantizing m can be constructed, as an alternative to the above action-based derivation.

The time component of [64] relates the electric and magnetic fields to the charge density:

$$D \cdot E - mB = \rho \qquad [67]$$

In the abelian case, the first term involves a total derivative and its spatial integral vanishes, leaving a formula that identifies magnetic flux with a total charge. At low energy, the mass term dominates the conventional kinetic term in [64], and the flux–charge relation becomes a local field-current identity,

$$m^* F^\nu \approx J^\nu \qquad [68]$$

These formulas have made Chern–Simons-modified gauge theories relevant to issues in condensed matter physics, for example, the quantum Hall effect. In the abelian case, m need not be quantized.

Adding Fermions

Three-dimensional Dirac matrices are minimally realized by 2×2 Pauli matrices. As a consequence, a mass term is not parity invariant; also, there is no γ_5 matrix, since the product of the three Dirac (= Pauli) matrices is proportional to I. While there are no chiral anomalies, there is the so-called parity anomaly: integrating a single doublet of massless SU(2) fermions one obtains $\Delta(A) \equiv \det[\gamma^\mu(i\partial_\mu + A_\mu)]$, which should preserve parity and gauge invariance.

Since there are no anomalies in current divergences, $\Delta(A)$ is certainly invariant against infinitesimal gauge transformations. But for finite gauge transformations (categorized by $\Pi^3(\mathrm{SU}(2) = Z)$ one finds that $\Delta(A)$ is not invariant: when the gauge transformation belongs to an odd-numbered homotopy class, $\Delta(A)$ changes sign. To regain gauge invariance, one must either work with an even number of fermion doublets or, if only one doublet (more generally, odd number) is to be used, one must add to the gauge Lagrangian a parity-violating Chern–Simons term with half the correctly quantized coefficient, to neutralize the gauge non-invariance of $\Delta(A)$.

Alternatively, $\Delta(A)$ can be regularized in a gauge-invariant manner. But this requires massive, Pauli–Villars regulator fields, which produce a parity-violating expression for $\Delta(A)$. One cannot avoid the parity anomaly.

Adding Bosons

There are a variety of bosonic field models that one may consider: Abelian or nonabelian; with conventional kinetic term or supplemented by the Chern–Simons topological mass; or, for low energy, no kinetic term but only the Chern–Simons term, as in [68]. Abelian charged Bose fields in a Maxwell theory lead to vortex solitons, based on $\Pi^1(U(1)) = Z$. These are just the instantons of the $(1+1)$-dimensional bosonic gauge theory discussed previously. With Maxwell kinematics there are no charged vortices, but these appear when the Chen–Simons mass is added; see [67]. Pure Chern–Simons kinematics, with no Maxwell term, can produce completely integrable soliton equations (Liouville, Toda) when the Bose field dynamics is appropriately chosen.

Conclusion

Topological effects in field theory are associated with the infinities and regularization that beset quantum field theories. These give rise to the chiral anomaly, parity anomaly (and scale symmetry anomalies, not discussed here). Yet the anomalies themselves are finite quantities that have topological significance (Atiyah–Singer, Chern–Pontryagin, Chern–Simons). This paradoxical pairing has not been understood. Nor can we explain why the anomalies interfere in a topological manner with symmetries associated with masslessness.

Although the range of topological effects in gauge theory is large, and even larger in non-gauge theories (sigma models, Skyrme models) the relevance to actual fundamental physics is confined to the θ-angle phenomenon, which is analyzed accurately and abstractly by reference to $\Pi^3(G)$ and to the interplay with fermions through the chiral anomaly. Instantons are relevant only to an approximate, semiclassical discussion. Although after much mathematical work, general instanton configurations are well understood, only the 1-instanton solution enjoys physical significance.

Other topological entities that fascinate are either nonexistent in fundamental physics or are relevant to

condensed matter physics (vortices, Chern–Simons effects). But here too, we note that the fundamental equation of condensed matter physics – the many-body Schrödinger equation – carries no evident topological structure. Only the phenomenological equations, which replace the fundamental one, give rise to topological intricacies.

Acknowledgment

This work is supported in part by funds provided by the US Department of Energy (DOE) under cooperative research agreement DE-FC02-94ER40818.

See also: Abelian and Nonabelian Gauge Theories Using Differential Forms; Abelian Higgs Vortices; Anomalies; BF Theories; Gauge Theories from Strings; Gauge Theory: Mathematical Applications; Quantum Field Theory: A Brief Introduction; Seiberg–Witten Theory.

Further Reading

Adler SL (1970) Perturbation theory anomalies. In: Deser S, Grisaru M, and Pendleton H (eds.) *Lectures on Elementary Particles and Quantum Field Theory*, vol. 1, pp. 3–164. Cambridge: MIT Press.

Adler SL (2005) Anomalies to all orders (ArXiv: hep-th/0405040). In: 't Hooft G (ed.) *Fifty Years of Yang–Mills Theory*. Singapore: Word Scientific.

Bertlmann RA (1996) *Anomalies in Quantum Field Theory*. Oxford: Clarendon.

Coleman S (1985) Classical lumps and their quantum descendants and the uses of instantons. In: Coleman S *Aspects of Symmetry*, pp. 185–350. Cambridge: Cambridge University Press.

Fujikawa K and Suzuki H (2004) *Path Integrals and Quantum Anomalies*. Oxford: Oxford University Press.

Jackiw R (1977) Quantum meaning of classical field theory. *Reviews of Modern Physics* 49: 681–706.

Jackiw R (1979) Introduction to the Yang–Mills quantum theory. *Reviews of Modern Physics* 52: 661–673.

Jackiw R (1985) Field theoretic investigations in current algebra and topological investigations in quantum gauge theories. In: Treiman S, Jackiw R, Zumino B, and Witten E (eds.) *Current Algebra and Anomalies*, pp. 81–359. Princeton: Princeton University Press; Singapore: World Scientific.

Jackiw R (1995) *Diverse Topics in Theoretical and Mathematical Physics*. Singapore: World Scientific.

Jackiw R (2005) Fifty years of Yang–Mills theory and our moments of triumph (ArXiv: physics/0403109). In: 't Hooft G (ed.) *Fifty Years of Yang–Mills Theory*. Singapore: World Scientific.

Jackiw R and Pi S-Y (1992) Chern–Simons solitons. *Progress of Theoretical Physics* 107: 1–40.

Rajaraman R (1982) *Solitons and Instantons*. Amsterdam: North-Holland.

Shifman M (1994) *Instantons in Gauge Theories*. Singapore: World Scientific.

't Hooft G (1976) Symmetry breaking through Bell–Jackiw anomalies. *Physical Review Letters* 37: 8–11.

Weinberg S (1996) *The Quantum Theory of Fields*. vol. II, chs. 22 and 23. Cambridge: Cambridge University Press.

Perturbative Renormalization Theory and BRST

K Fredenhagen, Universität Hamburg, Hamburg, Germany
M Dütsch, Universität Zürich, Zürich, Switzerland

© 2006 Elsevier Ltd. All rights reserved.

Main Problems in the Perturbative Quantization of Gauge Theories

Gauge theories are field theories in which the basic fields are not directly observable. Field configurations yielding the same observables are connected by a gauge transformation. In the classical theory, the Cauchy problem is well posed for the observables, but in general not for the nonobservable gauge-variant basic fields, due to the existence of time-dependent gauge transformations.

Attempts to quantize the gauge-invariant objects directly have not yet been completely satisfactory. Instead, one modifies the classical action by adding a gauge-fixing term such that standard techniques of perturbative quantization can be applied and such that the dynamics of the gauge-invariant classical fields is not changed. In perturbation theory, this

Notes:

problem shows up already in the quantization of the free gauge fields (see the section "Quantization of free gauge fields"). In the final (interacting) theory the physical quantities should be independent on how the gauge fixing is done ("gauge independence").

Traditionally, the quantization of gauge theories is mostly analyzed in terms of path integrals (e.g., by Faddeev and Popov), where some parts of the arguments are only heuristic. In the original treatment of Becchi, Rouet, and Stora (cf. also Tyutin) (which is called "BRST-quantization"), a restriction to purely massive theories was necessary; the generalization to the massless case by Lowenstein's method is cumbersome.

The BRST quantization is based on earlier work of Feynman, Faddeev, and Popov (introduction of "ghost fields"), and of Slavnov. The basic idea is that after adding a term to the Lagrangian which makes the Cauchy problem well posed but which is not gauge-invariant one enlarges the number of fields by infinitesimal gauge transformations ("ghosts") and their duals ("anti-ghosts"). One then adds a further term to the Lagrangian which contains a coupling of the anti-ghosts and ghosts. The BRST transformation acts as an infinitesimal gauge transformation on the original fields and on the gauge transformations themselves and maps the anti-ghosts to the gauge-fixing terms. This is done in such a way that the total Lagrangian is invariant and that the BRST transformation is nilpotent. The hard problem in the perturbative construction of gauge theories is to show that BRST symmetry can be maintained during renormalization (see the section on perturbative renormalization). By means of the "quantum action principle" of Lowenstein (1971) and Lam (1972, 1973) a cohomological classification of anomalies was worked out (an overview is given, e.g., in the book of Piguet and Sorella (1995)). For more details, *see* BRST Quantization.

The BRST quantization can be carried out in a transparent way in the framework of algebraic quantum field theory (AQFT, *see* Algebraic Approach to Quantum Field Theory). The advantage of this formulation is that it allows one to separate the three main problems of perturbative gauge theories:

1. the elimination of unphysical degrees of freedom,
2. positivity (or "unitarity"), and
3. the problem of infrared divergences.

In AQFT, the procedure is the following: starting from an algebra of all local fields, including the unphysical ones, one shows that after perturbative quantization the algebra admits the BRST transformation as a graded nilpotent derivation. The algebra of observables is then defined as the cohomology of the BRST transformation. To solve the problem of positivity, one has to show that the algebra of observables, in contrast to the algebra of all fields, has a nontrivial representation on a Hilbert space. Finally, one can attack the infrared problem by investigating the asymptotic behavior of states. The latter problem is nontrivial even in quantum electrodynamics (since an electron is accompanied by a "cloud of soft photons") and may be related to confinement in quantum chromodynamics.

The method of BRST quantization is by no means restricted to gauge theories, but applies to general constrained systems. In particular, massive vector fields, where the masses are usually generated by the Higgs mechanism, can alternatively be treated directly by the BRST formalism, in close analogy to the massless case (cf. the section on quantization of free gauge fields).

Local Operator BRST Formalism

In AQFT, the principal object is the family of operator algebras $\mathcal{O} \to \mathcal{A}(\mathcal{O})$ (where $\mathcal{O}$ runs, e.g., through all double cones in Minkowski space), which fulfills the Haag–Kastler axioms (cf. Algebraic Approach to Quantum Field Theory). To construct these algebras, one considers the algebras $\mathcal{F}(\mathcal{O})$ generated by all local fields including ghosts u and anti-ghosts $\tilde{u}$. Ghosts and anti-ghosts are scalar fermionic fields. The algebra gets a $\mathbb{Z}_2$ grading with respect to even and odd ghost numbers, where ghosts get ghost numbers +1 and anti-ghosts ghost number −1. The BRST transformation s acts on these algebras as a $\mathbb{Z}_2$-graded derivation with $s^2 = 0$, $s(\mathcal{F}(\mathcal{O})) \subset \mathcal{F}(\mathcal{O})$, and $s(F^*) = -(-1)^{\delta_F} s(F)^*$, δ_F denoting the ghost number of F.

The observables should be s-invariant and may be identified if they differ by a field in the range of s. Since the range $\mathcal{A}_{00}$ of s is an ideal in the kernel $\mathcal{A}_0$ of s, the algebra of observables is defined as the quotient

$$\mathcal{A} := \mathcal{A}_0/\mathcal{A}_{00} \qquad [1]$$

and the local algebras $\mathcal{A}(\mathcal{O}) \subset \mathcal{A}$ are the images of $\mathcal{A}_0 \cap \mathcal{F}(\mathcal{O})$ under the quotient map $\mathcal{A}_0 \to \mathcal{A}$.

To prove that $\mathcal{A}$ admits a nontrivial representation by operators on a Hilbert space, one may use the BRST operator formalism (Kugo and Ojima 1979, Dütsch and Fredenhagen 1999): one starts from a representation of $\mathcal{F}$ on an inner-product space $(\mathcal{K}, \langle \cdot, \cdot \rangle)$ such that $\langle F^* \phi, \psi \rangle = \langle \phi, F\psi \rangle$

and that s is implemented by an operator Q on $\mathcal{K}$, that is,

$$s(F) = [Q, F] \qquad [2]$$

with $[\cdot, \cdot]$ denoting the graded commutator, such that Q is symmetric and nilpotent. One may then construct the space of physical states as the cohomology of Q, $\mathcal{H} := \mathcal{K}_0/\mathcal{K}_{00}$, where $\mathcal{K}_0$ is the kernel and $\mathcal{K}_{00}$ the range of Q. The algebra of observables now has a natural representation π on $\mathcal{H}$:

$$\pi([A])[\phi] := [A\phi] \qquad [3]$$

(where $A \in \mathcal{A}_0, \phi \in \mathcal{K}_0, [A] := A + \mathcal{A}_{00}, [\phi] := \phi + \mathcal{K}_{00}$). The crucial question is whether the scalar product on $\mathcal{H}$ inherited from $\mathcal{K}$ is positive definite.

In free quantum field theories $(\mathcal{K}, \langle \cdot, \cdot \rangle)$ can be chosen in such a way that the positivity can directly be checked by identifying the physical degrees of freedom (see next section). In interacting theories (see the section on perturbative construction of gauge theories), one may argue in terms of scattering states that the free BRST operator on the asymptotic fields coincides with the BRST operator of the interacting theory. This argument, however, is invalidated by infrared problems in massless gauge theories. Instead, one may use a stability property of the construction.

Namely, let $\tilde{\mathcal{F}}$ be the algebra of formal power series with values in $\mathcal{F}$, and let $\tilde{\mathcal{K}}$ be the vector space of formal power series with values in $\mathcal{K}$. $\tilde{\mathcal{K}}$ possesses a natural inner product with values in the ring of formal power series $\mathbb{C}[[\lambda]]$, as well as a representation of $\tilde{\mathcal{F}}$ by operators. One also assumes that the BRST transformation $\tilde{s}$ is a formal power series $\tilde{s} = \sum_n \lambda^n s_n$ of operators s_n on $\mathcal{F}$ and that the BRST operator $\tilde{Q}$ is a formal power series $\tilde{Q} = \sum_n \lambda^n Q_n$ of operators on $\mathcal{K}$. The algebraic construction can then be done in the same way as before, yielding a representation $\tilde{\pi}$ of the algebra of observables $\tilde{\mathcal{A}}$ by endomorphisms of a $\mathbb{C}[[\lambda]]$ module $\tilde{\mathcal{H}}$, which has an inner product with values in $\mathbb{C}[[\lambda]]$.

One now assumes that at $\lambda = 0$ the inner product is positive, in the sense that

(Positivity)

(i) $\langle \phi, \phi \rangle \geq 0 \quad \forall \phi \in \mathcal{K}$ with $Q_0\phi = 0$, and

(ii) $Q_0\phi = 0 \wedge \langle \phi, \phi \rangle = 0 \Longrightarrow \phi \in Q_0\mathcal{K}$ [4]

Then the inner product on $\tilde{\mathcal{H}}$ is positive in the sense that for all $\tilde{\phi} \in \tilde{\mathcal{H}}$ the inner product with itself, $\langle \tilde{\phi}, \tilde{\phi} \rangle$, is of the form $\tilde{c}^* \tilde{c}$ with some power series $\tilde{c} \in \mathbb{C}[[\lambda]]$, and $\tilde{c} = 0$ iff $\tilde{\phi} = 0$.

This result guarantees that, within perturbation theory, the interacting theory satisfies positivity, provided the unperturbed theory was positive and BRST symmetry is preserved.

Quantization of Free Gauge Fields

The action of a classical free gauge field A,

$$S_0(A) = -\frac{1}{4}\int \mathrm{d}x\, F^{\mu\nu}(x) F_{\mu\nu}(x)$$

$$= \frac{1}{2}\int \mathrm{d}k \hat{A}_\mu(k)^* M^{\mu\nu}(k) \hat{A}_\nu(k) \qquad [5]$$

(where $F^{\mu\nu} := \partial^\mu A^\nu - \partial^\nu A^\mu$ and $M^{\mu\nu}(k) := k^2 g^{\mu\nu} - k^\mu k^\nu$) is unsuited for quantization because $M^{\mu\nu}$ is not invertible: due to $M^{\mu\nu}k_\mu = 0$, it has an eigenvalue 0. Therefore, the action is usually modified by adding a Lorentz-invariant gauge-fixing term: $M^{\mu\nu}$ is replaced by $M^{\mu\nu}(k) + \lambda k^\mu k^\nu$, where $\lambda \in \mathbb{R} \setminus \{0\}$ is an arbitrary constant. The corresponding Euler–Lagrange equation reads

$$\Box A^\mu - (1 - \lambda)\partial^\mu \partial_\nu A^\nu = 0 \qquad [6]$$

For simplicity, let us choose $\lambda = 1$, which is referred to as Feynman gauge. Then the algebra of the free gauge field is the unital $\star$-algebra generated by elements $A^\mu(f)$, $f \in \mathcal{D}(\mathbb{R}^4)$, which fulfill the relations:

$$f \mapsto A^\mu(f) \quad \text{is linear} \qquad [7]$$

$$A^\mu(\Box f) = 0 \qquad [8]$$

$$A^\mu(f)^* = A^\mu(\bar{f}) \qquad [9]$$

$$[A^\mu(f), A^\nu(g)] = \mathrm{i}g^{\mu\nu}\int \mathrm{d}x\, \mathrm{d}y\, f(x) D(x - y) g(y) \qquad [10]$$

where D is the massless Pauli–Jordan distribution.

This algebra does not possess Hilbert space representations which satisfy the microlocal spectrum condition, a condition which in particular requires the singularity of the two-point function to be of the so-called Hadamard form. It possesses, instead, representations on vector spaces with a nondegenerate sequilinear form, for example, the Fock space over a one-particle space with scalar product

$$\langle \phi, \psi \rangle = (2\pi)^{-3}\int \frac{\mathrm{d}^3 \boldsymbol{p}}{2|\boldsymbol{p}|} \overline{\phi^\mu(p)} \psi_\mu(p)|_{p^0 = |\boldsymbol{p}|} \qquad [11]$$

Gupta and Bleuler characterized a subspace of the Fock space on which the scalar product is semidefinite; the space of physical states is then obtained

by dividing out the space of vectors with vanishing norm.

After adding a mass term

$$\frac{m^2}{2}\int \mathrm{d}x A_\mu(x)A^\mu(x)$$

to the action [5], it seems to be no longer necessary to add also a gauge-fixing term. The fields then satisfy the Proca equation

$$\partial_\mu F^{\mu\nu} + m^2 A^\nu = 0 \qquad [12]$$

which is equivalent to the equation $(\Box + m^2)A^\mu = 0$ together with the constraint $\partial_\mu A^\mu = 0$. The Cauchy problem is well posed, and the fields can be represented in a positive-norm Fock space with only physical states (corresponding to the three physical polarizations of A). The problem, however, is that the corresponding propagator admits no power-counting renormalizable perturbation series.

The latter problem can be circumvented in the following way: for the algebra of the free quantum field, one takes only the equation $(\Box + m^2)A^\mu = 0$ into account (or, equivalently, one adds the gauge-fixing term $(1/2)(\partial_\mu A^\mu)^2$ to the Lagrangian) and goes over from the physical field A^μ to

$$B^\mu := A^\mu + \frac{\partial^\mu \phi}{m} \qquad [13]$$

where ϕ is a real scalar field, to the same mass m where the sign of the commutator is reversed ("bosonic ghost field" or "Stückelberg field"). The propagator of B^μ yields a power-counting renormalizable perturbation series; however, B^μ is an unphysical field. One obtains four independent components of B which satisfy the Klein–Gordon equation. The constraint $0 = \partial_\mu A^\mu = \partial_\mu B^\mu + m\phi$ is required for the expectation values in physical states only. So, quantization in the case $m > 0$ can be treated in analogy with [8]–[10] by replacing A^μ by B^μ, the wave operator by the Klein–Gordon operator $(\Box + m^2)$ in [8], and D by the corresponding massive commutator distribution Δ_m in [10]. Again, the algebra can be nontrivially represented on a space with indefinite metric, but not on a Hilbert space.

One can now use the method of BRST quantization in the massless as well as in the massive case. One introduces a pair of fermionic scalar fields (ghost fields) $(u, \tilde{u})$. $u, \tilde{u}$, and (for $m > 0$) ϕ fulfill the Klein–Gordon equation to the same mass $m \geq 0$ as the vector field B. The free BRST transformation reads

$$\begin{aligned} s_0(B^\mu) &= \mathrm{i}\partial^\mu u, \qquad & s_0(\phi) &= \mathrm{i}mu \\ s_0(u) &= 0, \qquad & s_0(\tilde{u}) &= -\mathrm{i}(\partial_\nu B^\nu + m\phi) \end{aligned} \qquad [14]$$

(see, e.g., Scharf (2001)). It is implemented by the free BRST charge

$$Q_0 = \int_{x^0=\text{const.}} \mathrm{d}^3 x j_0^{(0)}(x^0, x) \qquad [15]$$

where

$$j_\mu^{(0)} := (\partial_\nu B^\nu + m\phi)\partial_\mu u - \partial_\mu(\partial_\nu B^\nu + m\phi)u \qquad [16]$$

is the free BRST current, which is conserved. (The interpretation of the integral in [15] requires some care.) Q_0 satisfies the assumptions of the (local) operator BRST formalism, in particular it is nilpotent and positive [4]. Distinguished representatives of the equivalence classes $[\phi] \in \mathrm{Ke}\, Q_0/\mathrm{Ra}\, Q_0$ are the states built up only from the three spatial (two transversal for $m = 0$, respectively) polarizations of A.

Perturbative Renormalization

The starting point for a perturbative construction of an interacting quantum field theory is Dyson's formula for the evolution operator in the interaction picture. To avoid conflicts with Haag's theorem on the nonexistence of the interaction picture in quantum field theory, one multiplies the interaction Lagrangian $\mathcal{L}$ with a test function g and studies the local S-matrix,

$$\begin{aligned} S(g\mathcal{L}) = 1 + \sum_{n=1}^{\infty} \frac{\mathrm{i}^n}{n!}\int \mathrm{d}x_1 \cdots \mathrm{d}x_n g(x_1)\cdots g(x_n) \\ \times T(\mathcal{L}(x_1)\cdots\mathcal{L}(x_n)) \end{aligned} \qquad [17]$$

where T denotes a time-ordering prescription. In the limit $g \to 1$ (adiabatic limit), $S(g\mathcal{L})$ tends to the scattering matrix. This limit, however, is plagued by infrared divergences and does not always exist. Interacting fields $F_{g\mathcal{L}}$ are obtained by the Bogoliubov formula:

$$F_{g\mathcal{L}}(x) = \frac{\delta}{\delta h(x)}|_{h=0} S(g\mathcal{L})^{-1} S(g\mathcal{L} + hF) \qquad [18]$$

The algebraic properties of the interacting fields within a region $\mathcal{O}$ depend only on the interaction within a slightly larger region (Brunetti and Fredenhagen 2000), hence the net of algebras in the sense of AQFT can be constructed in the adiabatic limit without the infrared problems (this is called the "algebraic adiabatic limit").

The construction of the interacting theory is thus reduced to a definition of time-ordered products of fields. This is the program of causal perturbation theory (CPT), which was developed by Epstein and Glaser (1973) on the basis of previous work by Stückelberg and Petermann (1953) and Bogoliubov

and Shirkov (1959). For simplicity, we describe CPT only for a real scalar field. Let φ be a classical real scalar field which is not restricted by any field equation. Let $\mathcal{P}$ denote the algebra of polynomials in φ and all its partial derivatives $\partial^a\varphi$ with multi-indices $a \in \mathbb{N}_0^4$. The time-ordered products $(T_n)_{n\in\mathbb{N}}$ are linear and symmetric maps $T_n : (\mathcal{P} \otimes \mathcal{D}(\mathbb{R}^4))^{\otimes n} \to L(\mathcal{D})$, where $L(\mathcal{D})$ is the space of operators on a dense invariant domain $\mathcal{D}$ in the Fock space of the scalar free field. One often uses the informal notation

$$T_n(g_1F_1 \otimes \cdots \otimes g_nF_n) = \int \mathrm{d}x_1 \cdots \mathrm{d}x_n T_n(F_1(x_1), \ldots, F_n(x_n)) \times g_1(x_1) \cdots g_n(x_n) \qquad [19]$$

where $F_j \in \mathcal{P}$, $g_j \in \mathcal{D}(\mathbb{R}^4)$.

The sequence (T_n) is constructed by induction on n, starting with the initial condition

$$T_1\left(\prod_j \partial^{a_j}\varphi(x)\right) =: \prod_j \partial^{a_j}\phi(x) : \qquad [20]$$

where the right-hand side is a Wick polynomial of the free field ϕ. In the inductive step the requirement of causality plays the main role, that is, the condition that

$$T_n(f_1 \otimes \cdots \otimes f_n) = T_k(f_1 \otimes \cdots \otimes f_k) \times T_{n-k}(f_{k+1} \otimes \cdots \otimes f_n) \qquad [21]$$

if

$$(\operatorname{supp} f_1 \cup \cdots \cup \operatorname{supp} f_k) \cap ((\operatorname{supp} f_{k+1} \cup \cdots \cup \operatorname{supp} f_n) + \bar{V}_-) = \emptyset$$

(where $\bar{V}_-$ is the closed backward light cone). This condition expresses the composition law for evolution operators in a relativistically invariant and local way. Causality determines T_n as an operator-valued distribution on $\mathbb{R}^{4n}$ in terms of the inductively known T_l, $l < n$, outside of the total diagonal $\Delta_n := \{(x_1, \ldots, x_n) \mid x_1 = \cdots = x_n\}$, that is, on test functions from $\mathcal{D}(\mathbb{R}^{4n} \setminus \Delta_n)$.

Perturbative renormalization is now the extension of T_n to the full test function space $\mathcal{D}(\mathbb{R}^{4n})$. Generally, this extension is nonunique. In contrast to other methods of renormalization, no divergences appear, but the ambiguities correspond to the finite renormalizations that persist after removal of divergences by infinite counter terms. The ambiguities can be reduced by (re-)normalization conditions, which means that one requires that certain properties which hold by induction on $\mathcal{D}(\mathbb{R}^{4n} \setminus \Delta_n)$ are maintained in the extension, namely:

- (N0) a bound on the degree of singularity near the total diagonal;
- (N1) Poincaré covariance;
- (N2) unitarity of the local S-matrix;
- (N3) a relation to the time-ordered products of subpolynomials;
- (N4) the field equation for the interacting field $\varphi_{g\mathcal{L}}$ [18];
- (AWI) the "action Ward identity" (Stora 2002, Dütsch and Fredenhagen 2003): $\partial^\mu T(\cdots F_l(x)\cdots) = T(\cdots \partial^\mu F_l(x)\cdots)$. This condition can be understood as the requirement that physics depends on the action only, so total derivatives in the interaction Lagrangian can be removed; and
- further symmetries, in particular in gauge theories, Ward identities expressing BRST invariance. A universal formulation of all symmetries which can be derived from the field equation in classical field theory is the "master Ward identity" (which presupposes (N3) and (N4)) (Boas and Dütsch 2002, Dütsch and Fredenhagen 2003); see next section.

The problem of perturbative renormalization is to construct a solution of all these normalization conditions. Epstein and Glaser have constructed the solutions of (N0)–(N3). Recently, the conditions (N4) and (AWI) have been included. The master Ward identity cannot always be fulfilled, the obstructions being the famous "anomalies" of perturbative quantum field theory.

Perturbative Construction of Gauge Theories

In the case of a purely massive theory, the adiabatic limit $S = \lim_{g\to 1} S(g\mathcal{L})$ exists (Epstein and Glaser 1976), and one may adopt a formalism due to Kugo and Ojima (1979), who use the fact that in these theories the BRST charge Q can be identified with the incoming (free) BRST charge Q_0 [15]. For the scattering matrix S to be a well-defined operator on the physical Hilbert space of the free theory, $\mathcal{H} = \operatorname{Ke} Q_0 / \operatorname{Ra} Q_0$, one then has to require

$$\lim_{g\to 1}[Q_0, T((g\mathcal{L})^{\otimes n})]|_{\ker Q_0} = 0 \qquad [22]$$

This is the motivation for introducing the condition of "perturbative gauge invariance" (Dütsch *et al.* 1993, 1994); see Scharf (2001)): according to this condition, there should exist a Lorentz

vector $\mathcal{L}_1^\nu \in \mathcal{P}$ associated with the interaction $\mathcal{L}$, such that

$$[Q_0, T_n(\mathcal{L}(x_1)\cdots\mathcal{L}(x_n)] = \mathrm{i}\sum_{l=1}^{n}\partial_\nu^{x_l} T_n(\mathcal{L}(x_1)\cdots\mathcal{L}_1^\nu(x_l)\cdots\mathcal{L}(x_n)) \qquad [23]$$

This is a somewhat stronger condition than [22] but has the advantage that it can be formulated independently of the adiabatic limit. The condition [22] (or perturbative gauge invariance) can be satisfied for tree diagrams (i.e., the corresponding requirement in classical field theory can be fulfilled). In the massive case, this is impossible without a modification of the model; the inclusion of additional physical scalar fields (corresponding to Higgs fields) yields a solution. It is gratifying that, by making a polynomial ansatz for the interaction $\mathcal{L} \in \mathcal{P}$, perturbative gauge invariance [23] for tree diagrams, renormalizability (i.e., the mass dimension of $\mathcal{L}$ is ≤ 4), and some obvious requirements (e.g., the Lorentz invariance) determine $\mathcal{L}$ to a far extent. In particular, the Lie-algebraic structure needs not to be put in, as it can be derived in this way (Stora 1997, unpublished). Including loop diagrams (i.e., quantum effects), it has been proved that (N0)–(N2) and perturbative gauge invariance can be fulfilled to all orders for massless SU(N) Yang–Mills theories.

Unfortunately, in the massless case, it is unlikely that the adiabatic limit exists and, hence, an S-matrix formalism is problematic. One should better rely on the construction of local observables in terms of couplings with compact support. However, then the selection of the observables [1] has to be done in terms of the BRST transformation $\tilde{s}$ of the interacting fields.

For the corresponding BRST charge, one makes the ansatz

$$\tilde{Q} = \int \mathrm{d}^4x\, \tilde{j}^\mu_{g\mathcal{L}}(x) b_\mu(x), \quad \mathcal{L} = \sum_{n\geq 1}\mathcal{L}_n\lambda^n \qquad [24]$$

where (b_μ) is a smooth version of the δ-function characterizing a Cauchy surface and $\tilde{j}^\mu_{g\mathcal{L}}$ is the interacting BRST-current [18] (where $\tilde{j}_\mu = \sum_n j_\mu^{(n)}\lambda^n$ ($j_\mu^{(n)} \in \mathcal{P}$) is a formal power series with $j_\mu^{(0)}$ given by [16]). (Note that there is a volume divergence in this integral, which can be avoided by a spatial compactification. This does not change the abstract algebra $\mathcal{F}_\mathcal{L}(\mathcal{O})$.) A crucial requirement is that $\tilde{j}^\mu_{g\mathcal{L}}$ is conserved in a suitable sense. This condition is essentially equivalent to perturbative gauge invariance and hence its application to classical field theory determines the interaction $\mathcal{L}$ in the same way, and in addition the deformation $j^{(0)} \to \tilde{j}_{g\mathcal{L}}$. The latter also gives the interacting BRST charge and transformation, $\tilde{Q}$ and $\tilde{s}$, by [24] and [2]. The so-obtained $\tilde{Q}$ is often nilpotent in classical field theory (and hence this holds also for $\tilde{s}$). However, in QFT conservation of $\tilde{j}_{g\mathcal{L}}$ and $\tilde{Q}^2 = 0$ requires the validity of additional Ward identities, beyond the condition of perturbative gauge invariance [23]. All the necessary identities can be derived from the master Ward identity

$$T_{n+1}(A, F_1, \ldots, F_n) = -\sum_{k=1}^{n} T_n(F_1, \ldots, \delta_A F_k, \ldots, F_n) \qquad [25]$$

where $A = \delta_A S_0$ with a derivation δ_A. The master Ward identity is closely related to the quantum action principle which was formulated in the formalism of generating functionals of Green's functions. In the latter framework, the anomalies have been classified by cohomological methods. The vanishing of anomalies of the BRST symmetry is a selection criterion for physically acceptable models.

In the particular case of QED, the Ward identity

$$\partial_\mu^y T(j^\mu(y)F_1(x_1)\cdots F_n(x_n)) = \mathrm{i}\sum_{j=1}^{n}\delta(y-x_j) \times T\big(F_1(x_1)\cdots(\theta F_j)(x_j)\cdots F_n(x_n)\big) \qquad [26]$$

for the Dirac current $j^\mu := \bar{\psi}\gamma^\mu\psi$, is sufficient for the construction, where $(\theta F) := \mathrm{i}(r-s)F$ for $F = \psi^r\bar{\psi}^s B_1\cdots B_l$ ($B_1, \ldots, B_l$ are nonspinorial fields) and $F_1, \ldots, F_n$ run through all subpolynomials of $\mathcal{L} = j^\mu A_\mu$, (N0)–(N4) and [26] can be fulfilled to all orders (Dütsch and Fredenhagen, 1999).

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Batalin–Vilkovisky Quantization; BRST Quantization; Constrained Systems; Indefinite Metric; Perturbation Theory and its Techniques; Quantum Chromodynamics; Quantum Field Theory: A Brief Introduction; Quantum Fields with Indefinite Metric: Non-Trivial Models; Renormalization: General Theory; Renormalization: Statistical Mechanics and Condensed Matter; Standard Model of Particle Physics.

Further Reading

Becchi C, Rouet A, and Stora R (1975) Renormalization of the abelian Higgs–Kibble model. *Communications in Mathematical Physics* 42: 127.

Becchi C, Rouet A, and Stora R (1976) Renormalization of gauge theories. *Annals of Physics (NY)* 98: 287.

Bogoliubov NN and Shirkov DV (1959) *Introduction to the Theory of Quantized Fields.* New York: Interscience Publishers Inc.

Brunetti R and Fredenhagen K (2000) Microlocal analysis and interacting quantum field theories: renormalization on physical backgrounds. *Communications in Mathematical Physics* 208: 623.

Dütsch M, Hurth T, Krahe K, and Scharf G (1993) Causal construction of Yang–Mills theories. I. *N. Cimento A* 106: 1029.

Dütsch M, Hurth T, Krahe K, and Scharf G (1994) Causal construction of Yang–Mills theories. II. *N Cimento A* 107: 375.

Dütsch M and Fredenhagen K (1999) A local (perturbative) construction of observables in gauge theories: the example of QED. *Communications in Mathematical Physics* 203: 71.

Dütsch M and Boas F-M (2002) The Master Ward identity. *Reviews of Mathematical Physics* 14: 977–1049.

Dütsch M and Fredenhagen K (2003) The master Ward identity and generalized Schwinger–Dyson equation in classical field theory. *Communications in Mathematical Physics* 243: 275.

Dütsch M and Fredenhagen K (2004) Causal perturbation theory in terms of retarded products, and a proof of the action Ward identity. *Reviews in Mathematical Physics* 16: 1291–1348.

Epstein H and Glaser V (1973) *Annals Institut Henri Poincaré A* 19: 211.

Epstein H and Glaser V (1976) Adiabatic limit in perturbation theory. In: Velo G and Wightman AS (eds.) *Renormalization Theory*, pp. 193–254.

Henneaux M and Teitelboim C (1992) *Quantization of Gauge Systems*. Princeton: Princeton University Press.

Kugo T and Ojima I (1979) Local covariant operator formalism of nonabelian gauge theories and quark confinement problem. *Supplement of the Progress of Theoritical Physics* 66: 1.

Lam Y-MP (1972) Perturbation Lagrangian theory for scalar fields – Ward–Takahashi identity and current algebra. *Physics Reviews D* 6: 2145.

Lam Y-MP (1973) Equivalence theorem on Bogoliubov–Parasiuk–Hepp–Zimmermann – renormalized Lagrangian field theories. *Physics Reviews D* 7: 2943.

Lowenstein JH (1971) Differential vertex operations in Lagrangian field theory. *Communications in Mathematical Physics* 24: 1.

Piguet O and Sorella S (1995) Algebraic Renormalization: Perturbative Renormalization, Symmetries and Anomalies, Lecture Notes in Physics. Berlin: Springer.

Scharf G (1995) *Finite Quantum Electrodynamics. The Causal Approach*, 2nd edn. Berlin: Springer.

Scharf G (2001) *Quantum Gauge Theories – A True Ghost Story*. New York: Wiley.

Stora R (2002) Pedagogical experiments in renormalized perturbation theory. Contribution to the conference. *Theory of Renormalization and Regularization*. Germany: Hesselberg.

Stückelberg ECG and Petermann A (1953) La normalisation des constantes dans la theorie des quanta. *Helvetica Physica Acta* 26: 499–520.

Weinberg S (1996) *The Quantum Theory of Fields*. Cambridge: Cambridge University Press.

Quantum Chromodynamics

G Sterman, Stony Brook University, Stony Brook, NY, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Quantum chromodynamics, or QCD, as it is normally called in high-energy physics, is the quantum field theory that describes the strong interactions. It is the SU(3) gauge theory of the current standard model for elementary particles and forces, $SU(3) \times SU(2)_L \times U(1)$, which encompasses the strong, electromagnetic, and weak interactions. The symmetry group of QCD, with its eight conserved charges, is referred to as color SU(3). As is characteristic of quantum field theories, each field may be described in terms of quantum waves or particles.

Because it is a gauge field theory, the fields that carry the forces of QCD transform as vectors under the Lorentz group. Corresponding to these vector fields are the particles called "gluons," which carry an intrinsic angular momentum, or spin, of 1 in units of $\hbar$. The strong interactions are understood as the cumulative effects of gluons, interacting among themselves and with the quarks, the spin-1/2 particles of the Dirac quark fields.

There are six quark fields of varying masses in QCD. Of these, three are called "light" quarks, in a sense to be defined below, and three "heavy." The light quarks are the up (u), down (d), and strange (s), while the heavy quarks are the charm (c), bottom (b),

Notes:

and top (t). Their well-known electric charges are $e_f = 2e/3(u, c, t)$ and $e_f = -e/3(d, s, b)$, with e the positron charge. The gluons interact with each quark field in an identical fashion, and the relatively light masses of three of the quarks provide the theory with a number of approximate global symmetries that profoundly influence the manner in which QCD manifests itself in the standard model.

These quark and gluon fields and their corresponding particles are enumerated with complete confidence by the community of high-energy physicists. Yet, none of these particles has ever been observed in isolation, as one might observe a photon or an electron. Rather, all known strongly interacting particles are colorless; most are "mesons," combinations with the quantum numbers of a quark q and a antiquark $\bar{q}'$, or "baryons" with the quantum numbers of (possibly distinct) combinations of three quarks $qq'q''$. This feature of QCD, that its underlying fields never appear as asymptotic states, is called "confinement." The very existence of confinement required new ways of thinking about field theory, and only with these was the discovery and development of QCD possible.

The Background of QCD

The strong interactions have been recognized as a separate force of nature since the discovery of the neutron as a constituent of atomic nuclei, along with the proton. Neutrons and protons (collectively, nucleons) possess a force, attractive at intermediate distances and so strong that it overcomes the electric repulsion of the protons, each with charge e. A sense of the relative strengths of the electromagnetic and strong interactions may be inferred from the typical distance between mutually repulsive electrons in an atom, $\sim 10^{-8}$ cm, and the typical distance between protons in a nucleus, of order 10^{-13} cm.

The history that led up to the discovery of QCD is a fascinating one, beginning with Yukawa's 1935 theory of pion exchange as the source of the forces that bind nuclei, still a useful tool for low-energy scattering. Other turning points include the creation of nonabelian gauge theories by Yang and Mills in 1954, the discovery of the quantum number known as strangeness, the consequent development of the quark model, and then the proposal of color as a global symmetry. The role of pointlike constituents in hadrons was foreshadowed by the identification of electromagnetic and weak currents and the analysis of their quantum-mechanical algebras. Finally, the observation of "scaling" in deep-inelastic scattering, which we will describe below, made QCD, with color as a local symmetry, the unique explanation of the strong interactions, through its property of asymptotic freedom.

The Lagrangian and Its Symmetries

The QCD Lagrangian may be written as

$$\mathcal{L} = \sum_{f=1}^{n_f} \bar{q}_f (\mathrm{i}\, \not{D}[A] - m_f) q_f - \frac{1}{2}\,\mathrm{tr}\left[F_{\mu\nu}^2(A)\right] - \frac{\lambda}{2}(B_a(A))^2 + \bar{c}_b \left[\frac{\delta B_b(A)}{\delta \alpha_a}\right] c_a \qquad [1]$$

with $\not{D}[A] = \gamma \cdot \partial + \mathrm{i} g_s \gamma \cdot A$ the covariant derivative in QCD. The γ^μ are the Dirac matrices, satisfying the anticommutation relations, $[\gamma^\mu, \gamma^\nu]_+ = 2g^{\mu\nu}$. The SU(3) gluon fields are $A^\mu = \sum_{a=1}^{8} A_a^\mu T_a$, where T_a are the generators of SU(3) in the fundamental representation. The field strengths $F_{\mu\nu}[A] = \partial_\mu A_\nu - \partial_\nu A_\mu + \mathrm{i} g_s [A_\mu, A_\nu]$ specify the three- and four-point gluon couplings of nonabelian gauge theory. In QCD, there are $n_f = 6$ flavors of quark fields, q_f, with conjugate $\bar{q}_f = q_f^\dagger \gamma^0$.

The first two terms in the expression [1] make up the classical Lagrangian, followed by the gauge-fixing term, specified by a (usually, but not necessarily linear) function $B_a(A)$, and the ghost Lagrangian. The ghost (anti-ghost) fields $c_a(\bar{c}_a)$ carry the same adjoint index as the gauge fields.

The classical QCD Lagrangian before gauge fixing is invariant under the local gauge transformations

$$\begin{aligned} A'_\mu(x) &= \frac{\mathrm{i}}{g_s}\partial_\mu \Omega(x)\Omega^{-1}(x) + \Omega(x) A'_\mu(x)\Omega^{-1}(x) \\ &= A_\mu(x) - \partial_\mu \delta\alpha(x) \\ &\quad + \mathrm{i} g_s [\delta\alpha(x), A_\mu(x)] + \cdots \\ \psi'_i(x) &= \Omega(x)_{ij}\psi_j(x) = \psi_i(x) \\ &\quad + \mathrm{i} g_s \delta\alpha(x)_{ij}\psi_j(x) + \cdots \\ \delta\alpha(x) &= \sum_{a=1}^{8} \delta\alpha_a(x) T_a \end{aligned} \qquad [2]$$

The full QCD action including gauge-fixing and ghost terms is also invariant under the Bechi, Rouet, Stora, Tyutin (BRST) transformations with $\delta\xi$ an anticommuting variable.

$$\begin{aligned} \delta A_{\mu,a} &= (\delta_{ab}\partial_\mu + g A_{\mu c} f_{abc}) c_b \delta\xi \\ \delta c_a &= -\tfrac{1}{2} g C_{abc}\, c_b\, c_c \delta\xi, \quad \delta\bar{c} = \lambda B_a \delta\xi \\ \delta\psi_i &= \mathrm{i} g [T_b]_{ij}\, c_b \psi_j \end{aligned} \qquad [3]$$

with f_{abc} the SU(3) structure constants. The Jacobian of these transformations is unity.

In addition, neglecting masses of the light quarks, u, d, and s, the QCD Lagranian has a class of global flavor and chiral symmetries, the latter connecting left- and right-handed components of the quark fields, $\psi_{\mathrm{L,R}} \equiv (1/2)(1 \mp \gamma_5)\psi$,

$$\psi'(x) = \mathrm{e}^{\mathrm{i}\alpha\gamma_5^P}\psi(x), \quad P = 0, 1 \qquad [4]$$

Here, power $P=0$ describes phase, and $P=1$ chiral, transformations. Both transformations can be extended to transformations among the light flavors, by letting ψ become a vector, and α an element in the Lie algebra of SU(M), with $M=2$ if we take only the u and d quarks, and $M=3$ if we include the somewhat heavier strange quark. These symmetries, not to be confused with the local symmetries of the standard model, are strong isospin and its extension to the "eightfold way," which evolved into the (3-)quark model of Gell-Mann and Zweig. The many successes of these formalisms are automatically incorporated into QCD.

Green Functions, Phases, and Gauge Invariance

In large part, the business of quantum field theory is to calculate Green functions,

$$G_n(x_1 \ldots x_n) = \langle 0|T(\Phi_1(x_1)\ldots\Phi_i(x_i)\ldots\Phi_n(x_n))|0\rangle \quad [5]$$

where T denotes time ordering. The $\Phi_i(x)$ are elementary fields, such as A or q_f, or composite fields, such as currents like $J^\mu=\bar{q}_f\gamma^\mu q_f$. Such a Green function generates amplitudes for the scattering of particles of definite momenta and spin, when in the limit of large times the x_i-dependence of the Green function is that of a plane wave. For example, we may have in the limit $x_i^0 \to \infty$,

$$G_n(x_1 \ldots x_n) \to \phi_i(p,\lambda)\, e^{ip\cdot x_i}\langle(p,\lambda)|T(\Phi_1(x_1)\ldots \Phi_{i-1}(x_{i-1})\Phi_{i+1}(x_{i+1})\ldots\Phi_n(x_n))|0\rangle \quad [6]$$

where $\phi_i(p,\lambda)$ is a solution to the free-field equation for field Φ_i, characterized by momentum p and spin λ. (An inegral over possible momenta p is understood.) When this happens for field i, the vacuum state is replaced by $|(p,\lambda)\rangle$, a particle state with precisely this momentum and spin; when it occurs for all fields, we derive a scattering (S)-matrix amplitude. In essence, the statement of confinement is that Green functions with fields $q_f(x)$ never behave as plane waves at large times in the past or future. Only Green functions of color singlet composite fields, invariant under gauge transformations, are associated with plane wave behavior at large times.

Green functions remain invariant under the BRST transformations [3], and this invariance implies a set of Ward identities

$$\frac{\delta}{\delta\xi(z)}\sum_{i=1}^{n}\langle 0|T(\Phi_1(x_1)\ldots\delta_{\mathrm{BRS}}\Phi_i(x_i)\ldots \Phi_n(x_n))|0\rangle = 0 \quad [7]$$

The variation of the anti-ghost as in [3] is equivalent to an infinitesimal change in the gauge-fixing term; variations in the remaining fields all cancel single-particle plane wave behavior in the corresponding Green functions. These identities then ensure the gauge invariance of the perturbative S-matrix, a result that turns out to be useful despite confinement.

To go beyond a purely perturbative description of QCD, it is useful to introduce a set of nonlocal operators that are variously called nonabelian phases, ordered exponentials, and Wilson lines,

$$U_C(z,y) = P\,\exp\left[-ig_s\int_y^z dx^\mu A_\mu(x)\right] \quad [8]$$

where C is some self-avoiding curve between y and z. The U's transform at each end linearly in nonabelian gauge transformations $\Omega(x)$ at that point,

$$U'_C(z,y) = \Omega(z)U_C(z,y)\Omega^{-1}(y) \quad [9]$$

Especially interesting are closed curves C, for which $z=y$. The phases about such closed loops are, like their abelian counterparts, sensitive to the magnetic flux that they enclose, even when the field strengths vanish on the curve.

QCD at the Shortest and Longest Distances

Much of the fascination of QCD is its extraordinary variation of behavior at differing distance scales. Its discovery is linked to asymptotic freedom, which characterizes the theory at the shortest scales. Asymptotic freedom also suggests (and in part provides) a bridge to longer distances.

Most analyses in QCD begin with a path-integral formulation in terms of the elementary fields $\Phi_a = q_f \ldots$,

$$G_n(x_i,(z_j,y_j)) = \int\left[\prod_{a=q,\bar{q},G,c,\bar{c}}\mathcal{D}\Phi_a\right]\prod_i\Phi_i(x_i) \times\prod_j U_{C_j}(z_j,y_j)\,e^{iS_{\mathrm{QCD}}} \quad [10]$$

with S_{QCD} the action. Perturbation theory keeps only the kinetic Lagrangian, quadratic in fields, in the exponent, and expands the potential terms in the coupling. This procedure produces Feynman diagrams, with vertices corresponding to the cubic and quartic terms in the QCD Lagrangian [1].

Most nonperturbative analyses of QCD require studying the theory on a Euclean, rather than Minkowski space, related by an analytic continuation in the times x^0, y^0, z^0 in G_n from real to imaginary values. In Euclidean space, we find, for example,

classical solutions to the equations of motion, known as instantons, that provide nonperturbative contributions to the path integral. Perhaps the most flexible nonperturbative approach approximates the action and the measure at a lattice of points in four-dimensional space. For this purpose, integrals over the gauge fields are replaced by averages over "gauge links," of the form of eqn [8] between neighboring points.

Perturbation theory is most useful for processes that occur over short timescales and at high relative energies. Lattice QCD, on the other hand, can simulate processes that take much longer times, but is less useful when large momentum transfers are involved. The gap between the two methods remains quite wide, but between the two they have covered enormous ground, enough to more than confirm QCD as the theory of strong interactions.

Asymptotic Freedom

QCD is a renormalizable field theory, which implies that the coupling constant g must be defined by its value at a "renormalization scale," and is denoted $g(\mu)$. Usually, the magnitude of $\alpha_s(\mu) \equiv g^2/4\pi$, is quoted at $\mu = m_Z$, where it is ~ 0.12. In effect, $g(\mu)$ controls the amplitude that connects any state to another state with one more or one fewer gluon, including quantum corrections that occur over timescales from zero up to $\hbar/\mu$ (if we measure μ in units of energy). The QCD Ward identities mentioned above ensure that the coupling is the same for both quarks and gluons, and indeed remains the same in all terms in the Lagrangian, ensuring that the symmetries of QCD are not destroyed by renormalization.

Quantum corrections to gluon emission are not generally computable directly in renormalizable theories, but their dependence on μ is computable, and is a power series in $\alpha_s(\mu)$ itself,

$$\mu^2 \frac{d\alpha_s(\mu)}{d\mu^2} = -b_0 \frac{\alpha_s^2(\mu)}{4\pi} - b_1 \frac{\alpha_s^3(\mu)}{(4\pi)^2} + \cdots \equiv \beta(\alpha_s) \quad [11]$$

where $b_0 = 11 - 2n_f/3$ and $b_1 = 2(31 - 19n_f/3)$. The celebrated minus signs on the right-hand side are associated with both the spin and self-interactions of the gluons.

The solution to this equation provides an expression for α_s at any scale μ_1 in terms of its value at any other scale μ_0. Keeping only the lowest-order, b_0, term, we have

$$\alpha_s(\mu_1) = \frac{\alpha_s(\mu_0)}{1 + (b_0/4\pi)\ln(\mu_1^2/\mu_0^2)} = \frac{4\pi}{b_0 \ln\left(\mu_1^2/\Lambda_{QCD}^2\right)} \quad [12]$$

where in the second form, we have introduced Λ_{QCD}, the scale parameter of the theory, which embodies the condition that we get the same coupling at scale μ_1 no matter which scale μ_0 we start from. Asymptotic freedom consists of the observation that at larger renormalization masses μ, or correspondingly shorter timescales, the coupling weakens, and indeed vanishes in the limit $\mu \to \infty$. The other side of the coin is that over longer times or lower momenta, the coupling grows. Eventually, near the pole at $\mu_1 = \Lambda_{QCD}$, the lowest-order approximation to the running fails, and the theory becomes essentially nonperturbative. Thus, the discovery of asymptotic freedom suggested, although it certainly does not prove, that QCD is capable of producing very strong forces, and confinement at long distances. Current estimates of Λ_{QCD} are $\sim 200\,\text{MeV}$.

Spontaneous Breaking of Chiral Symmetry

The number of quarks and their masses is an external input to QCD. In the standard model masses are provided by the Higgs mechanism, but in QCD they are simply parameters. Because the standard model has chosen several of the quarks to be especially light, QCD incorporates the chiral symmetries implied by eqn [4] (with $P = 1$). In the limit of zero quark masses, these symmetries becomes exact, respected to all orders of perturbation theory, that is, for any finite number of gluons emitted or absorbed.

At distances on the order to $1/\Lambda_{QCD}$, however, QCD cannot respect chiral symmetry, which would require each state to have a degenerate partner with the opposite parity, something not seen in nature. Rather, QCD produces, nonperturbatively, nonzero values for matrix elements that mix right- and left-handed fields, such as $\langle 0|\bar{u}_L u_R|0\rangle$, with u the up-quark field. Pions are the Goldstone bosons of this symmetry, and may be thought of as ripples in the chiral condensate, rotating it locally as they pass along. The observation that these Goldstone bosons are not exactly massless is due to the "current" masses of the quarks, their values in $\mathcal{L}_{QCD}$. The (chiral perturbation theory) expansion in these light-quark masses also enables us to estimate them quantitatively: $1.5 \le m_u \le 4\,\text{MeV}$, $4 \le m_d \le 8\,\text{MeV}$, and $80 \le m_s \le 155\,\text{MeV}$. These are the light quarks, with masses smaller than Λ_{QCD}. (Like α_s, the masses are renormalized; these are quoted from Eidelman (2004) with $\mu = 2\,\text{GeV}$.) For comparison, the heavy quarks have masses $m_c \sim 1$–$1.5\,\text{GeV}$, $m_b \sim 4$–$4.5\,\text{GeV}$, and $m_t \sim 180\,\text{GeV}$ (the giant among the known elementary particles).

Although the mechanism of the chiral condensate (and in general other nonperturbative aspects of

QCD) has not yet been demonstrated from first principles, a very satisfactory description of the origin of the condensate, and indeed of much hadronic structure, has been given in terms of the attractive forces between quarks provided by instantons. The actions of instanton solutions provide a dependence $\exp[-8\pi^2/g_s^2]$ in Euclidean path integrals, and so are characteristically nonperturbative.

Mechanisms of Confinement

As described above, confinement is the absence of asymptotic states that transform nontrivially under color transformations. The full spectrum of QCD, however, is a complex thing to study, and so the problem has been approached somewhat indirectly. A difficulty is the same light-quark masses associated with approximate chiral symmetry. Because the masses of the light quarks are far below the scale Λ_{QCD} at which the perturbative coupling blows up, light quarks are created freely from the vacuum and the process of "hadronization," by which quarks and gluons form mesons and baryons, is both nonperturbative and relativistic. It is therefore difficult to approach in both perturbation theory and lattice simulations.

Tests and studies of confinement are thus normally formulated in truncations of QCD, typically with no light quarks. The question is then reformulated in a way that is somewhat more tractable, without relativistic light quarks popping in and out of the vacuum all the time. In the limit that its mass becomes infinite compared to the natural scale of fluctuations in the QCD vacuum, the propagator of a quark becomes identical to a phase operator, [8], with a path C corresponding to a constant velocity. This observation suggests a number of tests for confinement that can be implemented in the lattice theory. The most intuitive is the vacuum expectation value of a "Wilson loop," consisting of a rectangular path, with sides along the time direction, corresponding to a heavy quark and antiquark at rest a distance R apart, and closed at some starting and ending times with straight lines. The vacuum expectation value of the loop then turns out to be the exponential of the potential energy between the quark pair, multiplied by the elapsed time,

$$\left\langle 0 \left| P \exp\left[-ig_s \oint_C A_\mu(x)\, dx^\mu\right] \right| 0 \right\rangle = \exp(-V(R)T/\hbar) \qquad [13]$$

When $V(R) \propto R$ ("area law" behavior), there is a linearly rising, confining potential. This behavior, not yet proven analytically yet well confirmed on the lattice, has an appealing interpretation as the energy of a "string," connecting the quark and antiquark, whose energy is proportional to its length.

Motivation for such a string picture was also found from the hadron spectrum itself, before any of the heavy quarks were known, and even before the discovery of QCD, from the observation that many mesonic $(\bar{q}q')$ states lie along "Regge trajectories," which consist of sets of states of spin J and mass m_J^2 that obey a relation

$$J = \alpha' m_J^2 \qquad [14]$$

for some constant α'. Such a relation can be modeled by two light particles ("quarks") revolving around each other at some constant (for simplicity, fixed nonrelativistic) velocity v_0 and distance $2R$, connected by a "string" whose energy per unit length is a constant ρ.

Suppose the center of the string is stationary, so the overall system is at rest. Then neglecting the masses, the total energy of the system is $M = 2R\rho$. Meanwhile, the momentum density per unit length at distance r from the center is $v(r) = (r/R)v_0$, and the total angular momentum of the system is

$$J = 2\rho v_0 \int_0^R dr\, r^2 = \frac{2\rho v_0}{3} R^2 = \frac{v_0}{6\rho} M^2 \qquad [15]$$

and for such a system, [14] is indeed satisfied. Quantized values of angular momentum J give quantized masses m_J, and we might take this as a sort of "Bohr model" for a meson. Indeed, string theory has its origin in related consideration in the strong interactions.

Lattice data are unequivocal on the linearly rising potential, but it requires further analysis to take a lattice result and determine what field configurations, stringlike or not, gave that result. Probably the most widely accepted explanation is in terms of an analogy to the Meissner effect in superconductivity, in which type II superconductors isolate magnetic flux in quantized tubes, the result of the formation of a condensate of Cooper pairs of electrons. If the strings of QCD are to be made of the gauge field, they must be electric ($F^{\mu 0}$) in nature to couple to quarks, so the analogy postulates a "dual" Meissner effect, in which electric flux is isolated as the result of a condensate of objects with magnetic charge (producing nonzero F^{ij}). Although no proof of this mechanism has been provided yet, the role of magnetic fluctuations in confinement has been widely investigated in lattice simulations, with encouraging results. Of special interest are magnetic field configurations, monopoles or vortices, in the Z_3 center of $SU(3)$, $\exp[i\pi k/3]I_{3\times 3}$, $k = 0, 1, 2$. Such configurations, even when localized, influence closed gauge loops [13] through the nonabelian Aharonov–Bohm effect. Eventually, of course, the role of light quarks must be crucial for any complete

description of confinement in the real world, as emphasized by Gribov.

Another related choice of closed loop is the "Polyakov loop," implemented at finite temperature, for which the path integral is taken over periodic field configurations with period $1/T$, where T is the temperature. In this case, the curve C extends from times $t=0$ to $t=1/T$ at a fixed point in space. In this formulation it is possible to observe a phase transition from a confined phase, where the expectation is zero, to a deconfined phase, where it is nonzero. This phase transition is currently under intense experimental study in nuclear collisions.

Using Asymptotic Freedom: Perturbative QCD

It is not entirely obvious how to use asymptotic freedom in a theory that should (must) have confinement. Such applications of asymptotic freedom go by the term perturbative QCD, which has many applications, not the least as a window to extensions of the standard model.

Lepton Annihilation and Infrared Safety

The electromagnetic current, $J_\mu = \sum_f e_f \bar{q}_f \gamma_\mu q_f$, is a gauge-invariant operator, and its correlation functions are not limited by confinement. Perhaps, the simplest application of asymptotic freedom, yet of great physical relevance, is the scalar two-point function,

$$\pi(Q) = \frac{-i}{3}\int d^4x\, e^{-iQ\cdot x}\langle 0|T(J^\mu(0)J_\mu(x))|0\rangle \quad [16]$$

The imaginary part of this function is related to the total cross section for the annihilation process $e^+e^- \rightarrow$ hadrons in the approximation that only one photon takes part in the reaction. The specific relation is $\sigma_{QCD} = (e^4/Q^2)\ \mathrm{Im}\,\pi(Q^2)$, which follows from the optical theorem, illustrated in **Figure 1**. The perturbative expansion of the function $\pi(Q)$ depends, in general, on the mass scales Q and the quark masses m_f as well as on the strong coupling $\alpha_s(\mu)$ and on the renormalization scale μ. We may also worry about the influence of other, truly nonperturbative scales, proportional to powers of Λ_{QCD}. At large values of Q^2, however, the situation simplifies greatly, and dependence on all scales below Q is suppressed by powers of Q. This may be expressed in terms of the operator product expansion,

$$\langle 0|T(J^\mu(0)J_\mu(x))|0\rangle = \sum_{O_I}(x^2)^{-3+d_I/2}C_I(x^2\mu^2, \alpha_s(\mu)) \times \langle 0|O_I(0)|0\rangle \quad [17]$$

where d_I is the mass dimension of operator O_I, and where the dimensionless coefficient functions C_I incorporate quantum corrections. The sum over operators begins with the identity ($d_I = 0$), whose coefficient function is identified with the sum of quantum corrections in the approximation of zero masses. The sum continues with quark mass corrections, which are suppressed by powers of at least m_f^2/Q^2, for those flavors with masses below Q. Any QCD quantity that has this property, remaining finite in perturbation theory when all particle masses are set to zero, is said to be "infrared safe."

The effects of quarks whose masses are above Q are included indirectly, through the couplings and masses observed at the lower scales. In summary, the leading power behavior of $\pi(Q)$, and hence of the cross section, is a function of Q, μ, and $\alpha_s(\mu)$ only. Higher-order operators whose vacuum matrix elements receive nonperturbative corrections include the "gluon condensate," identified as the product $\alpha_s(\mu)G_{\alpha\beta}G^{\alpha\beta} \propto \Lambda^4_{QCD}$.

Once we have concluded that Q is the only physical scale in π, we may expect that the right choice of the renormalization scale is $\mu = Q$. Any observable quantity is independent of the choice of renormalization scale, μ, and neglecting quark masses, the chain rule gives

$$\mu\frac{d\sigma(Q/\mu, \alpha_s(\mu))}{d\mu} = \mu\frac{\partial\sigma}{\partial\mu} + 2\beta(\alpha_s)\frac{\partial\sigma}{\partial\alpha_s} = 0 \quad [18]$$

which shows that we can determine the beta function directly from the perturbative expansion of the cross section. Defining $a \equiv \alpha_s(\mu)/\pi$, such a perturbative calculation gives

$$\mathrm{Im}\,\pi(Q^2) = \frac{3}{4\pi}\sum_f e_f^2\left(1 + a + a^2\left(1.986 - 0.115n_f - (b_0/4\pi)\ln\frac{Q^2}{\mu^2}\right)\right) \quad [19]$$

with b_0 as above. Now, choosing $\mu = Q$, we see that asymptotic freedom implies that when Q is large, the total cross section is given by the lowest order,

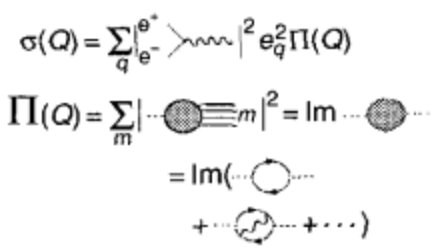


Figure 1 First line: schematic relation of lowest order e^+e^- annihilation to sum over quarks q, each with electric charge e_q. Second line: perturbative unitarity for the current correlation function $\pi(Q)$.

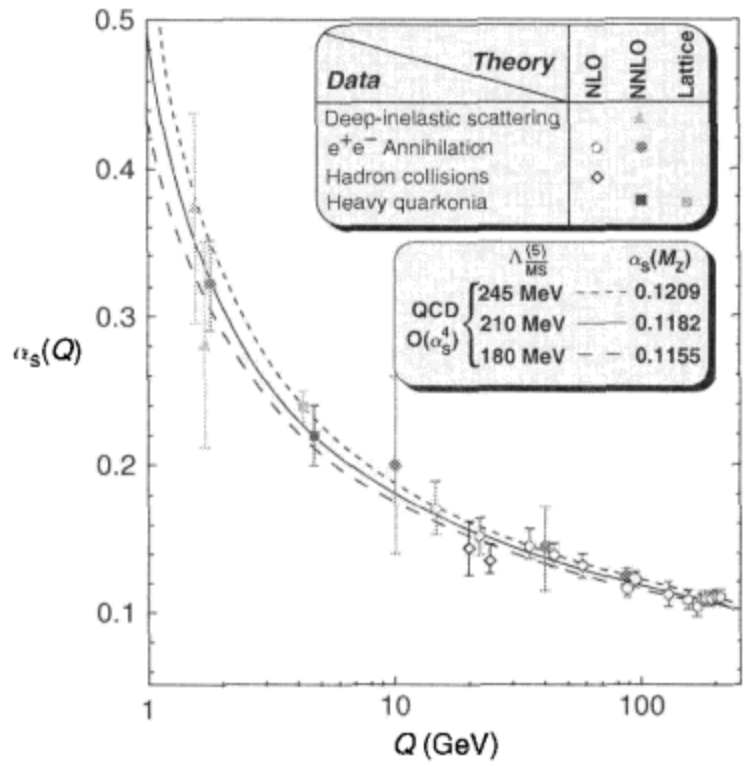


Figure 2 Experimental variation of the strong coupling with scales. Reproduced from Bethke S (2004) Alpha(s) at Zinnowitz. *Nuclear Physics Proceedings Supplements* 135: 345–352, with permission from Elsevier.

plus small and calculable QCD corrections, a result that is borne out in experiment. Comparing experiment to an expression like [19], one can measure the value of $\alpha_s(Q)$, and hence, with eqn [12], $\alpha_s(\mu)$ for any $\mu \gg \Lambda_{QCD}$. **Figure 2** shows a recent compilation of values of α_s from this kind of analysis in different experiments at different scales, clearly demonstrating asymptotic freedom.

Factorization, Scaling, and Parton distributions

One step beyond vacuum matrix elements of currents are their expectation values in single-particle states, and here we make contact with the discovery of QCD, through scaling. Such expectations are relevant to the class of experiments known as deep-inelastic scattering, in which a high-energy electron exchanges a photon with a nucleon target. All QCD information is contained in the tensor matrix element

$$W_N^{\mu\nu}(p,q) \equiv \frac{1}{8\pi}\sum_{\sigma}\int d^4x\, e^{-iq\cdot x}\langle p,\sigma|J^{\mu}(0)J^{\nu}(x)|p,\sigma\rangle \quad [20]$$

with q the momentum transfer carried by the photon, and p,σ the momentum and spin of the target nucleon, N. This matrix element is not infrared safe, since it depends in principle on the entire history of the nucleon state. Thus, it is not accessible to direct perturbative calculation.

Nevertheless, when the scattering involves a large momentum transfer compared to Λ_{QCD}, we may

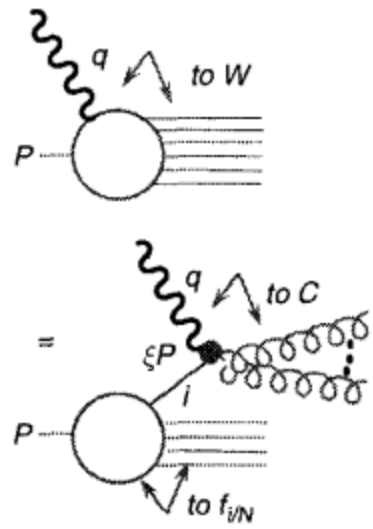


Figure 3 Schematic depiction of factorization in deep-inelastic scattering.

expect a quantum-mechanical incoherence between the scattering reaction, which occurs (by the uncertainty principle) at short distances, and the forces that stabilize the nucleon. After all, we have seen that the latter, strong forces, should be associated with long distances. Such a separation of dynamics, called factorization, can be implemented in perturbation theory, and is assumed to be a property of full QCD. Factorization is illustrated schematically in **Figure 3**. Of course, short and long distances are relative concepts, and the separation requires the introduction of a so-called factorization scale, μ_F, not dissimilar to the renormalization scale described above. For many purposes, it is convenient to choose the two equal, although this is not required.

The expression of factorization for deep-inelastic scattering is

$$W_N^{\mu\nu}(p,q) = \sum_{i=q_f,\bar{q}_f,G}\int_x^1 d\xi C_i^{\mu\nu}(\xi p,q,\mu_F,\alpha_s(\mu_F)) \times f_{i/N}(\xi,\mu_F) \quad [21]$$

where the functions $C_i^{\mu\nu}$ (the coefficient functions) can be computed as an expansion in $\alpha_s(\mu_F)$, and describe the scattering of the "partons," quarks, and gluons, of which the target is made. The variable ξ ranges from unity down to $x \equiv -q^2/2p\cdot q > 0$, and has the interpretation of the fractional momentum of the proton carried by parton i. (Here $-q^2 = Q^2$ is positive.) The parton distributions $f_{i/N}$ can be defined in terms of matrix elements in the nucleon, in which the currents are replaced by quark (or antiquark or gluon) fields, as

$$f_{q/N}(x,\mu) = \frac{1}{4\pi}\int_{-\infty}^{\infty} d\lambda\, e^{-i\lambda x p^+} \times \langle p,\sigma|\bar{q}(\lambda n)U_n(n\lambda,0)n\cdot\gamma q(0)|p,\sigma\rangle \quad [22]$$

n^μ is a light-like vector, and U_n a phase operator whose path C is in the n-direction. The dependence of the parton distribution on the factorization scale is through the renormalization of the composite operator consisting of the quark fields, separated along the light cone, and the nonabelian phase operator $U_n(n\lambda, 0)$, which renders the matrix element gauge invariant by eqn [9]. By combining the calculations of the C's and data for $W_N^{\mu\nu}$, we can infer the parton distributions, $f_{i/N}$. Important factorizations of a similar sort also apply to some exclusive processes, including amplitudes for elastic pion or nucelon scattering at large momentum transfer.

Equation [21] has a number of extraordinary consequences. First, because the coefficient function is an expansion in α_s, it is natural to choose $\mu_F^2 \sim Q^2 \sim p \cdot q$ (when x is of order unity). When Q is large, we may approximate $C_i^{\mu\nu}$ by its lowest order, which is first order in the electromagnetic coupling of quarks to photons, and zeroth order in α_s. In this approximation, dependence on Q is entirely in the parton distributions. But such dependence is of necessity weak (again for x not so small as to produce another scale), because the μ_F dependence of $f_{i/N}(\xi, \mu_F)$ must be compensated by the μ_F dependence of $C_i^{\mu\nu}$, which is order α_s. This means that the overall Q dependence of the tensor $W_N^{\mu\nu}$ is weak for Q large when x is moderate. This is the scaling phenomenon that played such an important role in the discovery of QCD.

Evolution: Beyond Scaling

Another consequence of the factorization [21], or equivalently of the operator definition [22], is that the μ_F-dependence of the coefficient functions and the parton distributions are linked. As in the lepton annihilation cross section, this may be thought of as due to the independence of the physically observable tensor $W_N^{\mu\nu}$ from the choice of factorization and renormalization scales. This implies that the μ_F-dependence of $f_{i/N}$ may be calculated perturbatively since it must cancel the corresponding dependence in C_i. The resulting relation is coventionally expressed in terms of the "evolution equations,"

$$\mu \frac{\mathrm{d}f_{a/N}(x,\mu)}{\mathrm{d}\mu} = \sum_c \int_x^1 \mathrm{d}\xi \; P_{ac}(x/\xi, \alpha_s(\mu)) f_{c/N}(\xi, \mu) \qquad [23]$$

where $P_{ac}(\xi)$ are calculable as power series, now known up to α_s^3. This relation expands the applicability of QCD from scales where parton distributions can be inferred directly from experiment, to arbitrarily high scales, reachable in accelerators under construction or in the imagination, or even on the cosmic level.

At very high energy, however, the effective values of the variable x can become very small and introduce new scales, so that eventually the evolution of eqn [23] fails. The study of nuclear collisions may provide a new high-density regime for QCD, which blurs the distinction between perturbative and nonperturbative dynamics.

Inclusive Production

Once we have evolution at our disposal, we can take yet another step, and replace electroweak currents with any operator from any extension of QCD, in the standard model or beyond, that couples quarks and gluons to the particles of as-yet unseen fields. Factorization can be extended to these situations as well, providing predictions for the production of new particles, F of mass M, in the form of factorized inclusive cross sections,

$$\sigma_{AB\to F(M)}(M, p_A, p_B) = \sum_{i,j=q_f,\bar{q}_f,G} \int \mathrm{d}\xi_a \, \mathrm{d}\xi_b f_{i/A}(\xi_a, \mu) f_{j/B}(\xi_b, \mu) \times H_{ij\to F(M)}(x_a p_A, x_b p_B, M, \mu, \alpha_s(\mu)) \qquad [24]$$

where the functions $H_{ij\to F}$ may be calculated perturbatively, while the $f_{i/A}$ and $f_{j/B}$ parton distributions are known from a combination of lower-energy observation and evolution. In this context, they are said to be "universal," in that they are the same functions in hadron–hadron collisions as in the electron–hadron collisions of deep-inelastic scattering. In general, the calculation of hard-scattering functions H_{ij} is quite nontrivial beyond lowest order in α_s. The exploration of methods to compute higher orders, currently as far as α_s^2, has required extraordinary insight into the properties of multidimensional integrals.

The factorization method helped predict the observation of the W and Z bosons of electroweak theory, and the discovery of the top quark. The extension of factorization from deep-inelastic scattering to hadron production is nontrivial; indeed, it only holds in the limit that the velocities, β_i, of the colliding particles approach the speed of light in the center-of-momentum frame of the produced particle. Corrections to the relation [24] are then at the level of powers of $\beta_i - 1$, which translates into inverse powers of the invariant mass(es) of the produced particle(s) M. Factorizations of this sort do not apply to low-velocity collisions. Arguments for this

result rely on relativistic causality and the uncertainty principle. The creation of the new state happens over timescales of order $1/M$. Before that well-defined event, the colliding particles are approaching at nearly the speed of light, and hence cannot affect the distributions of each others' partons. After the new particle is created, the fragments of the hadrons recede from each other, and the subsequent time development, when summed over all possible final states that include the heavy particle, is finite in perturbation theory as a direct result of the unitarity of QCD.

Structure of Hadronic Final States

A wide range of semi-inclusive cross sections are defined by measuring properties of final states that depend only on the flow of energy, and which bring QCD perturbation theory to the threshold of nonperturbative dynamics. Schematically, for a state $N = |k_1 \dots k_N\rangle$, we define $\mathcal{S}(N) = \sum_i s(\Omega_i) k_i^0$, where $s(\Omega)$ is some smooth function of directions. We generalize the e^+e^- annihilation case above, and define a cross section in terms of a related, but highly nonlocal, matrix element,

$$\frac{d\sigma(Q)}{d\mathcal{S}} \equiv \sigma_0 \int d^4x\, e^{-iQ\cdot x} \left\langle 0 \middle| J^\mu(0) \times \delta\left(\int d^2\Omega s(\Omega)\mathcal{E}(\Omega) - \mathcal{S} \right) J_\mu(x) \middle| 0 \right\rangle \quad [25]$$

where σ_0 is a zeroth-order cross section, and where $\mathcal{E}$ is an operator at spatial infinity, which measures the energy flow of any state in direction Ω: $\mathcal{E}(\Omega)$ $|k_1 \dots k_N\rangle = (1/Q) \sum_i k_i^0 \delta^2(\Omega - \Omega_i)$. This may seem a little complicated, but like the total annihilation cross section, the only dimensional scale on which it depends is Q. The operator $\mathcal{E}$ can be defined in a gauge-invariant manner, through the energy–momentum tensor for example, and has a meaning independent of partonic final states. At the same time, this sort of cross section may be implemented easily in perturbation theory, and like the total annihilation cross section, it is infrared safe. To see why, notice that when a massless ($k^2 = 0$) particle decays into two particles of momenta xk and $(1-x)k$ ($0 \le x \le 1$), the quantity $\mathcal{S}$ is unchanged, since the sum of the new energies is the same as the old. This makes the observable $\mathcal{S}(N)$ insensitive to processes at low momentum transfer.

For the case of leptonic annihilation, the lowest-order perturbative contribution to energy flow requires no powers of α_s, and consists of an oppositely moving quark and antiquark pair. Any measure of energy flow that includes these configurations will dominate over correlations that require α_s corrections. As a result, QCD predicts that in most leptonic annihilation events, energy will flow in two back-to-back collimated sets of particles, known as "jets." In this way, quarks and gluons are observed clearly, albeit indirectly.

With varying choices of $\mathcal{S}$, many properties of jets, such as their distributions in invariant mass, and the probabilities and angular distributions of multijet events, and even the energy dependence of their particle multiplicities, can be computed in QCD. This is in part because hadronization is dominated by the production of light quarks, whose production from the vacuum requires very little momentum transfer. Paradoxically, the very lightness of quarks is a boon to the use of perturbative methods. All these considerations can be extended to hadronic scattering, and jet and other semi-inclusive properties of final states also computed and compared to experiment.

Conclusions

QCD is an extremely broad field, and this article has hardly scratched the surface. The relation of QCD-like theories to supersymmetric and string theories, and implications of the latter for confinement and the computation of higher-order perturbative amplitudes, have been some of the most exciting developments of recent years. As another example, we note that the reduction of the heavy-quark propagator to a nonabelian phase, noted in our discussion of confinement, is related to additional symmetries of heavy quarks in QCD, with many consequences for the analysis of their bound states. Of the bibliography given below, one may mention the four volumes of Shifman (2001, 2002), which communicate in one place a sense of the sweep of work in QCD.

Our confidence in QCD as the correct description of the strong interactions is based on a wide variety of experimental and observational results. At each stage in the discovery, confirmation, and exploration of QCD, the mathematical analysis of relativistic quantum field theory entered new territory. As is the case for gravity or electromagnetism, this period of exploration is far from complete, and perhaps never will be.

See also: AdS/CFT Correspondence; Aharonov–Bohm Effect; BRST Quantization; Current Algebra; Dirac Operator and Dirac Field; Euclidean Field Theory; Effective Field Theories; Electroweak Theory; Lattice Gauge Theory; Operator Product Expansion in Quantum Field Theory; Perturbation Theory and its Techniques; Perturbative Renormalization Theory and BRST; Quantum Field Theory: A Brief Introduction; Random

Matrix Theory in Physics; Renormalization: General Theory; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Scattering, Asymptotic Completeness and Bound States; Seiberg–Witten Theory; Standard Model of Particle Physics.

Further Reading

Bethke S (2004) α_s at Zinnowitz, 2004. *Nuclear Physics Proceeding Supplements* 135: 345–352.

Brodsky SJ and Lepage P (1989) Exclusive processes in quantum chromodynamics. In: Mueller AH (ed.) *Perturbative Quantum Chromodynamics*. Singapore: World Scientific.

Collins JC, Soper DE, and Sterman G (1989) Factorization. In: Mueller AH (ed.) *Perturbative Quantum Chromodynamics*. Singapore: World Scientific.

Dokshitzer Yu L and Kharzeev DE (2004) Gribov's conception of quantum chromodynamics. *Annual Review of Nuclear and Particle Science* 54: 487–524.

Dokshitzer Yu L, Khoze V, Troian SI, and Mueller AH (1988) QCD coherence in high-energy reactions. *Reviews of Modern Physics* 60: 373.

Eidelman S *et al.* (2004) Review of particle physics. *Physics Letters B* 592: 1–1109.

Ellis RK, Stirling WJ, and Webber BR (1996) QCD and Collider Physics, *Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology*, vol. 8. Cambridge: Cambridge University Press.

Greensite J (2003) The confinement problem in lattice gauge theory. *Progress in Particle and Nuclear Physics* 51: 1.

Mandelstam S (1976) Vortices and quark confinement in nonabelian gauge theories. *Physics Reports* 23: 245–249.

Muta T (1986) *Foundations of Quantum Chromodynamics*. Singapore: World Scientific.

Neubert H (1994) Heavy quark symmetry. *Physics Reports* 245: 259–396.

Polyakov AM (1977) Quark confinement and topology of gauge groups. *Nuclear Physics B* 120: 429–458.

Schafer T and Shuryak EV (1998) Instantons and QCD. *Reviews of Modern Physics* 70: 323–426.

Shifman M (ed.) (2001) *At the Frontier of Particle Physics: Handbook of QCD*, vols. 1–3. River Edge, NJ: World Scientific.

Shifman M (ed.) (2002) *At the Frontier of Particle Physics: Hand book of QCD*, vol. 4. River Edge, NJ: World Scientific.

Sterman G (1993) *An Introduction to Quantum Field Theory*. Cambridge: Cambridge University Press.

't Hooft G (1977) On the phase transition towards permanent quark confinement (1978). *Nuclear Physics B* 138: 1.

't Hooft G (ed.) (2005) *Fifty Years of Yang–Mills Theories*. Hackensack: World Scientific.

Weinberg S (1977) The problem of mass. *Transactions of the New York Academy of Science* 38: 185–201.

Wilson KG (1974) Confinement of quarks. *Physical Review D* 10: 2445–2459.



Quantum Electrodynamics and Its Precision Tests

S Laporta, Università di Parma, Parma, Italy
E Remiddi, Università di Bologna, Bologna, Italy

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Quantum electrodynamics (QED) describes the interaction of the electromagnetic field (EMF) with charged particles. Any physical particle interacts, directly or indirectly, with any other particle (including itself); in the case of the electron, however, at low and medium energy (say, up to a few GeV) the interaction with the EMF is by and far the most important, so that QED describes with great precision the dynamics of the electron, and at the same time the electron provides with the most stringent tests of QED currently available.

In the various sections of this article we will discuss, in the following order, the origin of QED, the structure of the radiative corrections, the application of QED to various bound states problems (the hydrogen-like atoms, the muonium, and positronium) and the anomalous magnetic moments of the leptons (the muon and the electron).

Origin of QED

The origin of QED can ideally be traced back to the very beginning of quantum mechanics, the black-body formula by M Planck (1900), which was soon understood as pointing to a discretization of the energy and momentum associated to the EMF into *quanta of light or photons (Einstein 1905)*.

The quantization of the EMF was first worked out by P Jordan, within the article (1926) by M Born, W Heisenberg, and P Jordan (usually referred to as the *Dreimännerarbeit*) and then in the paper "The quantum theory of emission and absorption of radiation" by PAM Dirac, commonly considered the beginning of the so-called second quantization formalism.

In the subsequent year (1928) Dirac published the famous equation for the relativistic electron, from which it was immediately deduced, on a firmer basis, that the electron has spin 1/2, that its spin gyromagnetic ratio (the ratio between spin and associated magnetic moment in suitable dimensionless units; see below for more details) is twice the value predicted by classical physics (a result expressed as $g_e = 2$) and that the levels of atomic hydrogen with the same principal quantum number n are not fully degenerate, as in the nonrelativistic limit, but do possess the so-called fine structure splitting. In particular, the energy of the $n = 2$ levels splits into two values, one value for $2P_{3/2}$ states with total angular momentum $J = 3/2$ and another value for the states $2S_{1/2}$ and $2P_{1/2}$, which have $J = 1/2$; note that the $2S_{1/2}$ and $2P_{1/2}$ states are still degenerate.

Very soon it was realized that Dirac's equation also requires that each particle must be accompanied by its antiparticle, with exactly the same mass and opposite charge. The antiparticle of the electron, the positron, was indeed discovered by C Anderson

Notes:

(1932), establishing Dirac's equation as one of the cornerstones of theoretical physics.

All the ingredients needed for the evaluation of the perturbative corrections to the QED theory (usually called radiative corrections) were already present at that moment, but radiative corrections were not systematically investigated for several years, due perhaps to the length and difficulty of the calculations and the absence of important disagreements between theoretical predictions and experimental results.

The situation changed in 1947, when two experiments were carried out, measuring the energy difference between the $2^2S_{1/2}$ and $2^2P_{1/2}$ levels of the hydrogen atom and the gyromagnetic ratios of the electron.

Lamb and Retherford (1947), by using the "great wartime advances in microwaves techniques," succeeded in establishing that in the hydrogen atom "the $2^2S_{1/2}$ state is higher than the $2^2P_{1/2}$ by about 1000 Mc/sec.," while (as observed above) according to the Dirac theory the two states are expected to have exactly the same energy. Subsequent refinements of the experiment (Triebwasser *et al.* 1953) gave for the difference (now referred to as Lamb shift) the value 1057.77 ± 0.10 MHz, with a relative error 1×10^{-4}.

The authors of the second 1947 experiment (Kusch and Foley 1947) measured the frequencies associated with the Zeeman splitting of two different states of gallium, finding an inconsistency with the theoretical values of the gyromagnetic ratios of the electron. More exactly, write the magnetic moments μ_L, μ_S associated to the (dimensionless) orbital and spin angular momenta L, S of the electron as

$$\mu_L = -g_L \frac{e\hbar}{2m_e c} L, \quad \mu_S = -g_S \frac{e\hbar}{2m_e c} S \qquad [1]$$

where $(-e)$ is the charge of the electron $(e > 0)$, m_e its mass, c the speed of light and g_L, g_S, respectively, the orbital and spin gyromagnetic ratios; the Dirac theory then predicts $g_L = 1$ and $g_S = 2$, while the results of Kusch and Foley (1947) gave a discrepancy which could be accounted for by taking $g_S = 2.00229 \pm 0.00008$ and $g_L = 1$, or alternatively $g_S = 2$ and $g_L = 0.99886 \pm 0.00004$. In modern notation the first conjecture can be rewritten as

$$g_S = g_e = 2(1 + a_e), \quad a_e = 0.001145 \pm 0.00004 \qquad [2]$$

where a_e is the anomalous magnetic moment (or magnetic anomaly) of the electron.

The need of explaining the two experimental results gave rise to a rapid development of covariant perturbation theory (which replaced the previous noncovariant "old fashioned" perturbation theory) and of the renormalization theory, which liberated the perturbative expansion from the divergences plaguing the older approach, opening the path to the evaluation of radiative corrections and to the great success of precision predictions of QED.

The formalism improved quickly, evolving in the more general quantum field theory (QFT) approach; three of the main contributors were Sin-Itiro Tomonaga, Julian Schwinger, and Richard P Feynman, awarded a few years later (1965) the Nobel price "for their fundamental work in quantum electrodynamics, with deep-ploughing consequences for the physics of elementary particles." QFT was then successfully used for describing the weak interactions in the electroweak model and later on also for the strong interactions theory, dubbed quantum chromodynamics (or QCD, in analogy with the popular QED acronym). For more details and references to original works, the reader is invited to look at any treatise on QED or QFT, such as, for instance, Weinberg (1995).

Initially, the Lamb shift was perhaps more important than the electron magnetic anomaly both for the establishment of renormalization theory and as a test of QED, but in the following years it was supplanted by the latter as a precision test of QED.

In 1947 the "best values" for some fundamental constants were indeed

$$\begin{aligned} c &= (2.99776 \pm 0.00004) \times 10^{10}\ \mathrm{cm\ s^{-1}} \\ R_\infty &= \frac{m_e c^2 \alpha^2}{2hc} = 109737.303 \pm 0.017\ \mathrm{cm^{-1}} \\ 1/\alpha &= 137.030 \pm 0.016 \end{aligned} \qquad [3]$$

where R_∞ is the Rydberg constant for infinite mass, h the Planck constant, and α the fine structure constant (let us observe here in passing that R_∞ was and is still known much better than the separate values of m_e, α, and h entering in its definition); for comparison, the current (2005) values for c and R_∞ are

$$\begin{aligned} c &= 299792458\ \mathrm{m\ s^{-1}} \\ R_\infty &= 109737.31568525(73)\ \mathrm{cm^{-1}} \end{aligned} \qquad [4]$$

where the value of c is exact (it is in fact the definition of the meter), and the relative error in R_∞ is 6.6×10^{-12} (the value of α will be discussed later). The measurement of the Lamb shift, repeated several times, gave results in nice agreement with the original value, and for several years it was providing either a test of QED or a precise value for α. But the Lamb shift is the energy difference between the metastable level $2S_{1/2}$ (whose lifetime

is about 1/7 s) and the $2P_{1/2}$ level, which has a lifetime of about 1.596 ns or a natural linewidth of 99.7 MHz. Such a large linewidth poses a strong intrinsic limitation to the precision attainable in the measure of the Lamb shift, which is just ten times larger; as a matter of fact, that precision could never reach the 1×10^{-6} relative error level, while in the meantime the relative precision in a_e reached the 10^{-9} range, replacing the Lamb shift in the role of the leading quantity in high-precision QED.

The Structure of Radiative Corrections

For obvious space problems we can only superficially sketch here the lines along which the perturbative expansion of QED leading to the evaluation of radiative corrections can be built, considering for simplicity only the photon and the electron. One can start from a QED Lagrangian, formally similar to the classical Lagrangian, involving the electron field and the vector potentials of the electromagnetic (or photon) field. The theory is a gauge theory (its physical content should not change if a gradient is added to the vector potentials); it is further an abelian gauge theory as the EMF does not interact directly with itself.

The QED Lagrangian is separated into a free part and an interaction part. From the free part, one derives the wave functions of the free-particle states and the corresponding time-evolution operators (free Green's functions or propagators; let us just recall here that to obtain a convenient photon propagator one has to break the gauge invariance by adding to the Lagrangian a suitable gauge-breaking term), while the interaction part of the Lagrangian gives the "interaction vertices" of the theory.

Aim of the theory is to build the Green's function for the various processes in the presence of the interaction; from these Green's functions, one then derives all the physical quantities of interest.

With the free propagators and the interaction vertices, one generates the perturbative expansion of the Green's functions. The result, namely the contributions to the perturbative expansion (or radiative corrections), can be depicted in terms of Feynman graphs: they consist of various particle lines joined in the interaction vertices, with external lines corresponding to the initial and final particles and internal lines corresponding to intermediate or virtual particle states. Each graph stands for an integral on the momenta of all the intermediate states, each vertex implying among other things an interaction constant, which is $(-e)$ in the case of electron QED, and a δ-function imposing the conservation of the momenta at that vertex. For each process, the Feynman graphs are naturally classified by the total number of the interaction vertices they contain. In the simplest graphs for a given process (the so-called tree graphs) the δ-functions at the vertices make the integrations trivial; but when the number of vertices increases, closed loops of virtual particle states appear, whose evaluation quickly becomes extremely demanding. In QED, each loop gives an extra factor $(-e)^2$ with respect to the tree graph; it is customary to express it in terms of $(\alpha/\pi) = (e/2\pi)^2$, so that the resulting power of (α/π) corresponds to the number of internal loops. The typical QED prediction for a physical quantity is then expressed as a series of powers of the fine structure constant α (and of its logarithm in bound-state problems). As α is small ($\alpha \simeq 1/137$), and the first coefficients of the expansions are usually of the order of 1, a small number of terms in the expansion is in general sufficient to match the precision of the available experimental data.

But the number of different graphs for a given number of loops grows quickly with the number of the loops; in turn, each graph consists in general of a great number of terms and the loop integrations become prohibitively difficult when the number of loops increases, so that the evaluation of radiative corrections proved to be one of the major computational challenges of theoretical physics. As a matter of fact, it prompted the development of computer programs (Veltman 1999) for processing the huge algebraic expressions usually encountered, and of many sophisticated numerical and analytical techniques for performing the loop integrations.

It should be further mentioned here that Feynman graphs written by naively following the above sketched rules are often mathematically ill-defined, taking the form of nonconvergent integrals on the loop momenta. A regularization procedure is needed to give an unambiguous meaning to all the integrals; currently the most powerful regularization is the continuous dimensional regularization scheme, in which the loop integrations are carried out in d continuous dimensions, with d unspecified; renormalization counter-terms are also evaluated in the same scheme, and the physical quantities are recovered in the $d \to 4$ limit (unrenormalized loop integrals and renormalization counterterms are usually singular as powers of $1/(d-4)$ in the $d \to 4$ limit, but all those divergences cancel out in the physical combinations of interest).

QED describes the main interaction of the charged leptons (e, μ, and τ) which have, however, weak interactions as well. Strictly speaking, pure

QED processes do not exist; it is an essential feature of QFT that any existing particle can contribute to the Feynman graphs for any process, when the approximation is pushed to a sufficiently high degree. In particular the photon, which is the main carrier of the QED interaction, is directly coupled also to the strongly interacting particles (the resulting contributions are referred to as "hadronic vacuum polarization" effects).

The precision tests of QED are then to be necessarily searched for in those phenomena where non-QED contributions are presumably small and which involve quantities already well known independently of QED itself. But such high-precision quantities are not always available, and as QED is known better than the rest of physics, very often it is taken to be correct by assumption, and used as a tool for extracting or measuring some of the non-QED quantities relevant to various physical processes.

In any case, as QED predictions are expressed in terms of the fine structure constant α, a determination of α independent of QED is needed; without it, the most precise predictions of QED would simply become measures of α and not tests of the theory.

Finally, it is to be recalled that, ironically, the problem of the convergence of the expansion in powers of α is still open, even if it is commonly accepted that convergence problems will matter only for precisions and corresponding perturbative orders (say at order $1/\alpha \simeq 137$) absolutely out of reach of present experimental and computational possibilities, involving further extremely high energies, where the other fundamental interactions are expected to be as important as QED, so that it would be meaningless to consider only QED.

In the following we will discuss only the QED predictions for bound states and the anomalous magnetic moments of μ and e.

The Bound States

A very good review of the current status of the theory of hydrogen-like atoms can be found in Eides *et al.* (2001), to which we refer for more details and citation of the original papers. The starting point for studying the bound-state problem in QED is the scattering amplitude of two charged particles, predicted by perturbative QED (pQED) as a (formal) series expansion in powers of α. In the static limit $v \to 0$, where v is the relative velocity of the two particles, some of the pQED terms behave as α/v, so that the naive expansion in α becomes meaningless. Fortunately, it is relatively easy to identify the origin of those terms (which are essentially due to the Coulomb interaction between the two charges) and to devise techniques for their resummation. Among them, one can quote the Bethe–Salpeter equation, formally very elegant and complete but difficult to use in practice. A great progress has been achieved by the NRQED (nonrelativistic QED) approach, which is a nonrelativistic theory designed to reproduce the full QED scattering amplitude in the nonrelativistic limit by the *ad hoc* definition, *a posteriori*, of a suitable effective Hamiltonian. The Hamiltonian is then divided into a part containing the Coulomb interaction, which is treated exactly and which gives rise to the bound states, and all the rest, to be treated perturbatively. The power of the NRQED approach was further boosted by the continuous dimensional regularization technique of Feynman graph integrals.

Traditionally, the results are expressed in terms of the energies of the bound states, but as in practice the precise measurements concern the transition frequencies between various levels, it is customary to express any energy contribution to some level, say ΔE, also in terms of the associated frequency $\nu = (\Delta E)/h$, where h is the Planck constant.

The Hydrogen-Like Atoms

Quite in general, a hydrogen-like atom consists of a single electron bound to a positively charge particle, which is a proton for the hydrogen atom, a deuteron nucleus for deuterium, a Helium nucleus for an He^+ ion, a μ^+ meson for muonium, or a positron for positronium. Even if QED alone is not sufficient to treat the dynamical properties of the nuclei, their strong interactions can be described by introducing suitable form factors and a few phenomenological parameters; weak interactions could be treated perturbatively, but are not yet required at the precision levels achieved so far.

The QED results for the hydrogen-like atoms can be expressed in terms of the mass M of the positive particle and of its charge Ze (of course $Z=1$ for hydrogen). When the electron mass m_e is smaller then M (which is always the case, except the positronium case) one can take as a starting point the QED electron moving in the external field of the positive particle, and treat all the other aspects of the relativistic two-body problem (the so-called recoil effects) perturbatively in m_e/M.

Neglecting the spin of the positive particle, the energy levels of the hydrogen-like atom are identified by the usual principal quantum number n, the orbital angular momentum l (with the convention of writing $S, P; D, \ldots$ instead of $l=0, l=1, l=2, \ldots$) and j, the total angular momentum including the spin of the electron. It turns out that the bound

levels consist of very many contributions of different kinds; dropping quantum number indices for simplicity, the energy levels can be written as an expression of the form

$$E = -\frac{m_e c^2 (Z\alpha)^2}{2}\left(\frac{m_r}{m_e}\right) \times \left[\frac{1}{n^2} + (Z\alpha)^2 f_4 + (Z\alpha)^4 f_6 + \cdots\right] + \Delta E_{\text{rad}} + \Delta E_{\text{rec}} + \Delta E_{\text{nucl}} + \cdots \qquad [5]$$

Let us observe that it is convenient to write explicitly the Z factors even when $Z=1$ for a better bookkeeping of the various corrections. As usual, m_r is the reduced mass of the electron, $m_r = m_e M/(m_e + M)$ the mass of the nucleus being M; the first term in the square bracket, $1/n^2$, the familiar Balmer term, is by and far the dominant one, giving for the $n=1$ level in the $Z=1$ case an energy of about 13.6 eV or a corresponding frequency of 3.3×10^{15} Hz. The other terms in the square bracket, f_4 and f_6, are known coefficients (depending also on the small parameter m_e/M; f_4 is essentially the fine structure).

The term ΔE_{rad}, is the bulk of the radiative QED corrections; it can be written as a multiple expansion on $(Z\alpha)$, α and $L = \ln[1/(Z\alpha)^2]$, which turns out to have the following explicit form:

$$\Delta E_{\text{rad}} = m_e c^2 (Z\alpha)^4 \frac{1}{n^3}\left\{\left(\frac{\alpha}{\pi}\right)\left[A_{41}L + A_{40} + (Z\alpha)A_{50} + (Z\alpha)^2(A_{62}L^2 + A_{61}L + A_{60}) + \cdots\right] + \left(\frac{\alpha}{\pi}\right)^2\left[B_{40} + (Z\alpha)B_{50} + (Z\alpha)^2(B_{63}L^3 + B_{62}L^2 + B_{61}L + B_{60}) + \cdots\right] + \left(\frac{\alpha}{\pi}\right)^3[C_{40} + (Z\alpha)C_{50} + \cdots] + \cdots\right\} \qquad [6]$$

The first index of the coefficients refers to the power of $(Z\alpha)$, the second to the power of L; as a rule, there are three powers of $(Z\alpha)$ due to the normalization of the wave function and one power of $(Z\alpha)$ for each interaction with the nucleus (in the leading term of eqn [5] one must subtract two powers of $(Z\alpha)$ due to the long-range nature of the Coulomb interaction), while the terms in $L = \ln[1/(Z\alpha)^2]$ are related to the infrared divergences of the scattering amplitude, with the binding energy acting as infrared cutoff. The A-coefficients refers to order (α/π) or one-loop virtual correction (we do not distinguish here between one-loop self-mass and vacuum-polarization contribution, as usually done in the literature), the B-coefficients to two loops, etc. The coefficients are pure numbers, entirely determined within QED, even if their actual calculation is an extremely demanding task. One of the first results obtained in 1947 was $A_{41} = (4/3)\delta_{l0}$, contributing to the $2S$ but not to the $2P$ states (quite in general, most corrections are much bigger for $l=0$ states than for higher-angular-momentum states), which is sufficient to give the right order of magnitude of the $(2S_{1/2}-2P_{1/2})$ Lamb shift (about 1000 MHz). The other coefficients are now known, thanks to the strenuous and continued efforts (Eides *et al.* 2001) since then, which is impossible to refer properly here in any detail. The current frontier of the theoretical calculation (around the dots in the previous formula) corresponds to 8–9 total powers of (α/π) and $(Z\alpha)$ or some kHz for the $1S$ state.

The next term in eqn [5], ΔE_{rec} contains contributions of order $m_e c^2 (Z\alpha)^5 (m_e/M)$ or smaller (some care must be done for classifying the contributions of order m_e/M, which can be accounted for by proper use of m_r rather than m_e and genuine m_e/M contributions), and are sufficiently known for practical purposes; the same is true for many other contributions discussed in Eides *et al.* (2001) and skipped in eqn [5]. A troublesome contribution comes however from ΔE_{nucl}; at leading order, one has

$$\Delta E_{\text{nucl}} \simeq \frac{2(Z\alpha)^4 mc^2}{3n^3}\left(\frac{mcR_p}{\hbar}\right)^2 \delta_{l0}$$

where R_p is the so-called root-mean-square charge radius of the proton, which is not well known experimentally (in the literature, there are indeed two direct measurements, $R_p = 0.805(11)$ fm and $R_p = 0.862(12)$ fm, in poor agreement with each other; a new independent measurement is strongly needed).

The hyperfine splitting The effect of the interaction of the electron with the spin of the positive particle introduces the so-called hyperfine splitting of all the levels. The order of magnitude of the hyperfine splitting of the $1S$ state is given by the Fermi energy

$$E_F = \frac{4}{3} m_e c^2 (Z\alpha)^4 g_p \frac{m_e}{m_p}$$

where $g_p \simeq 5.586$ is the g-factor of the proton, which gives $\simeq 1.42$ GHz. It was dubbed hyperfine because it is smaller than the fine structure terms by the factor m_e/m_p. Many classes of corrections can be worked out, with patterns similar to those of the previous subsection, and also in this case the nuclear contributions (this time mainly due to the theoretically unknown magnetic form factor and the

so-called polarizability of the proton) prevent from obtaining predictions with án error less than 1 kHz (or a relative precision better than 1×10^{-6}).

The comparison with the experiments Experimen- Experimentally, one measures transition frequencies among the various levels. For many years the precision record was given by the hyperfine splitting of the ground states of hydrogen $\nu_{\rm hfs}(1S)$ was measured long ago (see Hellwig *et al.* (1970) and Essen *et al.* (1971)),

$$\nu_{\rm hfs}(1S) = 1\,420\,405.751\,766\,7(9)\,\text{kHz} \qquad [7]$$

with a relative error 6×10^{-13}. The current record in the optical range is the value of the $(1S–2S)$ hydrogen transition frequency, obtained by means of two-photon Doppler-free spectroscopy Niering *et al.* (2000),

$$\nu(1S–2S) = 2\,466\,061\,413\,187.103(46)\,\text{kHz} \qquad [8]$$

with a relative precision 1.9×10^{-14}; other optical transitions, such as $(2S–8D)$, $(2S–12D)$ are measured with precision of about 1×10^{-11}.

The measurement of the Lamb shift was repeated several times, with results in nice agreement with the original value, such as Lundeen and Pipkin (1986), 1057.845(9) MHz. The most precise value, 1057.8514 ± 0.0019 MHz was given in Palchikov *et al.* (1985) (the result depends, however, on the theoretical value of the lifetime, and should be changed into 1057.8576 ± 0.0021 according to subsequent analysis (see Karshenboim (1996)). The experimental $(2S_{1/2}–2P_{1/2})$ Lamb shift was also obtained as the difference between the measured fine structure separation $(2P_{3/2}–2S_{1/2})$ and the theoretical value of the $(2P_{3/2}–2P_{1/2})$ frequency, and the radiative corrections $\Delta E_{\rm rad}$ to any level are now referred to as the Lamb shift of that level.

As a somewhat deceiving conclusion, the wonderful experimental results of eqns [7] and [8] cannot be used as a high-precision test of the theory or to obtain precise values of many fundamental constants, as the theoretical calculations depend, unfortunately, on hadronic quantities which are not known accurately. Combining theoretical predictions, the above transitions and Lamb shift data, and the available values of α and m_e/m_p, one can indeed obtain a measure of R_p ($R_p = 0.883 \pm 0.014$, according to Melnikov and van Ritbergen (2000)) and the value of R_∞ already quoted above.

Muonium

The muonium is the bound state of a positive μ^+ meson and an electron. At variance with the proton, the μ^+ lepton has no strong interactions, the μ^+e^- system can be studied theoretically within pure QED, with the weak interactions giving a known and small perturbation. Further, the ratio of the masses $m_e/m_\mu \simeq 4.8 \times 10^{-3}$ is small, so that the external field approximation holds. However, the μ is unstable (lifetime $\simeq 2.2\,\mu$s), which makes experiments more difficult to carry out. The best measured quantity is the hyperfine splitting of the $1S$ ground state (see Liu *et al.* (1999))

$$\nu_{\rm hfs}(\mu e, 1S) = 4\,463\,302\,765(53)\,\text{Hz}$$

with a relative precision of 12×10^{-9}. The theoretical treatment is similar to the case of hydrogen, with the important advantage that nuclear interactions are absent and everything can be evaluated within QED, so that the bulk of the contribution is given by a formula with the structure of eqn [6]. But the prediction depends, in any case, on the m_e/m_μ mass, which is not known with the required precision. Indeed, a recent theoretical calculation (Czarnecki *et al.* 2002) (which includes also a contribution of 0.233(3) kHz from hadronic vacuum polarization) gives 4 463 302 680(510)(30)(220) Hz, where the first (and biggest) error comes from m_e/m_μ, the second from α, and the third is the theoretical error (an estimation of higher-order contributions not yet evaluated).

Positronium

The positronium is the bound state of an electron and a positron. Theoretically, it is an ideal system to study, as it can be described entirely within QED, without any unknown parameter of non-QED origin. As the masses of the two constituents, positron and electron, are strictly equal, the reduced mass of the system is exactly equal to half of the electron mass, $m_r = m_e/2$, and the energy scale of the bound states is half of R_∞.

At variance with the muonium case, the external field approximation is not valid, so that positronium must be treated with the full two-body bound-state machinery of QFT, of which it provides an excellent test (Karshenboim 2004).

Experimentally, radioactive positron sources are available, so that positronium is easier to produce than muonium. It is, however, unstable; states with total spin S equal 0 (also called parapositronium states) annihilate into an even number (mainly two) of gammas, and states with $S = 1$ (orthopositronium) into an odd number (mainly three) of gammas, with short lifetimes (which make precise measurements difficult). Further, as positronium is the lightest

atom, Doppler-broadening effects are very important, reducing the precision of spectroscopical measurements.

Positronium decay rates There has been a long-time discrepancy between theory and experiment in decay rate of ground-state orthopositronium, which prompted thorough theoretical investigations looking for errors in the calculations or flaws in the formalism, but it turned out that the flaw was on the experimental side. The current theoretical prediction for the ground state $S=1$ decay is (Adkins *et al.* 2002)

$$\Gamma(1S,\text{ortho}) = \Gamma_0\left\{1 + A\left(\frac{\alpha}{\pi}\right) + \frac{1}{3}\alpha^2 \ln\alpha + B\left(\frac{\alpha}{\pi}\right)^2 - \frac{3\alpha^3}{2\pi}\ln^2\alpha + C\frac{\alpha^3}{\pi}\ln\alpha + \cdots\right\} = 7.039979(11)\,\mu s^{-1}$$

where $\Gamma_0 = 2(\pi^2 - 9)m_e\alpha^6/(9\pi) = 7.2111670(1)$, $A = -10.286606(10), B = 45.06(26), C = -5.517$, in nice agreement with the less precise experimental result of Karshenboim (2004, ref. 38) 7.0404 (10)(8) μs^{-1}. As a curiosity, the coefficients A, B above are among the greatest coefficients so far appeared in QED radiative corrections.

The agreement between theory and experiment for the ground-state parapositronium decay rate has always been good; the current status of Karshenboim (2004, ref. 41) is 7990.9(1.7) μs^{-1} for the experimental result and of Karshenboim (2004, ref. 43) 7989.64(2) μs^{-1} for the theoretical prediction.

Positronium levels The quantum number structure of the levels is similar to muonium, with the important difference, however, that the hyperfine splitting (which in hydrogen or muonium is small because it is proportional to the ratio of the masses of the two components) is in fact of the same order as the fine structure. The theoretical evaluation of the energy levels provides a very stringent check of QED and of the overall treatment of the bound-state problem. Corrections have been evaluated, typically, up to order $mc^2\alpha^7$. The best-known quantities are the ground state (hyper)fine splitting, experimental value (Ritter *et al.* 1984) 203.38910(74) GHz (3.6×10^{-6} relative error), theoretical (Karshenboim 2004) 203.3917(6), and the $1S$–$2S$ transition for orthopositronium, experiment (Fee *et al.* 1993) 1 233 607 216.4 (3.2) MHz, theory 1 233 607 222.2(6). The general agreement is good; the precisions achieved are, however, not yet sufficient to allow a determination of R_∞ or α competitive with other measurements.

The Anomalous Magnetic Moments of Leptons

The precision of the measurements requires, for both the e and μ leptons, to also take into account graphs with contributions from the other leptons as virtual intermediate states and those of hadronic and weak origin. Quite in general, if the mass of the virtual particle, say m_v, is smaller than the mass of the external lepton, say m_l, one can have an $\ln(m_l/m_v)$ behavior of the contributions; that is the case of the virtual electron contributions to the muon magnetic anomaly a_μ, which can be enhanced by powers of $\ln(m_\mu/m_e)$. In the opposite case, $m_v > m_l$, the contribution has the behavior $(m_l/m_v)^2$; that is the case of the $(m_\mu/m_\tau)^2$ contributions to a_μ from τ loops and of the $(m_e/m_\mu)^2$ contributions from μ loops to th electron magnetic anomaly, a_e. As strong and weak interactions are in general associated with heavy-mass particles, they are expected to be more important for a_μ than a_e; further, a given heavy particle contribution to a_e is smaller by a factor $(m_e/m_\mu)^2$ than the corresponding contribution to a_μ.

The Magnetic Anomaly a_μ of the μ

The a_μ has been reviewed in Passera (2005). The present (2005) world average experimental value is

$$a_\mu(\text{exp}) = 116\,592\,080(60) \times 10^{-11}$$

with a relative error 0.5×10^{-6}.

Theoretically, one can write

$$a_\mu = a_\mu(\text{QED}) + a_\mu(\text{had}) + a_\mu(\text{EW}) \qquad [9]$$

where the three terms stand for the contributions from pure QED, strong interacting hadrons and electroweak interactions. In turn, one can expand $a_\mu(\text{QED})$ in powers of α as

$$a_\mu(\text{QED}) = \sum_l C_l\left(\frac{\alpha}{\pi}\right)^l = \sum_l\left[A_1^{(l)} + A_2^{(l)}\left(\frac{m_\mu}{m_e}\right) + A_2^{(l)}\left(\frac{m_\mu}{m_\tau}\right) + A_3^{(l)}\left(\frac{m_\mu}{m_e}, \frac{m_\mu}{m_\tau}\right)\right]\left(\frac{\alpha}{\pi}\right)^l \qquad [10]$$

The coefficients $A_1^{(l)}$ involve only the photon and the external lepton as virtual states, are identically the same as in a_e; they are known up to $l=4$ included (but, strictly speaking, the contribution of $A_1^{(4)}$ is smaller than the experimental error of a_μ) and will be discussed later for the electron. The $A_2^{(l)}(m_\mu/m_e)$ are very large, being enhanced by powers of $\ln(m_\mu/m_e)$, and are required and known up to $l=5$; $A_2^{(l)}(m_\mu/m_\tau)$ starts with $A_2^{(2)}(m_\mu/m_\tau) \simeq$

$1/45(m_\mu/m_\tau)^2$, contributing 4.2×10^{-11} to a_μ, so that the $A_2^{(l)}(m_\mu/m_\tau)$ with higher values of l are not needed. $A_3^{(l)}(m_\mu/m_e,\ m_\mu/m_\tau)$, finally, starts from $l=3$, and gives a negligible contribution 0.7×10^{-11}. Summing up, one finds $C_1=1/2$, $C_2=0.765\,857\,410(27)$ (the error is from the experimental errors in the lepton masses) $C_3=24.050\,509\,64(43)$, $C_4=131.011(8)$, and $C_5=677(40)$. As already observed, the coefficients are large due to the presence of $\ln(m_\mu/m_e)$ factors. The last term C_5 contributes $4.6(0.3)\times10^{-11}$ to a_μ, and the total QED contribution is

$$a_\mu(\text{QED})=116\,584\,718.8(0.3)(0.4)\times10^{-11}$$

where the first error is due to the uncertainties in the coefficients C_2, C_3, and C_5 and the second from the value of α coming from atom interferometry measurements (see below).

The hadronic contributions are of two kinds, those due to vacuum polarization, a_μ(vac.pol), which can be evaluated by sound theoretical methods by using existing experimental data, and those due to light-by-light hadronic scattering, a_μ(lbl), whose evaluation relies on much less firmer grounds and are entirely model-dependent. The value of a_μ(vac.pol) varies slightly among the various authors (see Passera (2005) for reference to original work), let us take as a typical value $a_\mu(\text{vac.pol})=6834(92)\times10^{-11}$ (based on e^+e^- scattering data and including also first-order radiative corrections). The model-dependent value of the light-by-light contribution changed several times in the years (also in sign!) but now there is a general consensus that it should be positive; let us take, somewhat arbitrarily, $a_\mu(\text{lbl})=136(25)\times10^{-11}$, so that the total hadronic contribution becomes

$$a_\mu(\text{had})=6970(92)\times10^{-11}$$

The electroweak contribution, finally, is

$$a_\mu(\text{EW})=154(2)\times10^{-11}$$

which accounts for a one-loop purely weak contribution and a two-loop electromagnetic and weak contribution, which turns out to be very large (-42×10^{-11}) for the presence of logarithms in the masses (the error is due to the uncertainty in the Higgs boson mass).

Summing up, eqn [9] gives $a_\mu=116\,591\,842(92)\times10^{-11}$, so that

$$a_\mu(\text{exp})-a_\mu=138(60)(90)\times10^{-11}$$

The substantial agreement can be considered to be a good overall check of QED and electroweak interactions. But another attitude is often adopted in the scientific community: the validity of QED and electroweak models is taken for granted, and a disagreement, if any, is considered to be an indication of new physics. To obtain significant information in that direction, however, the experimental and the theoretical errors (dominated in turn by the experimental error in e^+e^- scattering data) should be significantly reduced.

The Magnetic Anomaly a_e of the Electron

Experimentally, one has the 1987 value (Kinoshita 2005, ref. 1).

$$a_e(\text{exp})=1\,159\,652\,188.4(4.3)\times10^{-12} \qquad [11]$$

with a relative error 3.7×10^{-9} and the preliminary Harvard (2004) measurement (Kinoshita 2005, ref. 3).

$$a_e(\text{Harvard})=1\,159\,652\,180.86(0.57)\times10^{-12} \qquad [12]$$

with 0.5×10^{-9} relative error, that is, an increase in precision by a factor 7.

Theoretically, eqns [9] and [10] apply also to the electron; given the smallness of the electron mass, the relevant terms up to the precision of the experimental data are

$$\begin{aligned}a_e=&A_1^{(1)}\left(\frac{\alpha}{\pi}\right)+A_1^{(2)}\left(\frac{\alpha}{\pi}\right)^2+A_1^{(3)}\left(\frac{\alpha}{\pi}\right)^3\\&+A_1^{(4)}\left(\frac{\alpha}{\pi}\right)^4+\cdots+A_2^{(2)}\left(\frac{m_e}{m_\mu}\right)\left(\frac{\alpha}{\pi}\right)^2\\&+a_e(\text{had})+a_e(\text{EW}) \qquad [13]\end{aligned}$$

The explicit calculation gives

$$A_1^{(1)}=\frac{1}{2}\quad\text{(Passera 2005, ref. 1)}$$

$$\begin{aligned}A_1^{(2)}&=\frac{197}{144}+\frac{1}{12}\pi^2-\frac{1}{2}\pi^2\ln2+\frac{3}{4}\zeta(3)\\&=-0.328\,478\,965\,579\ldots\end{aligned}$$

(Passera 2005, ref. 17)

$$\begin{aligned}A_1^{(3)}=&\frac{83}{72}\pi^2\zeta(3)-\frac{215}{24}\zeta(5)\\&+\frac{100}{3}\left[a_4+\frac{1}{24}\ln^4 2-\frac{1}{24}\pi^2\ln^2 2\right]\\&-\frac{239}{2160}\pi^4+\frac{139}{18}\zeta(3)-\frac{298}{9}\pi^2\ln2\\&+\frac{17101}{810}\pi^2+\frac{28259}{5184}\\=&1.181\,241\,456\ldots\end{aligned}$$

(Laporta and Remiddi 1996)

$$A_1^{(4)}=-1.7283(35)\quad\text{(Kinoshita 2005)}$$

and

$$A_2^{(2)}\left(\frac{m_e}{m_\mu}\right)\left(\frac{\alpha}{\pi}\right)^2 \simeq \frac{1}{45}\left(\frac{m_e}{m_\mu}\right)^2\left(\frac{\alpha}{\pi}\right)^2 \simeq 2.72\times 10^{-12}$$

$$a_e(\text{had}) = 1.67(0.02)\times 10^{-12}$$

$$a_e(\text{EW}) = 0.03\times 10^{-12} \qquad [14]$$

For obtaining a meaningful prediction, one needs now a precise value of α. The most precise value available at present is that of Passera (2005, ref. 49)

$$\alpha^{-1}(\text{aif}) = 137.036\,000\,3(10)$$

with relative error 7×10^{-9}, obtained by the atom interferometry method (which is independent of QED, depending only on the kinematics of the Doppler effect). With that value of α, the theoretical prediction for a_e becomes

$$a_e = 1\,159\,652\,175.9(8.5)(0.1)10^{-12}$$

where the first error comes from α and the second from C_4; conversely, one can use the QED prediction for a_e and a_e(Harvard) for obtaining α; one obtains in that way

$$\alpha^{-1}(\text{QED}, a_e) = 137.035\,999\,708(12)(67)$$

where the first uncertainty is from C_4 and the second from the experiment. We see that theory and experiment are in good agreement.

As a concluding remark, another independent and more precise (or analytic!) evaluation of C_4 contribution would be welcome. The five-loop term is not known; but as $(\alpha/\pi)^5 \approx 0.07\times 10^{-12}$, if C_5 is, say, not greater than 2, its contribution to a_e becomes equal to the contribution of the error ΔC_4 of C_4 and is not yet required to match the current precision of $a_e(\text{exp})$. The ultimate theoretical limit, the error of the hadronic contribution, $\Delta a_e(\text{had}) = 0.02\times 10^{-12}$, is still smaller, corresponding to a change $\Delta C_4 = 0.0007$ of C_4 or $\Delta C_5 = 0.3$ of C_5.

See also: Abelian and Nonabelian Gauge Theories Using Differential Forms; Anomalies; Effective Field Theories; Electroweak Theory; Quantum Field Theory: A Brief Introduction; Standard Model of Particle Physics.

Further Reading

Adkins GS, Fell RN, and Sapirstein J (2002) Two-loop correction to the orthopositronium decay rate. *Annals of Physics (NY)* 295: 136.

Czarnecki A, Eidelman SI, and Karshenboim SG (2002) Muonium hyperfine structure and hadronic effects. *Physical Review D* 65: 053004.

Dirac PAM (1927) The quantum theory of emission and absorption of radiation. *Proceedings of the Royal Society A* 114: 243.

Eides MI, Grotch H, and Shelyuto VA (2001) Theory of light hydrogenlike atoms. *Physics Reports* 342: 63 (arXiv:hep-ph/0002158).

Essen L, Donaldson RW, Bangman MJ, and Hope EG (1971) Frequency of the hydrogen maser. *Nature* 229: 110.

Fee MS *et al.* Measurement of the positronium 1^3s_1–2^3s_1 interval by continuous-wave two-photon excitation. *Physical Review Letters* 70: 1397.

Hellwig H, Vessot RFC, Levine MW, Zitzewitz PW, Allan DW, and Glaze DJ (1970) Measurement of the unperturbed hydrogen hyperfine transition frequency. *IEEE Transactions on Instrument Measurement* IM19: 200.

Karshenboim SG (1996) Leading logarithmic corrections and uncertainty of muonium hyperfine splitting calculations. *Z. Physics D* 36: 11.

Karshenboim SG (2004) Precision study of positronium: testing bound state QED theory. *International Journal of Modern Physics A* 19: 3879 (e-print hep-ph/0310099).

Kinoshita T (2005) Improved α^4 term of the electron anomalous magnetic moment (hep-ph/0507249).

Kusch P and Foley HM (1947) Precision measurements of the ratio of the atomic 'g values' in the $^2P_{3/2}$ and $^2P_{1/2}$ states of gallium. *Physical Review* 72: 1256(L).

Lamb WE Jr. and Retherford RC (1947) Fine structure of the hydrogen atom by a microwave method. *Physical Review* 72: 241.

Laporta S and Remiddi E (1996) The analytical value of the electron $(g-2)$ at order α^3 in QED. *Physics Letters B* 379: 283.

Liu W *et al.* (1999) High precision measurements of the ground state hyperfine structure interval of muonium and of the muon magnetic moment. *Physical Review Letters* 82: 711.

Lundeen SR and Pipkin FM (1986) Separated oscillatory field measurement of Lamb shift in atomic hydrogen, $n=2$. *Metrologia* 22(1): 9.

Melnikov K and van Ritbergen T (2000) Three-loop slope of the Dirac form factor and the 1S Lamb shift in hydrogen. *Physical Review Letters* 84: 1673 (hep-ph/9911277).

Niering M *et al.* (2000) Measurement of the hydrogen 1s–2s transition frequency by phase coherent comparison with a microwave cesium fountain clock. *Physics Review Letters* 84: 5496.

Pal'chikov VG, Sokolov YL, and Yakovlev VP (1985) Measurement of the Lamb shift in H, $n=2$. *Metrologia* 21: 99.

Passera M (2005) The standard model prediction of the muon anomalous magnetic moment. *Journal of Physics G* 31: R75 (hep-ph/0411168).

Ritter MW, Egan PO, Hughes VW, and Woodle KA (1984) Precision determination of the hyperfine-structure interval in the ground state of positronium, V. *Physical Review A* 30: 1331.

Triebwasser S, Dayhoff ES, and Lamb WE Jr. (1953) Fine structure of the hydrogen atom, V. *Physical Review* 89: 98.

Veltman MJC (1999) *Nobel Lecture* (http://nobelprize.org/physics/laureates/1999/veltman-lecture.pdf).

Weinberg S (1995) *The Quantum Theory of Fields.* Cambridge: Cambridge University Press.

Renormalization: General Theory

J C Collins, Penn State University, University Park, PA, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Quantum field theories (QFTs) provide a natural framework for quantum theories that obey the principles of special relativity. Among their most striking features are ultraviolet (UV) divergences, which at first sight invalidate the existence of the theories. The divergences arise from Fourier modes of very high wave number, and hence from the structure of the theories at very short distances. In the very restricted class of theories called "renormalizable," the divergences may be removed by a singular redefinition of the parameters of the theory. This is the process of renormalization that defines a QFT as a nontrivial limit of a theory with a UV cutoff.

A very important QFT is the standard model, an accurate and successful theory for all the known interactions except gravity. Calculations using renormalization and related methods are vital to the theory's success.

The basic idea of renormalization predates QFT. Suppose we treat an observed electron as a combination of a bare electron of mass m_0 and the associated classical electromagnetic field down to a radius a. The observed mass of the electron is its bare mass plus the energy in the field (divided by c^2). The field energy is substantial, for example, 0.7 MeV when $a = 10^{-15}$ m, and it diverges when $a \to 0$. The observed mass, 0.5 MeV, is the sum of the large (or infinite) field contribution compensated by a negative and large (or infinite) bare mass. This calculation needs replacing by a more correct version for short distances, of course, but it remains a good motivation.

In this article, we review the theory of renormalization in its classic form, as applied to weak-coupling perturbation theory, or Feynman graphs. It is this method, rather than the Wilsonian approach (*see* Exact Renormalization Group), that is typically used in practice for perturbative calculations in the standard model, especially its QCD part.

Much of the emphasis is on weak-coupling perturbation theory, where there are well-known algorithmic rules for performing calculations and renormalization. Applications (*see* Quantum Chromodynamics for some important nontrivial examples) involve further related results, such as the operator product expansion, factorization theorems, and the renormalization group (RG), to go far beyond simple fixed-order perturbation theory. The construction of fully rigorous mathematical treatments for the exact theory is a topic of future research.

Formulation of QFT

A QFT is specified by its Lagrangian density. A simple example is ϕ^4 theory:

$$\mathcal{L} \stackrel{?}{=} \frac{(\partial\phi)^2}{2} - \frac{m^2\phi^2}{2} - \frac{\lambda\phi^4}{4!} \qquad [1]$$

where $\phi(x) = \phi(t, \boldsymbol{x})$ is a single component Hermitian field. The Lagrangian density and the resulting equation of motion, $\partial^2\phi + m^2\phi + (1/6)\lambda\phi^3 = 0$, are local; they involve only products of fields at the same spacetime point. Such locality is characteristic of relativistic theories, where otherwise it is difficult or impossible to preserve causality, but it is also the source of the UV divergences. The question mark over the equality symbol in eqn [1] is a reminder that renormalization of UV divergences will force us to modify the equation.

The Feynman rules for perturbation theory are given by a free propagator $\mathrm{i}/(p^2 - m^2 + \mathrm{i}0)$ and an interaction vertex $-\mathrm{i}\lambda$. Although we will usually work in four spacetime dimensions, it is useful also to consider the theory in a general spacetime dimensionality n, where the coupling has energy dimension $[\lambda] = E^{4-n}$. We use "natural units," that is, with $\hbar = c = 1$. The "i0" in the propagator $\mathrm{i}/(p^2 - m^2 + \mathrm{i}0)$ symbolizes the location of the pole relative to the integration contour; it is often written as $\mathrm{i}\epsilon$.

The primary targets of calculations are the vacuum expectation values of time-ordered products of ϕ; in QFT these are called the Green functions of the theory. From these can be reconstructed the scattering matrix, scattering cross sections, and other measurable quantities.

One-Loop Calculations

Low-order graphs for the connected and amputated four-point Green function are shown in **Figure 1**. Each one-loop graph has the form

$$-\mathrm{i}\lambda^2 I(p^2) \stackrel{?}{=} \frac{\lambda^2}{2}\int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{1}{(k^2 - m^2 + \mathrm{i}0)[(p-k)^2 - m^2 + \mathrm{i}0]} \qquad [2]$$

where p is a combination of external momenta. There is a divergence from where the loop

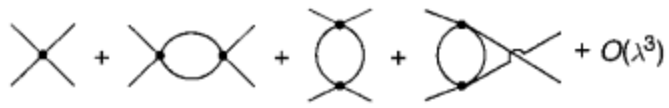


Figure 1 One-loop approximation to connected and amputated four-point function, before renormalization.

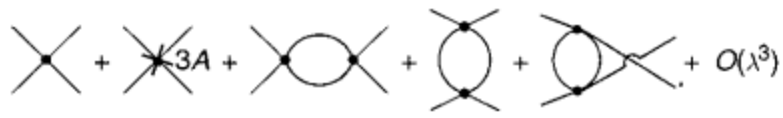


Figure 2 One-loop approximation to renormalized connected and amputated four-point function, with counter-term.

momentum k goes to infinity. We define the degree of divergence, Δ, by counting powers of k at large k, to get $\Delta=0$. In an n-dimensional spacetime we would have $\Delta=n-4$. The integral is divergent whenever $\Delta \geq 0$. Comparing the dimensions of the one-loop and tree graphs shows that Δ equals the negative of the energy dimension of the coupling λ. Thus, the dimensionlessness of λ at the physical spacetime dimension is equivalent to the integral being just divergent.

The infinity in the integral implies that the theory *in its naive formulation* is not defined. With the aid of RG methods, it has been shown that the problem is with the complete theory, not just perturbation theory.

The divergence only arises because we use a continuum spacetime. So suppose that we formulate the theory initially on a lattice of spacing a (in space or spacetime). Our loop graph is now

$$-\mathrm{i}\lambda^2 I(p;m,a) = \frac{-\lambda^2}{32\pi^4}\int \mathrm{d}^4k\, S(k,m;a)\, S(p-k,m;a) \qquad [3]$$

where the free propagator $S(k,m;a)$ approaches the usual value $\mathrm{i}/(k^2-m^2+\mathrm{i}0)$ when k is much smaller than $1/a$, and it falls off more rapidly for large k. The basic observation that propels the renormalization program is that the divergence as $a\to 0$ is independent of p. This is most easily seen by differentiating once with respect to p, after which the integral is convergent when $a=0$, because the differentiated integral has degree of divergence -1.

Thus we can cancel the divergence in eqn [2] by replacing the coupling in the first term in **Figure 1**, by the so-called bare coupling

$$\lambda_0 = \lambda + 3A(a)\lambda^2 + O(\lambda^3) \qquad [4]$$

Here $A(a)$ is chosen so that the renormalized value of our one-loop graph,

$$-\mathrm{i}\lambda^2 I_R(p^2,m^2) = -\mathrm{i}\lambda^2 \lim_{a\to 0}\left[I(p;m,a)+A(a)\right] \qquad [5]$$

exists, at $a=0$, with $A(a)$ in fact being real valued. The factor 3 multiplying $A(a)$ in eqn [4] is because there are three one-loop graphs, with equal divergent parts. The replacement for the coupling is made in the tree graph in **Figure 1**, but not yet at the vertices of the other graphs, because at the moment we are only doing a calculation accurate to order λ^2; the appropriate expansion parameter of the theory is the finite renormalized coupling λ, held fixed as $a\to 0$. We call the extra term in eqn [5] a counter-term. The diagrams for the correct renormalized calculation are represented in **Figure 2**, which has a counter-term graph compared with **Figure 1**.

In the physics terminology, used here, the cutting-off of the divergence by using a modified theory is called a regularization. This contrasts with the mathematics literature, where "regularized integral" usually means the same as a physicist's "renormalized integral."

There is always freedom to add a finite term to a counter-term. When we discuss the RG, we will see that this corresponds to a reorganization of the perturbation expansion and provides a powerful tool for improving perturbatively based calculations, especially in QCD. Contrary to the impression given in some parts of the literature, it is not necessary that a renormalized mass equal a corresponding physical particle mass, with similar statements for coupling and field renormalization. While such a prescription is common and natural in a simple theory like QED, it is by no means required and certainly may not always be best. If nothing else, the correspondence between fields and stable particles may be poor or nonexistent (as in QCD).

One classic possibility is to subtract the value of the graph at $p=0$, a prescription associated with Bogoliubov, Parasiuk, and Hepp (BPH), which leads to

$$-\mathrm{i}\lambda^2 I_{\mathrm{R,BPH}}(p^2) = \frac{-\mathrm{i}\lambda^2}{32\pi^2}\int_0^1 \mathrm{d}x \ln\left[1-p^2x(1-x)/m^2\right] \qquad [6]$$

In obtaining this from [2], we used a standard Feynman parameter formula,

$$\frac{1}{AB} = \int_0^1 \mathrm{d}x \frac{1}{[Ax+B(1-x)]^2} \qquad [7]$$

to combine the propagator denominators, after which the integral over the momentum variable k is elementary. We then obtain the renormalized one-loop (four-point and amputated) Green function

$$-\mathrm{i}\lambda - \mathrm{i}\lambda^2[I_R(s)+I_R(t)+I_R(u)] + O(\lambda^3) \qquad [8]$$

where s, t, and u are the three standard Mandelstam invariants for the Green function. (For a $2\to 2$

scattering process, or a corresponding off-shell Green function, in which particles of momenta p_1 and p_2 scatter to particles of momenta p_1' and p_2', the Mandelstam variables are defined as $s=(p_1+p_2)^2$, $t=(p_1-p_1')^2$, and $u=(p_1-p_2')^2$.)

In the general case, with a nonzero degree of divergence, the divergent part of an integral is a polynomial in p and m of degree D, where D is the smallest positive integer less than or equal to Δ. In a higher spacetime dimension, this implies that renormalization of the original, momentum-independent, interaction vertex is not sufficient to cancel the divergences. We would need higher derivative terms, and this is evidence that the theory is not renormalizable in higher than 4 spacetime dimensions. Even so, the terms needed would be local, because of the polynomiality in p.

Complete Formulation of Renormalization Program

The full renormalization program motivated by example calculations is:

- the theory is regulated to cut off the divergences;
- the numerical value of each coefficient in $\mathcal{L}$ is allowed to depend on the regulator parameter (e.g., a); and
- these dependences are adjusted so that finite results for Green functions are obtained after removal of the regulator.

In ϕ^4 theory, we therefore replace $\mathcal{L}$ by

$$\mathcal{L}=\frac{Z}{2}(\partial\phi)^2-\frac{Zm_0^2}{2}\phi^2-\frac{Z^2\lambda_0}{4!}\phi^4 \qquad [9]$$

with the bare parameters, Z, m_0 and λ_0, having a regulator dependence such that Green functions of ϕ are finite at $a=0$.

The slightly odd labeling of the coefficients in eqn [9] arises because observables like cross sections are invariant under a redefinition of the field by a factor. In terms of the bare field $\phi_0 \stackrel{\text{def}}{=} \sqrt{Z}\phi$, we have

$$\mathcal{L}=\frac{1}{2}(\partial\phi_0)^2-\frac{m_0^2}{2}\phi_0^2-\frac{\lambda_0}{4!}\phi_0^4 \qquad [10]$$

The unit coefficient of $(1/2)(\partial\phi_0)^2$ implies that ϕ_0 has canonical commutation relations (in the regulated theory). This provides a natural standard for the normalization of the bare mass m_0 and the bare coupling λ_0.

All terms in $\mathcal{L}$ have coefficients with dimension zero or larger. This is commonly characterized by saying that the terms $\mathcal{L}$ "have dimension 4 or less," which refers to the products of field operators and derivatives in each term. A generalization of the power-counting analysis shows that if we start with a theory whose $\mathcal{L}$ only has terms of dimension 4 or less, then no terms of higher dimension are needed as counter-terms, at least not in perturbation theory. This is a very powerful restriction on self-contained QFTs, and was critical in the discovery of the standard model.

Sometimes it is found that the description of some piece of physics appears to need higher-dimension operators, as was the case originally with weak-interaction physics. The lack of renormalizability of such theories indicates that they cannot be complete, and an upper bound on the scale of their applicability can be computed, for example, a few hundred GeV for the four-fermion theory of weak interactions. Eventually, this theory was superseded by the renormalizable Weinberg–Salam theory of weak interactions, now a part of the standard model, to which the four-fermion theory provides a low-energy approximation for charged current weak interactions.

Certain operators of allowed dimensions are missing in eqn [9]: the unit operator, and ϕ and ϕ^3. Symmetry under the transformation $\phi\rightarrow-\phi$ implies that Green functions with an odd number of fields vanish, so that no ϕ and ϕ^3 counter-terms are needed. Divergences with the unit operator do appear, but not for ordinary Green functions. In gravitational physics, the coefficient of the unit operator gives renormalization of the cosmological constant.

To implement renormalized perturbation theory, we partition $\mathcal{L}$ (nonuniquely) as

$$\mathcal{L}=\mathcal{L}_{\text{free}}+\mathcal{L}_{\text{basic interaction}}+\mathcal{L}_{\text{counter-term}} \qquad [11]$$

where the free, the basic interaction, and the counter-term Lagrangians are

$$\mathcal{L}_{\text{free}}=\frac{1}{2}(\partial\phi)^2-\frac{m^2}{2}\phi^2 \qquad [12]$$

$$\mathcal{L}_{\text{basic interaction}}=-\frac{\lambda}{4!}\phi^4 \qquad [13]$$

$$\mathcal{L}_{\text{counter-term}}=\frac{Z-1}{2}(\partial\phi)^2-\frac{(Zm_0^2-m^2)}{2}\phi^2-\frac{(Z^2\lambda_0-\lambda)}{4!}\phi^4 \qquad [14]$$

The renormalized coupling and mass, λ and m, are to be fixed and finite when the UV regulator is removed. Both the basic interaction and the counter-terms are treated as interactions. First we compute "basic graphs" for Green functions using only the basic

interaction. The counter-terms are expanded in powers of λ, and then all graphs involving counter-term vertices at the chosen order in λ are added to the calculation. The counter-terms are arranged to cancel all the divergences, so that the UV regulator can be removed, with m and λ held fixed. The counter-terms cancel the parts of the basic Feynman graphs associated with large loop momenta. An algorithmic specification of the otherwise arbitrary finite parts of the counter-terms is called a renormalization prescription or a renormalization scheme. Thus, it gives a definite relation between the renormalized and bare parameters, and hence a definite specification of the partitioning of $\mathcal{L}$ into its three parts.

It has been proved that this procedure works to all orders in λ, with corresponding results for other theories. Even in the absence of fully rigorous nonperturbative proofs, it appears clear that the results extend beyond perturbation theory, at least in asymptotically free theories like QCD: see the discussion on Wilsonian RG (*see* Exact Renormalization Group).

Dimensional Regularization and Minimal Subtraction

The final result for renormalized graphs does not depend on the particular regularization procedure. A particularly convenient procedure, especially in QCD, is dimensional regularization, where divergences are removed by going to a low spacetime dimension n. To make a useful regularization method, n is treated as a *continuous* variable, $n = 4 - 2\epsilon$.

Great advantages of the method are that it preserves Poincaré invariance and many other symmetries (including the gauge symmetry of QCD), and that Feynman graph calculations are minimally more complicated than for finite graphs at $n = 4$, particularly when all the lines are massless, as in many QCD calculations.

Although there is no such object as a genuine vector space of finite noninteger dimension, it is possible to construct an operation that behaves as if it were an integration over such a space. The operation was proved unique by Wilson, and explicit constructions have been made, so that consistency is assured at the level of all Feynman graphs. Whether a satisfactory definition beyond perturbation theory exists remains to be determined.

It is convenient to arrange that the renormalized coupling is dimensionless in the regulated theory. This is done by changing the normalization of λ with the aid of an extra parameter, the unit of mass μ:

$$\lambda_0 = \mu^{2\epsilon}(\lambda + \text{counter-terms}) \qquad [15]$$

with λ and μ being held fixed when $\epsilon \longrightarrow 0$. (Thus, the basic interaction in eqn [13] is changed to $-\lambda\mu^{2\epsilon}\phi^4/4!$.) Then for the one-loop graph of eqn [2], dimensionally regularized Feynman parameter methods give

$$-i\lambda^2 I(p; m, \epsilon) = \frac{i\lambda^2}{32\pi^2}(4\pi)^\epsilon \Gamma(\epsilon) \times \int_0^1 dx \left[\frac{m^2 - p^2 x(1-x) - i0}{\mu^2}\right]^{-\epsilon} \qquad [16]$$

A natural renormalization procedure is to subtract the pole at $\epsilon = 0$, but it is convenient to accompany this with other factors to remove some universally occurring finite terms. So $\overline{\text{MS}}$ renormalization ("modified minimal subtraction") is defined by using the counter-term

$$-iA(\epsilon)\lambda^2 = -i\frac{\lambda^2 S_\epsilon}{32\pi^2\epsilon} \qquad [17]$$

where $S_\epsilon \stackrel{\text{def}}{=} (4\pi e^{-\gamma_E})^\epsilon$, with $\gamma_E = 0.5772\ldots$ being the Euler constant. This gives a renormalized integral (at $\epsilon = 0$)

$$-\frac{i\lambda^2}{32\pi^2}\int_0^1 dx \ln\left[\frac{m^2 - p^2 x(1-x)}{\mu^2}\right] \qquad [18]$$

which can be evaluated easily. A particularly simple result is obtained at $m = 0$:

$$\frac{i\lambda^2}{32\pi^2}\left[-\ln\frac{-p^2}{\mu^2} + 2\right] \qquad [19]$$

This formula symptomizes important and very useful algorithmic simplifications in the higher-order massless calculations common in QCD.

The $\overline{\text{MS}}$ scheme amounts to a *de facto* standard for QCD. At higher orders a factor of $S_\epsilon{}^L$ is used in the counter-terms, with L being the number of loops.

Coordinate Space

Quantum fields are written as if they are functions of x, but they are in fact distributions or generalized functions, with quantum-mechanical operator values. This indicates that using products of fields is dangerous and in need of careful definition. The relation with ordinary distribution theory is simplest in the coordinate-space version of Feynman graphs. Indeed in the 1950s, Bogoliubov and Shirkov formulated renormalization as a problem of defining products of the singular numeric-valued distributions in coordinate-space Feynman graphs; theirs was perhaps the best treatment of renormalization in that era.

For example, the coordinate-space version of eqn [5] is

$$-\lambda^2 \lim_{a\to 0} \int \mathrm{d}^4x\, \mathrm{d}^4y\, f(x,y) \times \left[\tfrac{1}{2}\tilde{S}(x-y;m,a)^2 + \mathrm{i}A(a)\delta^{(4)}(x-y)\right] \quad [20]$$

where x and y are the coordinates for the interaction vertices, $f(x,y)$ is the product of external-line free propagators, and $\tilde{S}(x-y;m,a)$ is the coordinate-space free propagator, which at $a=0$ has a singularity

$$\frac{1}{4\pi^2[-(x-y)^2+\mathrm{i}0]} \quad [21]$$

as $(x-y)^2 \to 0$. We see in eqn [20] a version of the Hadamard finite part of a divergent integral, and renormalization theory generalizes this to particular kinds of arbitrarily high-dimension integrals. The physical realization and justification of the use of the finite-part procedure is in terms of renormalization of parameters in the Lagrangian; this also gives the procedure a significance that goes beyond the integrals themselves and involves the full nonperturbative formulation of QFT.

General Counter-Term Formulation

We have written $\mathcal{L}$ as a basic Lagrangian density plus counter-terms, and have seen in an example how to cancel divergences at one-loop order. In this section, we will see how the procedure works to all orders. The central mathematical tool is Bogoliubov's R-operation. Here the counter-terms are expanded as a sum of terms, one for each basic one-particle irreducible (1PI) graph with a non-negative degree of divergence. To each basic graph for a Green function is added a set of counter-term graphs associated with divergences for subgraphs. The central theorem of renormalization is that this procedure does in fact remove all the UV divergences, with the form of the counter-terms being determined by the simple computation of the degree of divergence for 1PI graphs.

To see the essential difficulty to be solved, consider a two-loop graph like the first one in **Figure 3**. Its divergence is not a polynomial in external momenta, and is therefore not canceled by an allowed counter-term. This is shown by differentiation with respect to external momenta, which does not produce a finite result because of the divergent one-loop subgraph. But for consistency of the theory, the one-loop counter-terms already computed must be themselves put into loop graphs. Among others, this gives the second graph of **Figure 3**, where the cross denotes that a counter-term contribution is used. The contribution used here is actually 2/3 of the total one-loop counter-term, for reasons of symmetry factors that are not fully evident at first sight. The remainder of the one-loop coupling renormalization cancels a subdivergence in another two-loop graph. It is readily shown that the divergence of the sum of the first two graphs in **Figure 3** is momentum independent, and thus can be canceled by a vertex counter-term.

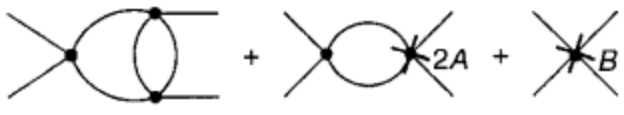


Figure 3 A two-loop graph and its counter-terms. The label B indicates that it is the two-loop overall counter-term for this graph.

This method is fully general, and is formalized in the Bogoliubov R-operation, which gives a recursive specification of the renormalized value $R(G)$ of a graph G:

$$R(G) \stackrel{\text{def}}{=} G + \sum_{\{\gamma_1,\ldots,\gamma_n\}} G|_{\gamma_i \to C(\gamma_i)} \quad [22]$$

The sum is over all sets of nonintersecting 1PI subgraphs of G, and the notation $G|_{\gamma_i \to C(\gamma_i)}$ denotes G with all the subgraphs γ_i replaced by associated counter-terms $C(\gamma_i)$. The counter-term $C(\gamma)$ of a 1PI graph γ has the form

$$C(\gamma) \stackrel{\text{def}}{=} -T(\gamma + \text{counter-terms for subdivergences}) \quad [23]$$

Here T is an operation that extracts the divergent part of its argument and whose precise definition gives the renormalization scheme. For example, in minimal subtraction we define

$$T(\Gamma) = \text{pole part at } \epsilon = 0 \text{ of } \Gamma \quad [24]$$

We formalize the term inside parentheses in eqn [23] as

$$\bar{R}(\gamma) \stackrel{\text{def}}{=} \gamma + \text{counterterms for subdivergences} = \gamma + {\sum_{\{\gamma_1,\ldots,\gamma_n\}}}' G|_{\gamma_i \to C(\gamma_i)} \quad [25]$$

where the prime on the $\sum'$ denotes that we sum over all sets of nonintersecting 1PI subgraphs except for the case that there is a single γ_i equal to the whole graph (i.e., the term with $n=1$ and $\gamma_1=\gamma$ is omitted).

Note that, for the $\overline{\text{MS}}$ scheme, we define the T operation to be applied to a factor of constant dimension obtained by taking the appropriate power of μ^ϵ outside of the pole-part operation. Moreover, it is not a strict pole-part operation; instead each

pole is to be multiplied by $S_\epsilon{}^L$, where L is the number of loops, and S_ϵ is defined after eqn [17].

Equations [22]–[25] give a recursive construction of the renormalization of an arbitrary graph. The recursion starts on one-loop graphs, since they have no subdivergences, that is, $C(\gamma) = -T(\gamma)$ for a one-loop 1PI graph.

Each counter-term $C(\gamma)$ is implemented as a contribution to the counter-term Lagrangian. The Feynman rules ensure that once $C(\gamma)$ has been computed, it appears as a vertex in bigger graphs in such a way as to give exactly the counter-terms for subdivergences used in the R-operation. It has been proved that the R-operation does in fact give finite results for Feynman graphs, and that basic power counting in exactly the same fashion as at one-loop determines the relevant operators.

In early treatments of renormalization, a problem was caused by graphs like **Figure 4**. This graph has three divergent subgraphs which overlap, rather than being nested. Within the R-operation approach, such cases are no harder to deal with than merely nested divergences.

The recursive specification of R-operation can be converted to a nonrecursive formulation by the forest formula of Zavyalov and Stepanov, later rediscovered by Zimmerman. It is normally the recursive formulation that is suited to all-orders proofs.

Whether these results, proved to all orders of perturbation theory, genuinely extend to the complete theory is not so easy to answer, certainly in a realistic four-dimensional QFT. One illuminating case is of a nonrelativistic quantum mechanics model with a delta-function potential in a two-dimensional space. Renormalization can be applied just as in field theory, but the model can also be treated exactly, and it has been shown that the results agree with perturbation theory.

Perturbation series in relativistic QFTs can at best be expected to be asymptotic, not convergent. So instead of a radius of convergence, we should talk about a region of applicability of a weak-coupling expansion. In a direct calculation of counter-terms, etc., the radius of applicability shrinks to zero as the regulator is removed. However, we can deduce the expansion for a renormalized quantity, whose expansion is expected to have a nonzero range of applicability. We can therefore appeal to the uniqueness of power series expansions to allow the calculation, at intermediate stages, to use bare quantities that are divergent as the regulator is removed.

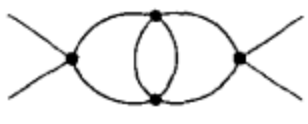


Figure 4 Graph with overlapping divergent subgraphs.

Renormalizability, Non-Renormalizability, and Super-Renormalizability

The basic power-counting method shows that if a theory with conventional fields (at $n=4$) has only operators of dimension 4 or less in its $\mathcal{L}$, then the necessary counter-term operators are also of dimension 4 or less. So if we start with a Lagrangian with all possible such operators, given the field content, then the theory is renormalizable. This is not the whole story, as we will see in the discussion of gauge theories.

If we start with a Lagrangian containing operators of dimension higher than 4, then renormalization requires operators of ever higher dimension as counter-terms when one goes to higher orders in perturbation theory. Therefore, such a theory is said to be perturbatively non-renormalizable. Some very powerful methods of cancelation or some nonperturbative effects are needed to evade this result.

In the case of dimension-4 interactions, there is only a finite set of operators given the set of basic fields, but divergences occur at arbitrarily high orders in perturbation theory. If, instead, all the operators have at most dimension 3, then only a finite number of graphs need counter-terms. Such theories are called super-renormalizable. The divergent graphs also occur as subgraphs inside bigger graphs, of course. There is only one such theory in a four-dimensional spacetime: ϕ^3 theory, which suffers from an energy density that is unbounded from below, so it is not physical. In lower spacetime dimension, where the requirements on operator dimension are different, there are many more known super-renormalizable theories, some with a very rigorous proof of existence.

All the above characterizations rely primarily on perturbative analysis, so they are subject to being not quite accurate in an exact theory, but they form a guide to the relevant issues.

Renormalization and Symmetries: Gauge Theories

In most physical applications, we are interested in QFTs whose Lagrangian is restricted to obey certain symmetry requirements. Are these symmetries preserved by renormalization? That is, is the Lagrangian with all necessary counter-terms still invariant under the symmetry?

We first discuss nonchiral symmetries; these are symmetries in which the left-handed and right-handed parts of Dirac fields transform identically.

For Poincaré invariance and simple global internal symmetries, it is simplest to use a regulator, like dimensional regularization, which respects the symmetries. Then it is easily shown that the symmetries are preserved under renormalization. This holds even if the internal symmetries are spontaneously broken (as happens with a "wrong-sign mass term," e.g., negative m^2 in eqn [1]).

The case of local gauge symmetries is harder. But their preservation is more important, because gauge theories contain vector fields which, without a gauge symmetry, generally give unphysical features to the theory. For perturbation theory, BRST quantization is usually used, in which, instead of gauge symmetry, there is a BRST supersymmetry. This is manifested at the Green function level by Slavnov–Taylor identities that are more complicated, in general, than the Ward identities for simple global symmetries and for abelian local symmetries.

Dimensional regularization preserves these symmetries and the Slavnov–Taylor identities. Moreover, the R-operation still produces finite results with local counter-terms, but cancelations and relations occur between divergences for different graphs in order to preserve the symmetry. A simple example is QED, which has an abelian U(1) gauge symmetry, and whose gauge-invariant Lagrangian is

$$\mathcal{L} = -\tfrac{1}{4}\left(\partial_\mu A_\nu^{(0)} - \partial_\nu A_\mu^{(0)}\right)^2 + \bar{\psi}_0\left(\mathrm{i}\gamma^\mu\partial_\mu - e_0 A_\mu^{(0)} - m_0\right)\psi_0 \qquad [26]$$

At the level of individual divergent 1PI graphs, we get counter-terms proportional to ${A_\mu}^2$ and to $({A_\mu}^2)^2$, operators not present in the gauge-invariant Lagrangian. The Ward identities and Slavnov–Taylor identities show that these counter-terms cancel when they are summed over all graphs at a given order of renormalized perturbation theory. Moreover, the renormalization of coupling and the gauge field are inverse, so that $e_0 A_\mu^{(0)}$ equals the corresponding object with renormalized quantities, $\mu^\epsilon e A_\mu$. Naturally, sums of contributions to a counter-term in $\mathcal{L}$ can only be quantified with use of a regulator.

In nonabelian theories, the gauge-invariance properties are not just the absence of certain terms in $\mathcal{L}$ but quantitative relations between the coefficients of terms with different numbers of fields. Even so, the argument with Slavnov–Taylor identities generalizes appropriately and proves renormalizability of QCD, for example. But note that the relation concerning the product of the coupling and the gauge field does not generally hold; the form of the gauge transformation is itself renormalized, in a certain sense.

Anomalies

Chiral symmetries, as in the weak-interaction part of the gauge symmetry of the standard model, are much harder to deal with. Chiral symmetries are ones for which the left-handed and right-handed components of Dirac field transform independently under different components of the symmetry group, local or global as the case may be. Occasionally, some or other of the left-handed or right-handed components may not even be present.

In general, chiral symmetries are not preserved by regularization, at least not without some other pathology. At best one can adjust the finite parts of counter-terms such that in the limit of the removal of the regulator, the Ward or Slavnov–Taylor identities hold. But in general, this cannot be done consistently, and the theory is said to suffer from an anomaly. In the case of chiral gauge theories, the presence of an anomaly prevents the (candidate) theory from being valid. A dramatic and nontrivial result (Adler–Bardeen theorem and some nontrivial generalizations) is that if chiral anomalies cancel at the one-loop level, then they cancel at all orders.

Similar results, but more difficult ones, hold for supersymmetries.

The anomaly cancelation conditions in the standard model lead to constraints that relate the lepton content to the quark content in each generation. For example, given the existence of the b quark, and the τ and ν_τ leptons (of masses around 4.5 GeV, 1.8 GeV, and zero respectively), it was strongly predicted on the grounds of anomaly cancelation that there must be a t quark partner of the b to complete the third generation of quark doublets. This prediction was much later vindicated by the discovery of the much heavier top quark with $m_t \simeq 175$ GeV.

Renormalization Schemes

A precise definition of the counter-terms entails a specification of the renormalization prescription (or scheme), so that the finite parts of the counter-terms are determined. This apparently induces extra arbitrariness in the results. However, in the ϕ^4 Lagrangian (for example), there are really only two independent parameters. (A scaling of the field does not affect any observables, so we do not count Z as a parameter here.) Thus, at fixed regulator parameter a or ϵ, renormalization actually just gives a reparametrization of a two-parameter collection of theories. A renormalization prescription gives the

change of variables between bare and renormalized parameters, a rather singular transformation when the regulator is removed. If we have two different prescriptions, we can deduce a transformation between the renormalized parameters in the two schemes. The renormalized mass and coupling m_1 and λ_1 in one scheme can be obtained as functions of their values m_2 and λ_2 in the other scheme, with the bare parameters, and hence the physics, being the same in both schemes. Since these are renormalized parameters, the removal of the regulator leaves the transformation well behaved.

Generalization to all renormalizable theories is immediate.

Renormalization Group and Applications and Generalizations

One part of the choice of renormalization scheme is that of a scale parameter such as the unit of mass μ of the $\overline{\text{MS}}$ scheme. The physical predictions of the theory are invariant if a change of μ is accompanied by a suitable change of the renormalized parameters, now considered as μ-dependent parameters $\lambda(\mu)$ and $m(\mu)$. These are called the effective, or running, coupling and mass. The transformation of the parametrization of the theory is called an RG transformation.

The bare coupling and mass λ_0 and m_0 are RG invariant, and this can be used to obtain equations for the RG evolution of the effective parameters from the perturbatively computed counter-terms. For example, in ϕ^4 theory, we have (in the renormalized theory after removal of the regulator)

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\ln\mu^2} = \beta(\lambda) \qquad [27]$$

with $\beta(\lambda) = 3\lambda^2/(16\pi^2) + O(\lambda^3)$. As exemplified in eqns. [18] and [19], Feynman diagrams depend logarithmically on μ. By choosing μ to be comparable to the physical external momentum scale, we remove possible large logarithms in this and higher orders. Thus, provided that the effective coupling at this scale is weak, we get an effective perturbation expansion.

This is a basic technique for exploiting perturbation theory in QCD, for the strong interactions, where the interactions are not automatically weak. In this theory the RG β function is negative so that the coupling decreases to zero as $\mu \to \infty$; this is the asymptotic freedom of QCD.

A closely related method is that associated with the Callan–Symanzik equation, which is a formulation of a Ward identity for anomalously broken scale invariance. However, RG methods are the actually used ones, normally, even if sometimes an RG equation is incorrectly labeled as a Callan–Symanzik equation.

The elementary use of the RG is not sufficient for most interesting processes, which involve a set of widely different scales. Then more powerful theorems come into play. Typical are the factorization theorems of QCD (*see* Quantum Chromodynamics). These express differential cross sections for certain important reactions as a product of quantities that involve a single scale:

$$\mathrm{d}\sigma = C(Q, \mu, \lambda(\mu)) \otimes f(m, \mu, \lambda(\mu)) + \text{small correction} \qquad [28]$$

The product is typically a matrix or a convolution product. The factors obey nontrivial RG equations, and these enable different values of μ to be used in the different factors. Predictions arise because some factors and the kernels of the RG equation are perturbatively calculable, with a weak effective coupling. Other factors, such as f in eqn [28], are not perturbative. These are quantities with names like "parton distribution functions," and they are universal between many different processes. Thus, the nonperturbative functions can be measured in a limited set of reactions and used to predict cross sections for many other reactions with the aid of calculations of the perturbative factors.

Ultimately, this whole area depends on physical phenomena associated with renormalization.

Concluding Remarks

The actual ability to remove the divergences in certain QFTs to produce consistent, finite, and nontrivial theories is a quite dramatic result. Moreover, associated with the integrals that give the divergences is behavior of the kind that is analyzed with RG methods and generalizations. So the properties of QFTs associated with renormalization get tightly coupled to many interesting consequences of the theories, most notably in QCD.

QFTs are actually very abstruse and difficult theories; only certain aspects currently lend themselves to practical calculations. So the reader should not assume that all aspects of their rigorous mathematical treatment are perfect. Experience, both within the theories and in their comparison with experiment, indicates, nevertheless, that we have a good approximation to the truth.

When one examines the mathematics associated with the R-operation and its generalizations with factorization theorems, there are clearly present some interesting mathematical structures that are not yet formulated in their most general terms. Some

indications of this can be seen in the work by Connes and Kreimer (*see* Hopf Algebra Structure of Renormalizable Quantum Field Theory), where it is seen that renormalization is associated with a Hopf algebra structure for Feynman graphs.

With such a deep subject, it is not surprising that it lends itself to other approaches, notably the Connes–Kreimer one and the Wilsonian one (*see* Exact Renormalization Group). Readers new to the subject should not be surprised if it is difficult to get a fully unified view of these different approaches.

Notes on Bibliography

Reliable textbooks on quantum field theory are Sterman (1993) and Weinberg (1995). A clear account of the foundations of perturbative QCD methods is given by Sterman (1996).

A pedagogical account of renormalization and related subjects may be found in Collins (1984). The best account of renormalization theory before the 1970s is given by Bogoliubov and Shirkov (1959); the viewpoint is very modern, including a coordinate-space distribution-theoretic view. A full account of the Wilsonian method as applied to renormalization is given by Polchinski (1984).

Manuel and Tarrach (1994) give an excellent account of renormalization for a theory with a non-relativistic delta-function potential in 2 space dimensions, which provides a fully tractable model.

Tkachov (1994) reviews a systematic application of distribution theoretic methods to asymptotic problems in QFT. Finally, Weinzierl (1999) provides a construction of dimensional regularization with the aid of K-theory using an underlying vector space of the physical integer dimension. Other constructions, referred to in this paper, follow Wilson and use an infinite-dimensional underlying space.

Acknowledgments

The author would like to thank Professor K Goeke for hospitality and support at the Ruhr-Universität-Bochum. This work was also supported by the US DOE.

See also: Anomalies; BRST Quantization; Effective Field Theories; Electroweak Theory; Euclidean Field Theory; Exact Renormalization Group; High T_c Superconductor Theory; Holomorphic Dynamics; Hopf Algebra Structure of Renormalizable Quantum Field Theory; Lattice Gauge Theory; Operator Product Expansion in Quantum Field Theory; Perturbation Theory and its Techniques; Perturbative Renormalization Theory and BRST; Quantum Chromodynamics; Quantum Field Theory: A Brief Introduction; Singularities of the Ricci Flow; Standard Model of Particle Physics; Supergravity.

Further Reading

Bogoliubov NN and Shirkov DV (1959) *Introduction to the Theory of Quantized Fields.* New York: Wiley-Interscience.

Collins JC (1984) *Renormalization.* Cambridge: Cambridge University Press.

Kraus E (1998) Renormalization of the electroweak standard model to all orders. *Annals of Physics* 262: 155–259.

Manuel C and Tarrach R (1994) Perturbative renormalization in quantum mechanics. *Physics Letters B* 328: 113–118.

Polchinski J (1984) Renormalization and effective Lagrangians. *Nuclear Physics B* 231: 269–295.

Sterman G (1993) *An Introduction to Quantum Field Theory.* Cambridge: Cambridge University Press.

Sterman G (1996) Partons, factorization and resummation. In: Soper DE (ed.) *QCD and Beyond*, pp. 327–406. (hep-ph/9606312). Singapore: World Scientific.

Tkachov FV (1994) Theory of asymptotic operation. A summary of basic principles. *Soviet Journal of Particles and Nuclei* 25: 649.

Weinberg S (1995) *The Quantum Theory of Fields, Vol. I. Foundations.* Cambridge: Cambridge University Press.

Weinzierl S (1999) Equivariant dimensional regularization, hep-ph/9903380.

Zavyalov OI (1990) *Renormalized Quantum Field Theory.* Dordrecht: Kluwer.

Seiberg–Witten Theory

Siye Wu, University of Colorado, Boulder, CO, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Gauge theory is the cornerstone of the standard model of elementary particles. The original motivation for studying supersymmetric gauge theories was phenomenological (such as the hierarchy problem). They display a large number of interesting phenomena and become the models for the dynamics of strongly coupled field theories. They also offer valuable insights to nonsupersymmetric models. In $N=1$ gauge theory, the low-energy effective superpotential is holomorphic both in the superfields and in the coupling constants. This powerful holomorphy principle, together with symmetry and various limits, often determines the effective superpotential completely. Such theories often have quantum moduli spaces where the classical singularities are smoothed out, continuous interpolation between Higgs and confinement phases, massless composite mesons and baryons, and dual theories weakly coupled at low energy. For $N=2$ pure gauge theory, the low-energy effective theory is an abelian gauge theory in which both the kinetic term and the coupling constant are determined by a holomorphic prepotential. The electric–magnetic duality is in the ambiguity of the low-energy description. Much physical information, such as the coupling constant, the Kähler metric on the quantum moduli, the monodromy around the singularities, can be incorporated in a family of elliptic curves. This low-energy exact solution is also useful to topological field theory that can be obtained from the $N=2$ theory by twisting. Much of the above was the work of Seiberg and Witten in the mid-1990s. In this article, we review some of the fascinating aspects of $N=1$ and $N=2$ supersymmetric gauge theories.

$N=1$ Gauge Theory and Seiberg Dualities

$N=1$ Yang–Mills Theory and QCD

Let G be a compact Lie group and let P be a principal G-bundle over the Minkowski space $\mathbb{R}^{3,1}$. In pure gauge theory, the dynamical variable is a connection A in P; two connections are equivalent if they are related by a gauge transformation. Let $F \in \Omega^2(\mathbb{R}^{3,1}, \mathrm{ad}\, P)$ be the curvature of A. It decomposes into the self-dual and anti-self-dual parts, that is, $F = F^+ + F^-$, where $F^{\pm} = (1/2)(F \mp \sqrt{-1} * F)$. With a suitably normalized nondegenerate bilinear form $\langle \cdot, \cdot \rangle$ on the Lie algebra $\mathfrak{g}$, the classical action is

$$\begin{aligned} S_{\mathrm{YM}}[A] &= \int_{\mathbb{R}^{3,1}} -\frac{1}{2g^2}\langle F \wedge *F\rangle + \frac{\theta}{16\pi^2}\langle F \wedge F\rangle \\ &= \int_{\mathbb{R}^{3,1}} -\frac{\tau}{8\pi}\langle F^+ \wedge F^+\rangle - \frac{\bar{\tau}}{8\pi}\langle F^- \wedge F^-\rangle \end{aligned}$$

Here $g > 0$ is the coupling constant and $\theta \in \mathbb{R}$, the θ angle, and

$$\tau = \frac{\theta}{2\pi} + \frac{4\pi\sqrt{-1}}{g^2}$$

is a complex number in the upper-half plane that incorporates both. Classically, the theory is conformally invariant and the dynamics is independent of the θ-term. At the quantum level, $\theta(\mathrm{mod}2\pi)$ appears in the path integral and parametrizes inequivalent vacua. The coupling constant runs as energy μ varies, satisfying the renormalization group equation

$$\mu\frac{\mathrm{d}g}{\mathrm{d}\mu} = -\frac{b_0}{(4\pi)^2}g^3 + o(g^5)$$

where the right-hand side is called the β-function $\beta(g)$. This introduces, when $b_0 \neq 0$, a mass scale Λ given by

$$(\Lambda/\mu)^{b_0} = \mathrm{e}^{-8\pi^2/g(\mu)^2}$$

Notes:

up to one-loop. Consequently, the classical scale invariance is lost. It is convenient to redefine Λ as a complex quantity such that

$$(\Lambda/\mu)^{b_0} = e^{2\pi\sqrt{-1}\tau(\mu)}$$

For pure gauge theory, $b_0 = (11/3)\check{h}$, where $\check{h}$ is the dual Coxeter number of $\mathfrak{g}$. At high energy ($\mu \to \infty$), the coupling becomes weak ($g \to 0$); this is known as asymptotic freedom. On the contrary, the interaction becomes strong at low energy. It is believed that the theory exhibits confinement and has a mass gap.

QCD, or quantum chromodynamics, is gauge theory coupled to matter fields. Suppose the boson ϕ and the fermion ψ are in the (complex) representations R_b and R_f of G, respectively. That is, $\phi \in \Gamma(P \times_G R_b)$, or ϕ is a section of the bundle $P \times_G R_b$, and $\psi \in \Gamma(S \otimes (P \times_G R_f))$, where S is the spinor bundle over $\mathbb{R}^{3,1}$. The classical action is

$$S_{\mathrm{QCD}}[A,\phi,\psi] = S_{\mathrm{YM}}[A] + \frac{1}{g^2}\int \mathrm{d}^4x \frac{1}{2}|\nabla\phi|^2 + \sqrt{-1}(\psi, \not{\nabla}\psi) + \cdots$$

where ∇ is the covariant derivative, $\not{\nabla}$ is the Dirac operator coupled to A, and we have omitted possible mass and potential terms. The quantum theory depends sensitively on the representations R_b and R_f. In the β-function, we have

$$b_0 = \tfrac{11}{3}\check{h} - \tfrac{1}{6}\nu(R_b) - \tfrac{2}{3}\nu(R_f)$$

where $\nu(R)$ is the Dynkin index of a representation R. If $b_0 < 0$, the theory is free in the infrared but strongly interacting in the ultraviolet. If $b_0 > 0$, the converse is true; in particular, the theory exhibits asymptotic freedom. If $b_0 = 0$, the situation depends on the sign of the two or higher-loop contributions.

Pure $N = 1$ supersymmetric gauge theory is one on the superspace $\mathbb{R}^{3,1|(2,2)}$ with a constraint that the curvature vanishes in the odd directions. The dynamical variables are in the superfield strength W, a $1|(1,0)$-form valued in ad P. In components, the theory is gauge field coupled to a Majorana or Weyl fermion in the adjoint representation. Let $S^{\pm}$ be spinor bundles of positive (negative) chiralities, respectively, and let λ be a section of $S^+ \otimes \mathrm{ad}P$. The action, written both in superspace and in ordinary spacetime, is

$$S^{N=1}_{\mathrm{SYM}}[A,\lambda] = \frac{1}{4\pi}\mathrm{Im}\left(\int \mathrm{d}^4x\, \mathrm{d}^2\theta\, \tau\langle W, W\rangle\right) = S_{\mathrm{YM}}[A] + \frac{1}{g^2}\int \mathrm{d}^4x \sqrt{-1}\langle\bar{\lambda}, \not{\nabla}^+\lambda\rangle$$

Since $b_0 = 3\check{h}$, the theory is asymptotically free but strongly coupled at low energy. Classically, the theory has a $\mathrm{U}(1)_R$ chiral symmetry. However, due to anomaly, only the subgroup $\mathbb{Z}_{2\check{h}}$ survives at the quantum level. Instanton effect yields gaugino condensation $\langle\lambda\lambda\rangle \sim \Lambda^3$. The symmetry is thus further broken to $\mathbb{Z}_2$, resulting $\check{h}$ inequivalent vacua.

The $N = 1$ QCD has additional chiral superfields Φ in a representation R, including the bosons $\phi \in \Gamma(P \times_G R)$ and the fermions $\psi \in \Gamma(S^+ \otimes (P \times_G R))$. In the absence of superpotential, the action is

$$S^{N=1}_{\mathrm{SQCD}}[A,\lambda,\phi,\psi] = S^{N=1}_{\mathrm{SYM}}[A,\lambda] + \frac{1}{g^2}\int \mathrm{d}^4x\, \mathrm{d}^2\theta\, d^2\bar{\theta}\ \tfrac{1}{2}|\Phi|^2$$

In components, the second term is

$$\frac{1}{g^2}\int \mathrm{d}^4x \left(\tfrac{1}{2}|\nabla\phi|^2 + \sqrt{-1}(\psi, \not{\nabla}^+\psi) - \tfrac{1}{2}\phi^*|D|^2 + \cdots\right)$$

where $D: R \to \mathfrak{g}^*$ is the moment map of the Hamiltonian G-action on R, and we have omitted other terms containing fermionic fields. The moduli space of classical vacua is the symplectic quotient $D^{-1}(0)/G = R/\!/G$. It is the same as the Kähler quotient $R^s/G^{\mathbb{C}}$, where the stable subset $R^s = \{\phi \in R | G^{\mathbb{C}} \cdot \phi \cap D^{-1}(0) \neq \emptyset\}$ is open and dense in R. Again, the quantum theory depends on the representation R. Since $b_0 = 3\check{h} - (1/2)\nu(R)$, the theory is asymptotically free, infrared free, scale invariant (to one-loop) when $\nu(R) < 6\check{h}, \nu(R) > 6\check{h}, \nu(R) = 6\check{h}$, respectively. The moduli space may be lifted by a superpotential or modified by other quantum effects.

$\mathrm{SU}(N_c)$ Theories at Low Energy

We now consider $N = 1$ QCD with $G = \mathrm{SU}(N_c)$; N_c is the number of colors. The matter field consists of N_f copies of quarks $Q^i (1 \le i \le N_f)$ in the fundamental representation of $\mathrm{SU}(N_c)$ and N_f copies of antiquarks $Q'_{i'} (1 \le i' \le N_f)$ in the conjugate representation. Using the isomorphism of $\mathfrak{su}(N_c)$ with its dual, the moment map is

$$D(Q,Q') = \text{traceless part of } \sqrt{-1}(QQ^\dagger - Q'Q'^\dagger)$$

So $(Q,Q') \in D^{-1}(0)$ if and only if $QQ^\dagger - Q'^\dagger = cI_{N_c}$ for some $c \in \mathbb{R}$. If $N_f < N_c$, then $c = 0$ and

$$Q, Q' \sim \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_{N_f} \end{pmatrix}$$

for some $a_k \geq 0$. Generically, these $a_k > 0$ and the gauge group $SU(N_c)$ is broken to $SU(N_c - N_f)$. If $N_f \geq N_c$, then

$$Q \sim \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_{N_c} \end{pmatrix}, \qquad Q' \sim \begin{pmatrix} a'_1 & & \\ & \ddots & \\ & & a'_{N_c} \end{pmatrix}$$

where $a_k, a'_k \geq 0$ satisfy $a_k^2 - a_k'^2 = c$ for some $c \in \mathbb{R}$. The gauge group is completely broken. The low-energy superfields are the mesons $M^i_{i'} = Q^i Q'_{i'}$ and, if $N_f \geq N_c$, the baryons

$$B_{i_{N_c+1}\cdots i_{N_f}} = \frac{1}{N_c!}\epsilon_{i_1\cdots i_{N_f}} Q^{i_1}\cdots Q^{i_{N_c}}$$
$$B'^{i'_{N_c+1}\cdots i'_{N_f}} = \frac{1}{N_c!}\epsilon^{i'_1\cdots i'_{N_f}} Q'_{i'_1}\cdots Q'_{i'_{N_c}}$$

When $N_f < N_c$, Affleck *et al.* (1984) found a dynamically generated superpotential

$$W_{\text{eff}}(\hat{M}) = (N_c - N_f)\left(\frac{\Lambda^{3N_c - N_f}}{\det M}\right)^{1/(N_c - N_f)}$$

generated by instanton effect when $N_f = N_c - 1$ and by gaugino condensation in the unbroken $SU(N_c - N_f)$ theory when $N_f < N_c - 1$. It is also the unique superpotential (up to a multiplicative constant) that is consistent with the global and supersymmetry. The potential pushes the vacuum to infinity. Therefore, contrary to the classical picture, theories with $N_f < N_c$ do not have a vacuum at the quantum level.

When $N_f \geq 3N_c$, the theory is not strongly interacting at low energy, and perturbation methods are reliable. (When $N_f = 3N_c$, the two-loop contribution to the β-function is negative.) We now look at the range $N_c \leq N_f < 3N_c$. The cases $N_f = N_c, N_c + 1$ and $N_c + 2 \leq N_f < 3N_c$ were studied in Seiberg (1994) and Seiberg (1995), respectively.

When $N_f = N_c$, the classical moduli space is $\det M = BB'$. The quantum theory at low energy consists of the fields M, B, B' satisfying the constraint $\det M - BB' = \Lambda^{2N_c}$. The quantum moduli space is smooth everywhere, and there are no additional massless particles. So the gluons are heavy throughout the moduli space. This is due to confinement near the origin, where the interaction is strong, and due to the Higgs mechanism far out in the flat direction, where the classical picture is a good approximation. We see a smooth transition between these two effects.

When $N_f = N_c + 1$, there is a dynamically generated superpotential

$$W_{\text{eff}} = \frac{1}{\Lambda^{2N_c - 1}}(B'MB - \det M)$$

The stationary points of W_{eff} are at $BB' - \wedge^{N_c} M = 0$, $BM = 0, MB' = 0$; these are precisely the constraints that the classical configuration satisfies. However, the moduli space is interpreted differently: it is embedded into a larger space, and the constraints are satisfied only at stationary points. At the singularity $\langle M \rangle = 0$, the whole global symmetry group is unbroken, and B, B' are the new massless fields resolving the singularity. So we have a continuous transition between confinement (without chiral symmetry breaking) and the Higgs mechanism in the semiclassical regime.

When $N_c + 2 \leq N_f \leq (3/2)N_c$, the original theory, called the electric theory, is still strongly coupled in the infrared. Seiberg (1995) proposed that there is a dual, magnetic theory, which is infrared free. The two theories are different classically, but are equivalent at the quantum level. The dual theory is an $N = 1 SU(\tilde{N}_c)$ gauge theory with $\tilde{N}_c = N_f - N_c$, coupled to dual quarks $\tilde{Q}_i, \tilde{Q}^{i'}$, where $1 \leq i; i' \leq N_f$ are flavor indices. In addition, the mesons $M^i_{i'}$ become fundamental fields. They are not coupled to the $SU(\tilde{N}_c)$ gauge field but interact with the dual quarks through the superpotential

$$W = \mu^{-1} M^i_{i'} \tilde{Q}_i \tilde{Q}^{i'}$$

The two theories have the same global symmetry and the same gauge-invariant operators. The dual quarks are fundamental in the magnetic theory but are solitonic excitations in the electric theory. At high energy, the electric theory is asymptotically free, while the magnetic theory is strongly coupled. At low energies, the converse is true. Therefore, reliable perturbative calculations can be performed by choosing an appropriate weakly coupled theory.

When $(3/2)N_c < N_f < 3N_c$, the theory has a nontrivial infrared fixed point. This is because up to two-loop,

$$\beta(g) = -\frac{g^3}{16\pi^2}(3N_c - N_f) + \frac{g^5}{128\pi^4}\left(2N_c N_f - 3N_c^2 - \frac{N_f}{N_c}\right) + o(g^7)$$

There is a solution $g_* > 0$ to $\beta(g) = 0$. We have $\beta(g) < 0$ when $0 < g < g_*, \beta(g) > 0$ when $g > g_*$. In the infrared limit, the coupling constant flows to $g = g_*$, where we have a nontrivial, interacting superconformal theory in four dimensions. The conformal dimension becomes anomalous and is equal to 3/2 of the charge of the chiral $U(1)_R$; for example, that of the meson $\mu^{-1}M$ is $3(N_f - N_c)/N_f > 1$ in this range.

Other Classical Gauge Groups

We now consider $N=1$ supersymmetric gauge theory and QCD with gauge groups $\mathrm{Sp}(N_c)$ and $\mathrm{SO}(N_c)$. The $\mathrm{Sp}(N_c)$ theories, studied by Intriligator and Pouliot (1995), are the simplest examples of the $N=1$ theories. We take $2N_f$ chiral superfields $Q_i (i=1,\ldots,2N_f)$ in the fundamental representation $\mathbf{C}^{2N_c} \cong \mathbf{H}^{N_c}$ of $\mathrm{Sp}(N_c)$. The number of copies must be even so that the quantum theory is free from global gauge anomaly. The gauge-invariant quantities are the mesons $M_{ij} = Q_i^a Q_j^b \omega_{ab}$, where ω is the symplectic form on $\mathbf{C}^{2N_c}$, subject to a constraint $\epsilon^{1,\ldots,2N_c+2} M_{1,2} \cdots M_{2N_c+1,2N_c+2} = 0$. Using the decomposition $\mathfrak{u}(2N_c) = \mathfrak{sp}(N_c) \oplus \sqrt{-1}\{\mathbf{H}\text{-self-adjoint matrices}\}$, the moment map $D(Q)$ is the projection of $\sqrt{-1}QQ^\dagger$ on $\mathfrak{sp}(N_c)$. So $D(Q)=0$ implies

$$Q \sim \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_{\min\{N_c,N_f\}} \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

where $a_k \geq 0$. At a generic point of the classical moduli space, the gauge group is broken to $\mathrm{Sp}(N_c - N_f)$ if $N_c > N_f$; it is completely broken if $N_c \leq N_f$.

Since $b_0 = 3(N_c+1) - N_f$, the quantum theory is infrared free if $N_f \geq 3(N_c+1)$. (When $b_0 = 0$, the two-loop β-function is negative.) When $N_f \leq N_c$, there is a dynamically generated superpotential

$$W_{\mathrm{eff}} = (N_c + 1 - N_f) \times \left(\frac{2^{N_c-1} \Lambda^{3(N_c+1)-N_f}}{\mathrm{Pf}\, M} \right)^{1/(N_c+1-N_f)}$$

pushing the vacuum to infinity.

When $N_f = N_c$, the classical moduli space $\mathrm{Pf} M = 0$ has singularities. The quantum moduli space is $\mathrm{Pf}\, M = 2^{N_c-1} \Lambda^{2(N_c+1)}$. The singularity is smoothed out and there are no light fields other than the mesons M. When $N_f = N_c + 1$, all components of M become dynamical in the low-energy theory, and there is a superpotential

$$W_{\mathrm{eff}} = -\frac{\mathrm{Pf}\, M}{2^{N_c-1} \Lambda^{2N_c+1}}$$

At the most singular point $\langle M \rangle = 0$, the global symmetry is unbroken, and all the light fields in M become massless. In both cases, there is a transition between confinement and Higgs mechanism.

When $N_c + 3 \leq N_f \leq (3/2)(N_c+1)$, there is a dual, magnetic theory which is free in the infrared. The dual theory has $2N_f$ quarks $\tilde{Q}^i$ in the fundamental representation of $\mathrm{Sp}(\tilde{N}_c)$, where $\tilde{N}_c = N_f - N_c - 2$. In addition, the mesons M_{ij} become elementary and couple to $\tilde{Q}$ through a superpotential $W = (2\mu)^{-1} M_{ij} \tilde{Q}^{ia} \tilde{Q}^{jb} \tilde{\omega}_{ab}$, where $\tilde{\omega}$ is the symplectic form on $\mathbf{C}^{2\tilde{N}_c}$. When $(3/2)(N_c+1) < N_f < 3(N_c+1)$, the theory flows to an interacting superconformal field theory in the infrared.

Theories with the $\mathrm{SO}(N_c)$ gauge group were studied by Seiberg (1995) and by Intriligator and Seiberg (1995). Since the fundamental representation is real, there is no constraint on the number N_f of quarks $Q^i (1 \leq i \leq N_f)$. The gauge invariants are the mesons $M^{ij} = Q_a^i Q_b^j \delta^{ab}$ and, if $N_f \geq N_c$, the baryons $B_{i_{N_c+1} \cdots i_{N_f}} = \epsilon_{i_1 \cdots i_{N_f}} Q^{i_1} \cdots Q^{i_{N_c}} / N_c!$ They satisfy rank $M \leq N_c$ and $BB = \wedge^{N_c} M$. Using the decomposition $\mathfrak{u}(N_c) = \mathfrak{so}(N_c) \oplus \sqrt{-1}\{\mathbf{R}\text{-self-adjoint matrices}\}$, the moment map $D(Q)$ is the projection $\sqrt{-1}QQ^\dagger$ on $\mathfrak{so}(N_c)$. If $D(Q)=0$, then up to gauge and global symmetries, Q is of the form

$$Q \sim \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_r \end{pmatrix}$$

where $a_1, \ldots, a_r > 0$ if $r = \mathrm{rank}\, Q \leq N_c$ and $a_1, \ldots, a_{N_c-1} > 0$ and $a_{N_c} \neq 0$ if $r = N_c$. Generically, the gauge group is broken to $\mathrm{SO}(N_c - N_f)$ if $N_c \geq N_f + 2$ and is totally broken if $N_c < N_f + 2$.

We have $b_0 = 3(N_c - 2) - N_f$ if $N_c \geq 5$. For $N_c = 4$, the group is $(\mathrm{SU}(2) \times \mathrm{SU}(2))/\mathbf{Z}_2$ and $b_0 = 6 - N_f$ for each SU(2) factor. If $N_c = 3$, the group is $\mathrm{SU}(2)/\mathbf{Z}_2 b_0 = 6 - 2N_f$. The theory is asymptotically free if $N_f > 3(N_c - 2)$ and infrared free if $N_f \geq 3(N_c - 2)$.

When $N_f \leq N_c - 5$, there is a dynamically generated superpotential

$$W_{\mathrm{eff}} = \frac{1}{2}(N_c - 2 - N_f) \times \left(\frac{16 \Lambda^{3(N_c-2)-N_f}}{\det M} \right)^{1/(N_c-2-N_f)}$$

lifting the classical vacuum degeneracy. The coefficient is fixed by mass deformation and by matching the SU(4) theory when $N_c = 6$.

When $N_f = N_c - 4$, the unbroken gauge group is $\mathrm{SO}(4) = (\mathrm{SU}(2) \times \mathrm{SU}(2))/\mathbf{Z}_2$ on the generic point of the moduli space. The superpotential of the original theory is

$$W_{\mathrm{eff}} = 2(\epsilon_+ + \epsilon_-) \left(\frac{\Lambda^{2(N_c-1)}}{\det M} \right)^{1/2}$$

where the choices $\epsilon_+, \epsilon_- = \pm 1$ correspond to the fact that each of the SU(2) theory has two vacua. There are two physically inequivalent branches: $\epsilon_+ = \epsilon_-$ and $\epsilon_+ = -\epsilon_-$. For $\epsilon_+ = \epsilon_-$, the superpotential pushes the vacuum to infinity. For $\epsilon_+ = -\epsilon_-, W_{\mathrm{eff}} = 0$. In the quantum theory, the singularity is smoothed out and

all the massless fermions are in M, even at the origin of the moduli space. Hence the quarks are confined.

When $N_f = N_c - 3$, the unbroken gauge group is SO(3) and the theory has two branches with

$$W_{\text{eff}} = 4(1+\epsilon)\frac{\Lambda^{2N_c-3}}{\det M}$$

where $\epsilon = \pm 1$. For $\epsilon = 1$, the quantum theory has no vacuum. For $\epsilon = -1$, $W_{\text{eff}} = 0$, but there are additional light fields $\tilde{Q}_i$ coupling to M via the superpotential $W \sim (2\mu)^{-1} M^{ij}\tilde{Q}_i\tilde{Q}_j$ near $M = 0$.

When $N_f = N_c - 2$, the low-energy theory is related to the $N = 2$ gauge theory and will be addressed in the subsection "Seiberg–Witten's low-energy solution."

When $N_f \geq N_c - 1$, we define a dual, magnetic theory whose gauge group is $\text{SO}(\tilde{N}_c)$, where $\tilde{N}_c = N_f - N_c + 4$. There are N_f dual quarks $\tilde{Q}_i (1 \leq i \leq N_f)$ in the fundamental representation. This theory is infrared free if $N_f \leq (3/2)(N_c - 2)$. In the effective theory, the mesons M^{ij} become fundamental and couple with the dual quarks through a superpotential $W = (2\mu)^{-1} M^{ij}\tilde{Q}_i\tilde{Q}_j$ if $N_f \geq N_c$; there is an additional term $\det M / 64\Lambda^{2\tilde{N}_c - 5}$ if $N_f = N_c - 1$. When $(3/2)(N_c - 2) < N_f < 3(N_c - 2)$, the theory flows to an interacting superconformal field theory in the infrared.

$N = 2$ Gauge Theory and Seiberg–Witten Duality

$N = 2$ Yang–Mills Theory

Pure $N = 2$ supersymmetric gauge theory is a special case of $N = 1$ QCD when $R = \mathfrak{g}^{\mathbb{C}}$ is the (complexified) adjoint representation of G. The moment map is $D(\phi) = (1/2\sqrt{-1})[\phi, \bar{\phi}] \in \mathfrak{g} \cong \mathfrak{g}^*(\phi \in \mathfrak{g})$. Since the fermionic fields λ and ψ are sections of the same bundle, there is a second set of supersymmetry transformations by interchanging the roles of λ and ψ. This makes the theory $N = 2$ supersymmetric. The classical action is

$$\begin{aligned} S^{N=2}_{\text{SYM}}[A, \lambda, \psi, \phi] = S_{\text{YM}}[A] \\ + \frac{1}{g^2}\int \mathrm{d}^4 x \sqrt{-1}(\langle \bar{\lambda}, \not{\nabla}\lambda\rangle \\ + \langle \bar{\psi}, \not{\nabla}\psi\rangle) + \frac{1}{2}|\nabla\phi|^2 \\ + \sqrt{-1}(\langle \bar{\phi}, [\lambda, \psi]\rangle + \langle \phi, [\bar{\lambda}, \bar{\psi}]\rangle) \\ - \frac{1}{8}|[\phi, \bar{\phi}]|^2 \end{aligned}$$

The energy reaches the minimum when ϕ takes a constant value $\phi \in \mathfrak{g}^{\mathbb{C}}$ that can be conjugated by G to the Cartan subalgebra $\mathfrak{t}^{\mathbb{C}}$. ($\mathfrak{t}$ is the Lie algebra of the maximal torus T.) The classical moduli space is $\mathfrak{g}^{\mathbb{C}}/G^{\mathbb{C}} = \mathfrak{t}^{\mathbb{C}}/W$, where W is the Weyl group. At a generic $\phi \in \mathfrak{t}^{\mathbb{C}}$, the gauge group is broken to T by the Higgs mechanism. Classically, the massless degrees of freedom are excitations of ϕ and components of the gauge field in $\mathfrak{t}$. So the low-energy physics can be described by these massless fields. However, the moduli space is singular when ϕ is on the walls of the Weyl chambers. At these values, the unbroken gauge group is larger and there are extra massless fields that resolve the singularities.

Since $b_0 = 2\check{h} > 0$, the quantum theory is asymptotically free but strongly interacting at low energy. It can be shown that $N = 1$ supersymmetry already forbids a dynamically generated superpotential on $\mathfrak{t}^{\mathbb{C}}/W$. Therefore, the vacuum degeneracy is not lifted and the quantum moduli space is still a continuum. However, there are corrections to the part of classical moduli space where strong interactions occur. The quantum theory has a dynamically generated mass scale Λ. We pick the renormalization scale μ to be $|\phi|$, the typical energy scale where spontaneous symmetry breaking occurs. Far away from the origin, that is, when $|\phi| \gg |\Lambda|$, the theory is weakly interacting and the classical description of the moduli space is a good approximation. However, when $|\phi|$ is comparable to $|\Lambda|$, the classical language and perturbation methods fail due to strong interaction. At $\phi = 0$, the full gauge symmetry is restored classically. But since the theory becomes strongly interacting at low energy, it cannot be the low-energy solution of the original theory.

The classical $\text{U}(1)_R$ symmetry extends to $\text{U}(2)_R$, mixing λ and ψ. The $\text{U}(1)_R$ subgroup in $\text{U}(2)_R$ is anomalous except for a subgroup $\mathbb{Z}_{4\check{h}}$. So we have a global $\text{SU}(2)_R \times_{\mathbb{Z}_2} \mathbb{Z}_{4\check{h}}$ symmetry at the quantum level. This is consistent with a continuous moduli space of vacua, if the group $\text{SU}(2)_R$ is to act nontrivially. Also, the space is not a single orbit of the global symmetry group. The generator of $\mathbb{Z}_{4\check{h}}$ acts on $\mathfrak{t}^{\mathbb{C}}$ by a phase $\mathrm{e}^{-\pi\sqrt{-1}/\check{h}}$. The group $\mathbb{Z}_{4\check{h}}$ is spontaneously broken to the subgroup which acts trivially on $\mathfrak{t}^{\mathbb{C}}/W$.

We study the general form of low-energy effective Lagrangian that is consistent with $N = 2$ supersymmetry. We assume that the quantum effect does not modify the topology of the moduli space $\mathfrak{t}^{\mathbb{C}}/W$, though it may alter the singularity and its nature. Suppose U is the quantum moduli. At a generic point in U, the residual gauge group is T. In the $N = 1$ language, the theory is a supersymmetric gauged sigma model with target space U. It contains $N = 1$ vector multiplets W^I and chiral multiplets Φ^I, where $1 \leq I \leq r$, $r = \dim T$ being the rank of G. $N = 1$ supersymmetry requires that U is Kähler, with

possible singularities where the effective theory breaks down. $N=2$ supersymmetry requires further that U is special Kähler, that is, there is a flat, torsion-free connection ∇ on TU such that the Kähler form ω is parallel and such that $d_\nabla J=0$, where the complex structure J is viewed as a 1-form valued in TU. See, for example, Freed (1999). Locally, there is a holomorphic prepotential $\mathcal{F}$ and special coordinates $\{z^I\}$. Let $\tilde{z}_I=\partial\mathcal{F}/\partial z^I$ be the dual coordinates and let $\tau_{IJ}=\partial^2\mathcal{F}/\partial z^I\partial z^J=\partial\tilde{z}_I/\partial z^J$. Then $K=\mathrm{Im}(\tilde{z}_I\bar{z}^I)$ is a Kähler potential and $\omega=(\sqrt{-1}/2)\,\mathrm{Im}(\tau_{IJ})\mathrm{d}z^I\wedge\mathrm{d}\bar{z}^J$ is the Kähler form. The effective action is

$$S_{\mathrm{eff}}^{N=2}[W,\Phi]=\frac{1}{4\pi}\mathrm{Im}\left(\int \mathrm{d}^4x\,\mathrm{d}^2\theta\,\tfrac{1}{2}\tau_{IJ}(\Phi)(W^I,W^J)+\int \mathrm{d}^4x\,\mathrm{d}^2\theta\mathrm{d}^2\bar{\theta}K(\Phi)\right)$$

Note that both the coupling constants τ_{IJ} and the metric $\mathrm{Im}\tau_{IJ}$ on U are determined by a holomorphic function $\mathcal{F}$, which is the hallmark of $N=2$ supersymmetry.

In the bare theory with abelian gauge group T, the action is given by choosing $\mathcal{F}_0(\Phi)=(1/2)\tau_{IJ}\langle\Phi^I,\Phi^J\rangle$, where the τ_{IJ} (and hence the metric $\mathrm{Im}\tau_{IJ}$) are constants. Due to one-loop and instanton effects, $\mathcal{F}$ is no longer quadratic in the effective theory. Since τ varies on U, it cannot be holomorphic (except at a few singular points), single valued, and having a positive-definite imaginary part. The solution to this apparent contradiction is that each set of special coordinates and the expression of $\mathcal{F}$ is valid only in part of U. Solving the $N=2$ gauge theory at low energy means understanding the singularity of U in the strong coupling regime and obtaining the explicit form of $\mathcal{F}$ or τ_{IJ} in various regions of the moduli space.

Seiberg–Witten's Low-Energy Solution

We consider $N=2$ gauge theory with $G=\mathrm{SU}(2)$. The Cartan subalgebra is $\mathfrak{t}\cong\mathbb{C}$; each $a\in\mathbb{C}$ determines an element $\phi=(1/2)\left(\begin{smallmatrix}a & 0\\ 0 & -a\end{smallmatrix}\right)$ in $\mathfrak{t}$. The Weyl group $W\cong\mathbb{Z}_2$ acts on $\mathbb{C}$ by $a\mapsto\pm a$. The moduli space of classical vacua is the u-plane $\mathbb{C}/\mathbb{Z}_2$ parametrized by $u=\mathrm{tr}\,\phi^2=(1/2)a^2$. When $u\neq 0$, the gauge group is broken to $\mathrm{U}(1)$. The generator of $\mathbb{Z}_{4h}=\mathbb{Z}_8\subset\mathrm{U}(1)_R$ acts as $a\mapsto\sqrt{-1}a, u\mapsto -u$. The $\mathbb{Z}_8$ symmetry is broken to $\mathbb{Z}_4$; the quotient $\mathbb{Z}_2=\mathbb{Z}_8/\mathbb{Z}_4$ acts on the u-plane by $u\mapsto\pm u$.

Abelian gauge theory and $N=4$ supersymmetric gauge theory exhibit exact electric–magnetic duality in the sense that the quantum theories are identical if the coupling constant τ undergoes an $\mathrm{SL}(2,\mathbb{Z})$ transformation. Seiberg and Witten (1994a, b) proposed that this is so for the low-energy effective theory of the $N=2$ gauge theory. An $\mathrm{SL}(2,\mathbb{Z})$ transformation maps one description of the low-energy theory to another, exchanging electricity and magnetism. It is however not an exact duality of the full SU(2) theory. Rather, duality is in the ambiguity of the choice of the low-energy description. More precisely, τ is a section of a flat $\mathrm{SL}(2,\mathbb{Z})$ bundle over U. Thus, τ is multivalued and exists as a function in local charts only. So we must use different Lagrangians in different regions of the u-plane. Around the singularities where τ is not defined, nontrivial monodromy can appear.

Away from infinity, the electric theory is strongly interacting but the magnetic theory is infrared free. The dual field is $\tilde{a}=\mathrm{d}\mathcal{F}(a)/\mathrm{d}a$, and $\tau_{\mathrm{eff}}(u)=\mathrm{d}\tilde{a}/\mathrm{d}a$. The group $\mathrm{SL}(2,\mathbb{Z})$ is generated by

$$P=\begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix},\qquad S=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}$$
$$T=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$$

To see its action on $\left(\begin{smallmatrix}\tilde{a}\\ a\end{smallmatrix}\right)$, we use the central extension of the $N=2$ super-Poincaré algebra. In the classical theory, the central charge is $Z=(n_{\mathrm{e}}+\tau n_{\mathrm{m}})a$ from the boundary terms at infinity. As the electric–magnetic duality transformation S interchanges n_{e} and n_{m}, we have for any $\gamma\in\mathrm{SL}(2,\mathbb{Z})$, $\gamma:(n_{\mathrm{m}},n_{\mathrm{e}})\mapsto(n_{\mathrm{m}},n_{\mathrm{e}})\gamma^{-1}$. When $n_m=0$, the classical formula $Z=n_e a$ is valid. Invariance of Z under $\mathrm{SL}(2,\mathbb{Z})$ requires that $Z=n_{\mathrm{m}}\tilde{a}+n_{\mathrm{e}}a$ at the quantum level and that $\mathrm{SL}(2,\mathbb{Z})$ acts on $\left(\begin{smallmatrix}\tilde{a}\\ a\end{smallmatrix}\right)$ homogeneously as a column vector.

When $u=(1/2)a^2$ is large, perturbation is reliable. The classical and one-loop results are $a(u)\sim\sqrt{2u}, \tilde{a}\sim(\sqrt{-1}/\pi)a\log a^2$. As u goes around infinity, the fields transforms as $a\mapsto -a, \tilde{a}\mapsto-\tilde{a}+2a$. The monodromy is $M_\infty=PT^{-2}$. The mass M of a monopole state is bounded by $M^2=P^\mu P_\mu\geq|Z|^2$, which is precisely the Bogomol'nyi bound. Now as a consequence of the $N=2$ supersymmetry, it receives no quantum corrections as long as supersymmetry is not broken at the quantum level. The states that saturate the bound are the BPS states. The BPS spectrum at $u\in U$ is a subset of $H_1(E_u,\mathbb{Z})\cong\mathbb{Z}^2$ containing the pairs $(n_{\mathrm{m}},n_{\mathrm{e}})$ realized by the dyon charges. Near infinity, the condition is that either $n_{\mathrm{e}}=\pm1, n_{\mathrm{m}}=0$ (for $W_\pm$ particles) or $n_{\mathrm{m}}=\pm1$ (for monopoles or dyons). This spectrum is invariant under the monodromy M_∞.

The nontrivial holonomy at infinity implies the existence of at least one singularity at a finite value $u=u_0$, where extra particles become massless. Seiberg and Witten (1994a, b) propose that these

particles are collective excitations in the perturbative regime. Suppose along a path connecting u_0 and some base point near infinity, a monopole of charges $(\pm 1, n_e) = (0,1)(T^{\mp n_e} S^{\pm 1})^{-1}$ becomes massless at u_0. Then by the renormalization group analysis and duality, the monodromy at u_0 is $M_{u_0} = (T^{\mp n_e} S^{\pm 1}) T^2 (T^{\mp n_e} S^{\pm 1})^{-1}$. It turns out that there are two singularities $u = \pm\Lambda^2$ with monodromies $M_{\Lambda^2} = ST^2S^{-1}$ and $M_{-\Lambda^2} = (TS)T^2(TS)^{-1}$. The particles that become massless at $\pm\Lambda^2$ are of charges $(n_m, n_e) = (1,0)$ and $(1,-1)$, respectively. The only BPS states in the strong coupling regime are those which become massless at the singularities; the others decay as u deforms towards strong interaction.

The monodromies $M_{\pm\Lambda^2}, M_\infty$ (or any two of them) generate the subgroup $\Gamma(2)$. The family of elliptic curves with these monodromies can be identified with $y^2 = (x - \Lambda^2)(x + \Lambda^2)(x - u)$ called the Seiberg–Witten curve. The singularities are at $u = \pm\Lambda^2$ and $u = \infty$, where the curve degenerates. Let

$$\lambda = \frac{\sqrt{2}}{2\pi} \frac{y\,dx}{x^2 - \Lambda^4}$$

be the Seiberg–Witten differential (of second kind on the total space E). Then in a suitable basis (α, β) of $H_1(E_u/U, \mathbb{Z})$, we have $a = \int_\alpha \lambda, \tilde{a} = \int_\beta \lambda$. At a singularity, if $\nu = n_m\beta + n_e\alpha$ is a vanishing cycle, then the dyon of charges (n_m, n_e) becomes massless. This is because its central charge is $Z = n_m\tilde{a} + n_e a = \int_\nu \lambda$. The monodromy at a singularity where ν is a vanishing cycle is given by the Picard–Lefshetz formula $M: \gamma \mapsto \gamma - 2(\gamma \cdot \nu)\nu$. At $u = \pm\Lambda^2$, the vanishing cycles are β and $\beta - \alpha$, respectively.

We return to the $N = 1$ $SO(N_c)$ gauge theory with $N_f = N_c - 2$. At a generic point in the moduli space, the gauge group is broken to $SO(2)$, which is abelian. Much of the above discussion applies to this case. By $N = 1$ supersymmetry, the effective coupling τ_{eff} is holomorphic in M but is not single valued. In fact, τ_{eff} depends on $u = \det M$, which is invariant under the (anomaly free) $SU(N_f)$ symmetry. For large u, we have $e^{2\pi\sqrt{-1}\tau_{\text{eff}}} = \Lambda^{4N_c - 8}/u^2$ and the monodromy around infinity is $M_\infty = PT^{-2}$. On the other hand, a large expectation value of M of rank $N_c - 3$ breaks the gauge group to $SO(3)$ and the theory is the $N = 2$ theory discussed earlier. Using these facts, Intriligator and Seiberg (1995) identified the family of elliptic curves as $y^2 = x(x - 16\Lambda^{2N_c - 4})(x - u)$. There are two singularities with inequivalent physics. At $u = 0$, the monodromy is ST^2S^{-1}. A pair of monopoles $\tilde{Q}^\pm$ becomes massless. They couple with M through the superpotential $W \sim (2\mu)^{-1} M^{ij} \tilde{Q}_i \tilde{Q}_j$. At $u = 16\Lambda^{2N_c - 4}$, the monodromy is $(T^2S)T^2(T^2S)^{-1}$. A pair of dyons $E^\pm$ of charges ± 1 become massless. The effective action is $W_{\text{eff}} \sim (u - 16\Lambda^{2N_c - 4})E^+E^-$.

Topological gauge theory is a twisted version of $N = 2$ Yang–Mills theory in which the observables at high energy are the Donaldson invariants. The work of Seiberg and Witten (1994a, b) yields new insight to it and has a tremendous impact on the geometry of 4-manifolds. See Witten (1994) for the initial steps.

After the work of Seiberg and Witten (1994a, b), there has been much progress on theories with other gauge groups. If the gauge group is a compact Lie group of rank r, the u-plane is replaced by $\mathfrak{t}^{\mathbb{C}}/W$; the singularities are modified by quantum effects. The duality group is $Sp(2r, \mathbb{Z})$ or its subgroup of finite index, acting on the coupling matrix $\tau = (\tau_{IJ})$ by fractional linear transformations. For example, for $G = SU(N_c)$, the moduli space is parametrized by gauge invariants $u_2, \ldots, u_{N_c}$ defined by $\det(xI - \phi) = x^{N_c} - \sum_{i=2}^{N_c} u_i x^{N_c - i} = P_{N_c}(x, u_i)$. Classically, the singular locus is a simple singularity of type $A_{N_c - 1}$. At the quantum level, the singularity consists of two copies of such locus shifted by $\pm\Lambda^n$ in the u_n direction. The monodromies correspond to a family of hyperelliptic curves $y^2 = P_{N_c}(x, u_i)^2 - \Lambda^{2N_c}$ of genus $N_c - 1$. The Seiberg–Witten differential is

$$\lambda = \frac{\sqrt{2}}{\pi\sqrt{-1}} \frac{\partial P_{N_c}(x, u_i)}{\partial x} \frac{x\,dx}{y} + \partial(\cdots)$$

The $N_c - 1$ independent eigenvalues a^i of ϕ and their duals $\tilde{a}_i = \partial\mathcal{F}/\partial a^i$ are the periods of λ along the $2N_c - 2$ homology cycles in the curve. For more details, the reader is referred to Klemm *et al.* (1995) and Argyes and Faraggi (1995).

$N = 2$ QCD

$N = 2$ supersymmetric QCD is $N = 2$ Yang–Mills theory coupled to $N = 2$ matter. The latter consists of $N = 1$ superfields Q that form a quarternionic representation R of the gauge group G. The space R has a G-invariant hyper-Kähler structure. The hyper-Kähler moment map $\mu_H: R \to \mathfrak{g}^* \otimes \operatorname{Im}\mathbb{H}$ consists of a real moment map $\mu_R: R \to \mathfrak{g}^*$ for the Kähler structure and a complex moment map $\mu_{\mathbb{C}}: R \to (\mathfrak{g}^*)^{\mathbb{C}}$ for the holomorphic symplectic structure. As an $N = 1$ theory, the matter superfields are valued in $R \times \mathfrak{g}^{\mathbb{C}}$ with a D-term $D(Q, \Phi) = \mu_R(Q) + (1/2\sqrt{-1})[\Phi, \bar{\Phi}]$ and a superpotential $W(Q, \Phi) = \sqrt{2}\langle \mu_{\mathbb{C}}(Q), \Phi \rangle + m(Q)$, where the mass term m is a G-invariant quadratic form on R. The classical moduli space of vacua has two branches. On the Coulomb branch where $Q = 0$ and $\Phi \neq 0$, the unbroken gauge group is abelian and the

photons are massless. If $Q \neq 0$ exists in the flat directions, the gauge group is broken according to the value of Q; these are the Higgs branches. If $m=0$, the moduli space of classical vacua is the hyper-Kähler quotient $\mu_{\mathrm{H}}^{-1}(0)/G$. The branches of two types touch at the origin, where the full gauge group is restored, and at other subvarieties in R. The global symmetry is the subgroup of $\mathrm{U}(R)$ that commutes with the G-action on R and preserves m; it contains $\mathrm{U}(2)_R$.

Quantum mechanically, such a theory is free from local gauge anomalies. Consistency under large gauge transformations puts a torsion condition on R, such as $\nu(R)=0(\mathrm{mod}\,2)$. Since $b_0=2\check{h}-(1/2)\nu(R)$, the theory is asymptotically free if $\nu(R)<4\check{h}$. If $\nu(R)=4\check{h}$, the quantum theory is scale invariant up to one-loop (and hence to all loops), and is expected to be so nonperturbatively. If $\nu(R)>4\check{h}$, the quantum theory may not be defined but it can be the low-energy solution of another asymptotically free theory. Due to the axial anomaly, the $\mathrm{U}(2)_R$ global symmetry reduces to the subgroup $\mathrm{SU}(2)_R\times_{\mathbb{Z}_2}\mathbb{Z}_{4\check{h}-\nu(R)}$. The metric on the Coulomb branch can be corrected by quantum effects, but those on the Higgs branches do not change because of the uniqueness of the hyper-Kähler metric. In the quantum theory, the Higgs branches still touch the Coulomb branch, but the photons of the Coulomb branch are the only massless gauge bosons at the point where they meet.

When $G=\mathrm{SU}(N_c)$ we take N_f quarks $Q^i(i=1,\ldots,N_f)$ in the fundamental representation and N_f antiquarks $\tilde{Q}_i(i=1,\ldots,N_f)$ in the complex-conjugate representation. The moment map is the same as in $N=1$ QCD whereas the superpotential is $W=\sqrt{2}\tilde{Q}_i\Phi Q^i+\sum_i m_i\tilde{Q}_iQ^i$. Consider the case $G=\mathrm{SU}(2)$ as in Seiberg and Witten (1994b). Since $b_0=4-N_f$, the asymptotically free theories have $N_f\leq 3$ whereas the $N_f=4$ theory is scale invariant. As the representations on Q^i and $\tilde{Q}_i$ are isomorphic, the classical global symmetry is $\mathrm{O}(2N_f)\times\mathrm{U}(2)_R$ when all $m_i=0$. The appearance of the even number of fundamental representations is necessary for the consistency of the theory at the quantum level. The $\mathrm{U}(1)_R$ symmetry is anomalous if $N_f\neq 4$. When $N_f>0$, $\mathrm{SO}(2N_f)$ is anomaly free, whereas $\mathrm{O}(2N_f)/\mathrm{SO}(2N_f)=\mathbb{Z}_2$ is anomalous. The anomaly free subgroup of $\mathbb{Z}_2\times\mathrm{U}(1)_R$ is $\mathbb{Z}_{4(4-N_f)}$. Its $\mathbb{Z}_2$ subgroup acts in the same way as $\mathbb{Z}_2\subset Z(\mathrm{SO}(2N_f))$. A nonzero expectation value of $u=\mathrm{tr}\,\phi^2$ further breaks the symmetry to $\mathbb{Z}_4$. The quotient group that acts effectively on the u-plane (the Coulomb branch) is $\mathbb{Z}_{4-N_f}$ if $N_f>0$ and $\mathbb{Z}_2$ if $N_f=0$. When $N_f=4$, the $\mathrm{U}(1)_R$ symmetry is anomaly free but $\mathbb{Z}_2=\mathrm{O}(8)/\mathrm{SO}(8)$ is still anomalous.

The $N_f=0$ theory is the $N=2$ pure gauge theory. In order to compare it to the $N_f>0$ theories, we multiply n_e by 2 so that it has integer values on Q^i and $\tilde{Q}_i$, and divide a by 2 to preserve the formula $Z=n_m\tilde{a}+n_e a$. The monodromies around the singularities become $M_{\Lambda^2}=STS^{-1}$, $M_{-\Lambda^2}=(T^2S)T(T^2S)^{-1}$, $M_\infty=PT^{-4}$. They generate the subgroup $\Gamma_0(4)$ of $\mathrm{SL}(2,\mathbb{Z})$. The coupling constant is

$$\tau=\frac{\theta}{\pi}+\frac{8\pi\sqrt{-1}}{g^2}$$

The Seiberg–Witten curve is $y^2=x^3-ux^2+(1/4)\Lambda_0^4x$, related to the earlier one $y^2=(x-u)(x^2-\Lambda_0^4)$ by an isogeny. Here and below, Λ_{N_f} is the dynamically generated scale.

For $N_f>0$, we consider the case with zero bare masses. The simplest BPS-saturated states are the elementary quarks with mass $\sqrt{2}|a|$, which form the vector representation of $\mathrm{SO}(2N_f)$. In addition, the quarks have fermion zero modes in the monopole background. When $n_m=1$, each SU(2) doublet of quarks has one zero mode. With N_f hypermultiplet, there are $2N_f$ zero modes in the vector representation of $\mathrm{SO}(2N_f)$. Upon quantization, the quantum states are in the spinor representation. So the flavor symmetry is really $\mathrm{Spin}(2N_f)$. The spectrum may also include states with $n_m>1$. For $N_f=2,3,4$, the center $Z(\mathrm{Spin}(2N_f))$ are $\mathbb{Z}_2\times\mathbb{Z}_2$, $\mathbb{Z}_4$, $\mathbb{Z}_2\times\mathbb{Z}_2$, whose generators act on states of charges (n_m,n_e) by $((-1)^{n_e+n_m},(-1)^{n_e})$, $\sqrt{-1}^{n_m+2n_e}$, $((-1)^{n_m},(-1)^{n_e})$, respectively.

Suppose at a singularity on the u-plane, the low-energy theory is QED with k hypermultiplets. Let m_i be the bare mass and S_i, the U(1) charge of the ith hypermultiplet. With the expectation value of ϕ, the actual masses are $|\sqrt{2}a+m_i|(1\leq i\leq k)$. As the states form a small representation of the $N=2$ algebra, the central charge is modified as $Z=n_m\tilde{a}+n_e a+S\cdot m/\sqrt{2}$, where $m=(m_1,\ldots,m_k)$ and $S=(S_1,\ldots,S_k)$. Under a duality transformation $M\in\mathrm{SL}(2,\mathbb{Z})$, the column vector $(m/\sqrt{2},\tilde{a},a)$ is multiplied by a matrix of the form $\hat{M}=\begin{pmatrix}1_k & 0\\ * & M\end{pmatrix}$. (For example, if $M=T$, $\hat{M}$ can be derived by one-loop analysis.) So the row vector $W=(S,n_m,n_e)$ transforms as $W\mapsto W\hat{M}^{-1}$. The transformation on (n_m,n_e) is not homogeneous when there are hypermultiplets. This phenomenon persists even when all the bare masses m_i are zero.

When $N_f=1$, the global symmetry of the u-plane is $\mathbb{Z}_3$. There are three singularities related by this symmetry, where monopoles with charges $(n_m,n_e)=(1,0),(1,1)$, and $(1,2)$ become massless. The low-energy theory at each singularity is QED with a single light hypermultiplet. Besides the photon, no other flat directions exist. This is consistent with the absence of Higgs branch in the original theory. The monodromies at the singularities are STS^{-1},

$(TS)T(TS)^{-1}, (T^2S)T(T^2S)^{-1}$, respectively, and the corresponding Seiberg–Witten family of curves is $y^2 = x^2(x-u) - (1/64)\Lambda_1^6$. The Seiberg–Witten differential is

$$\lambda = -\frac{\sqrt{2}}{4\pi}\frac{y\,dx}{x^2}$$

When $N_f = 2$, there are two singularities related by the global symmetry $\mathbb{Z}_2$ of the u-plane. The massless states at one singularity have $(n_m, n_e) = (1, 0)$ and form a spinor representation of SO(4) while those at the other have $(n_m, n_e) = (1, 1)$ and form the other spinor representation. The low-energy theory at each singularity is QED with two light hypermultiplets. There are additional flat directions along which $SO(4) \times SU(2)_R$ is broken. They form the two Higgs branches that touch the u-plane at the two singularities rather than at the origin. The metric and pattern of symmetry breaking are the same as classically. The monodromies are $ST^2S^{-1}, (TS)T^2(TS)^{-1}$. The Seiberg–Witten curve is $y^2 = (x^2 - u) - (1/64)\Lambda_2^4)(x-u)$ and the differential is

$$\lambda = -\frac{\sqrt{2}}{4\pi}\frac{y\,dx}{x^2 - \Lambda_2^4/64}$$

When $N_f = 3$, the u-plane has no global symmetry. There are two singularities. At one of them, a single monopole bound state with $(n_m, n_e) = (2, 1)$ becomes massless and there are no other light particles. At the other singularity, the massless states have $(n_m, n_e) = (1, 0)$ and form a (four-dimensional) spinor representation of SO(6) with a definite chirality. Thus, the low-energy theory is QED with four light hypermultiplets. Along the flat directions, the $SO(6) \times SU(2)_R$ symmetry is further broken. This corresponds to a single Higgs branch touching the u-plane at the singularity. Again, the metric on the Higgs branch is not modified by quantum effects. The monodromies at the two singularities are $(ST^2S)T(ST^2S)^{-1}$ and ST^4S^{-1}, respectively. The Seiberg–Witten curve is $y^2 = x^2(x-u) - (1/64)\Lambda_3^2(x-u)^2$ and the differential is

$$\lambda = \frac{\sqrt{2}}{\pi\Lambda_3}\log\left(y + \sqrt{-1}\frac{\Lambda_3}{8}\left(x - u - \frac{32}{\Lambda_3^2}x^2\right)\right)dx$$

When $N_f = 4$, the theory is characterized by classical coupling constant τ, and there are no corrections to $a = (1/2)\sqrt{2u}, \tilde{a} = \tau a$. There is only one singularity at $u = 0$, where the monodromy is P. Seiberg and Witten (1994b) postulate that the full quantum theory is $SL(2, \mathbb{Z})$ invariant, just like the $N = 4$ pure gauge theory. The elementary hypermultiplet has $(n_m, n_e) = (0, 1)$ and form the vector representation v of SO(8). Fermion zero modes give rise to hypermultiplets with $(n_m, n_e) = (1, 0), (1, 1)$ that transform under the spinor representations s, c of Spin(8). $SL(2, \mathbb{Z})$ acts on the spectrum via a homomorphism onto the outer-automorphism group S_3 of Spin(8), which then permutes v, s, and c. So duality is mixed in an interesting way with the SO(8) triality. In v, s, and c, the center $\mathbb{Z}_2 \times \mathbb{Z}_2$ acts as $((-1)^{n_m}, (-1)^{n_e}) = (1, -1), (-1, 1), (-1, -1)$, respectively. The full $SL(2, \mathbb{Z})$ invariance predicts the existence of multimonopole bound states: for every pair of relatively prime integers (p, q), there are eight states with $(n_m, n_e) = (p, q)$ that form a representation of Spin(8) on which the center acts as $((-1)^p, (-1)^q)$.

Solutions when the bare masses are nonzero are also obtained by Seiberg and Witten (1994b). The masses can be deformed to relate theories with different values of N_f. $N = 2$ QCD with a general classical gauge group has also been studied. By adding to these theories a mass term $m \operatorname{tr} \Phi^2$ that explicitly breaks the supersymmetry to $N = 1$, the dualities of Seiberg can be recovered. For $SU(N_c), SO(N_c)$ and $Sp(2N_c)$ gauge groups, see Hanany and Oz (1995), Argyes *et al.* (1996), Argyes *et al.* (1997) and references therein.

See also: Anomalies; Brane Construction of Gauge Theories; Donaldson–Witten Theory; Duality in Topological Quantum Field Theory; Effective Field Theories; Electric–Magnetic Duality; Floer Homology; Gauge Theories from Strings; Gauge Theory: Mathematical Applications; Nonperturbative and Topological Aspects of Gauge Theory; Quantum Chromodynamics; Topological Quantum Field Theory: Overview; Supersymmetric Particle Models.

Further Reading

Affleck I, Dine M, and Seiberg N (1984) Dynamical supersymmetric breaking in supersymmetric QCD. *Nuclear Physics B* 241: 493–534.

Argyes PC and Faraggi AE (1995) Vacuum structure and spectrum of $N = 2$ supersymmetric $SU(n)$ gauge theory. *Physical Review Letters* 74: 3931–3934.

Argyes PC, Plesser MR, and Seiberg N (1996) The moduli space of vacua of $N = 2$ SUSY QCD and duality in $N = 1$ SUSY QCD. *Nuclear Physics B* 471: 159–194.

Argyes PC, Plesser MR, and Shapere AD (1997) $N = 2$ moduli spaces and $N = 1$ dualities for $SO(n_c)$ and $USp(2n_c)$ super-QCD. *Nuclear Physics B* 483: 172–186.

Freed DS (1999) Special Kähler manifolds. *Communications in Mathematical Physics* 203: 31–52.

Hanany A and Oz Y (1995) On the quantum moduli space of vacua of $N = 2$ supersymmetric $SU(N_c)$ gauge theories. *Nuclear Physics B* 452: 283–312.

Klemm A, Lerche W, Yankielowicz S, and Theisen S (1995) Simple singularities and $N = 2$ supersymmetric Yang–Mills theory. *Physics Letters B* 344: 169–175.

Intriligator K and Pouliot P (1995) Exact superpotential, quantum vacua and duality in supersymmetric SP(N_c) gauge theories. *Physics Letters B* 353: 471–476.

Intriligator K and Seiberg N (1995) Duality, monopoles, dyons, confinement and oblique confinement in supersymmetric SO(N_c) gauge theories. *Nuclear Physics B* 444: 125–160.

Seiberg N (1994) Exact results on the space of vacua of four-dimensional SUSY gauge theories. *Physical Review D* 49: 6857–6863.

Seiberg N (1995) Electric–magnetic duality in supersymmetric non-Abelian gauge theories. *Nuclear Physics B* 435: 129–146.

Seiberg N and Witten E (1994a) Electric–magnetic duality, monopole condensation, and confinement in $N=2$ supersymmetric Yang–Mills theory. *Nuclear Physics B* 426: 19–52.

Seiberg N and Witten E (1994b) Monopoles, duality and chiral symmetry breaking in $N=2$ supersymmetric QCD. *Nuclear Physics B* 431: 484–550.

Witten E (1994) Monopoles and four-manifolds. *Mathematical Research Letters* 1: 769–796.

Semiclassical Approximation *see* Stationary Phase Approximation; Normal Forms and Semiclassical Approximation



Standard Model of Particle Physics

G Altarelli, CERN, Geneva, Switzerland

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The standard model (SM) is a consistent, finite, and – within the limitations of our present technical ability – computable theory of fundamental microscopic interactions that successfully explains most of the known phenomena in elementary particle physics. The SM describes strong, electromagnetic, and weak interactions. All microscopic phenomena observed to date can be attributed to one or the other of these interactions. For example, the forces that hold together the protons and the neutrons in the atomic nuclei are due to strong interactions; the binding of electrons to nuclei in atoms or of atoms in molecules is caused by electromagnetism; and the energy production in the Sun and the other stars occurs through nuclear reactions induced by weak interactions. In principle, gravitational forces should also be included in the list of fundamental interactions but their impact on fundamental particle processes at accessible energies is totally negligible.

The structure of the SM is a generalization of that of quantum electrodynamics (QED), in the sense that it is a renormalizable field theory based on a local symmetry (i.e., separately valid at each spacetime point x) that extends the gauge invariance of electrodynamics to a larger set of

Notes:

conserved currents and charges. There are eight strong charges, called "color" charges and four electroweak charges (which, in particular, include the electric charge). The commutators of these charges form the $SU(3) \otimes SU(2) \otimes U(1)$ algebra. In QED, the interaction between two matter particles with electric charges (e.g., two electrons) is mediated by the exchange of one (or more) photons emitted by one electron and reabsorbed by the second. In the SM the matter fields, all of spin 1/2, are the quarks, the constituents of protons, neutrons, and all hadrons, endowed with both color and electroweak charges, and the leptons (the electron e^-, the muon μ^-, the tauon τ^-, plus the three associated neutrinos ν_e, ν_μ, and ν_τ) with no color but with electroweak charges. The matter fermions come in three generations or families with identical quantum numbers but different masses. The pattern is as follows:

$$\begin{bmatrix} u & u & u & \nu_e \\ d & d & d & e \end{bmatrix}, \quad \begin{bmatrix} c & c & c & \nu_\mu \\ s & s & s & \mu \end{bmatrix}, \quad \begin{bmatrix} t & t & t & \nu_\tau \\ b & b & b & \tau \end{bmatrix} \qquad [1]$$

Each family contains a weakly charged doublet of quarks, in three color replicas, and a colorless weakly charged doublet with a neutrino and a charged lepton. At present, there is no explanation for this triple repetition of fermion families. The force carriers, of spin 1, are the photon γ, the weak interaction gauge bosons W^+, W^-, and Z_0 and the eight gluons g that mediate the strong interactions. The photon and the gluons have zero masses as a consequence of the exact conservation of the corresponding symmetry generators, the electric charge and the eight color charges. The weak bosons W^+, W^-, and Z_0 have large masses ($m_W \sim 80.4\,\text{GeV}$, $m_Z = 91.2\,\text{GeV}$), signaling that the corresponding symmetries are badly broken. In the SM, the spontaneous breaking of the electroweak gauge symmetry is induced by the Higgs mechanism, which predicts the presence of one (or more) spin 0 particles in the physical spectrum, the Higgs boson(s), not yet experimentally observed. A tremendous experimental effort is underway or planned to reveal the Higgs sector as the last crucial missing link in the SM verification.

Quantum Chromodynamics

The statement that quantum chromodynamics (QCD) is a renormalizable gauge theory based on the group SU(3) with color triplet quark matter fields fixes the QCD Lagrangian density to be

$$\mathcal{L} = -\frac{1}{4}\sum_{A=1}^{8} F^{A\mu\nu}F^{A}_{\mu\nu} + \sum_{j=1}^{n_f} \bar{q}_j(\mathrm{i}\not{D} - m_j)q_j \qquad [2]$$

Here q_j are the quark fields (of n_f different flavors) with mass m_j; $\not{D} = \mathrm{D}_\mu\gamma^\mu$, where γ^μ are the Dirac matrices and D_μ is the covariant derivative

$$D_\mu = \partial_\mu - \mathrm{i}e_s\sum_A t^A g^A_\mu \qquad [3]$$

e_s is the gauge coupling (in analogy with QED,

$$\alpha_s = \frac{e_s^2}{4\pi} \qquad [4]$$

here and throughout this article natural units, $\hbar = c = 1$, are used); g^A_μ, $A = 1, \ldots, 8$, are the gluon fields, and t^A are the SU(3) group generators in the triplet representation of quarks (i.e., t_A are 3×3 matrices acting on q); the generators obey the commutation relations $[t^A, t^B] = \mathrm{i}C_{ABC}t^C$, where C_{ABC} are the complete antisymmetric structure constants of SU(3) (the normalization of C_{ABC} and of e_s is specified by $\mathrm{tr}[t^A t^B] = 1/2\delta^{AB}$);

$$F^A_{\mu\nu} = \partial_\mu g^A_\nu - \partial_\nu g^A_\mu - e_s C_{ABC} g^B_\mu g^C_\nu \qquad [5]$$

The physical vertices in QCD include the gluon–quark–antiquark vertex, analogous to the QED photon–fermion–antifermion coupling, but also the three-gluon and four-gluon vertices, of order e_s and e_s^2, respectively, which have no analog in an abelian theory like QED. In QED, the photon (a neutral particle) is coupled to all electrically charged particles. In QCD, the gluons are colored, hence self-coupled. This is reflected in the fact that in QED $F_{\mu\nu}$ is linear in the gauge field, so that the term $F^2_{\mu\nu}$ in the Lagrangian is a pure kinetic term, while in QCD $F^A_{\mu\nu}$ is quadratic in the gauge field, so that in $F^{A2}_{\mu\nu}$ we find cubic and quartic vertices beyond the kinetic term.

The QCD Lagrangian in eqn [2] has a simple structure but a very rich dynamical content, including the observed complex spectroscopy with a large number of hadrons. The most prominent properties of QCD are asymptotic freedom and confinement. In field theory, the effective coupling of a given interaction vertex is modified by the interaction. As a result, the measured intensity of the force depends on the transferred (four)momentum squared, Q^2, among the participants. In QCD, the relevant coupling parameter that appears in physical processes is α_s (see eqn [4]). Asymptotic freedom means that the effective coupling becomes a function of

Q^2: $\alpha_s(Q^2)$ decreases for increasing Q^2 and vanishes asymptotically. Thus, the QCD interaction becomes very weak in processes with large Q^2, called hard processes or deep inelastic processes (i.e., with a final-state distribution of momenta and a particle content very different from that in the initial state). One can prove that in four spacetime dimensions all gauge theories based on a noncommuting group of symmetry are asymptotically free, and conversely. The effective coupling decreases very slowly at large momenta with the inverse logarithm of Q^2: $\alpha_s(Q^2) = 1/b \log Q^2/\Lambda^2$, where b is a known constant and Λ is an energy of the order of a few hundred MeV. Since in quantum mechanics large momenta imply short wavelengths, the result is that at short distances the potential between two color charges is similar to the Coulomb potential, that is, proportional to $\alpha_s(r)/r$, with an effective color charge which is small at short distances. On the contrary the interaction strength becomes large at large distances or small transferred momenta, of order $Q \lesssim \Lambda$. In fact, the observed hadrons are tightly bound composite states of quarks, with compensating color charges so that they are overall neutral in color. The property of confinement is the impossibility of separating color charges, like individual quarks and gluons. This is because in QCD the interaction potential between color charges increases, at long distances, linearly in r. When we try to separate the quark and the antiquark that form a color-neutral meson the interaction energy grows until pairs of quarks and antiquarks are created from the vacuum and new neutral mesons are coalesced instead of free quarks. For example, consider the process $e^+e^- \to q\bar{q}$ at large center-of-mass energies. The final-state quark and antiquark have large energies, so they separate in opposite directions very fast. But the color-confinement forces create new pairs in between them. Two back-to-back jets of colorless hadrons are observed with a number of slow pions that make the exact separation of the two jets impossible. In some cases, a third well-separated jet of hadrons is also observed: these events correspond to the radiation of an energetic gluon from the parent quark–antiquark pair.

Electroweak Interactions

We split the electroweak Lagrangian into two parts by separating the Higgs boson couplings:

$$\mathcal{L} = \mathcal{L}_{\text{symm}} + \mathcal{L}_{\text{Higgs}} \qquad [6]$$

We start by specifying $\mathcal{L}_{\text{symm}}$, which involves only gauge bosons and fermions (a sum over all flavors of quarks and leptons, generally indicated by ψ is understood):

$$\mathcal{L}_{\text{symm}} = -\frac{1}{4}\sum_{A=1}^{3} F^A_{\mu\nu}F^{A\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} + \bar{\psi}_L i\gamma^\mu D_\mu \psi_L + \bar{\psi}_R i\gamma^\mu D_\mu \psi_R \qquad [7]$$

This is the Yang–Mills Lagrangian for the gauge group SU(2) ⊗ U(1) with fermion matter fields. Here

$$\begin{aligned} B_{\mu\nu} &= \partial_\mu B_\nu - \partial_\nu B_\mu \\ F^A_{\mu\nu} &= \partial_\mu W^A_\nu - \partial_\nu W^A_\mu - g\epsilon_{ABC} W^B_\mu W^C_\nu \end{aligned} \qquad [8]$$

are the gauge antisymmetric tensors constructed out of the gauge field B_μ associated with U(1), and W^A_μ corresponding to the three SU(2) generators; ϵ_{ABC} are the group structure constants (see eqn [11]), which, for SU(2), coincide with the totally antisymmetric Levi-Civita tensor (recall the familiar angular-momentum commutators).

The fermion fields are described through their left- and right-hand components:

$$\psi_{L,R} = [(1 \mp \gamma_5)/2]\psi, \qquad \bar{\psi}_{L,R} = \bar{\psi}[(1 \pm \gamma_5)/2] \qquad [9]$$

Note that, as given in eqn [9],

$$\begin{aligned} \bar{\psi}_L &= \psi_L^\dagger \gamma_0 = \psi^\dagger[(1-\gamma_5)/2]\gamma_0 \\ &= \bar{\psi}[\gamma_0(1-\gamma_5)/2]\gamma_0 = \bar{\psi}[(1+\gamma_5)/2] \end{aligned}$$

The matrices $P_\pm = (1 \pm \gamma_5)/2$ are projectors. They satisfy the relations $P_\pm P_\pm = P_\pm, P_\pm P_\mp = 0$, $P_+ + P_- = 1$.

The standard electroweak theory is a chiral theory, in the sense that ψ_L and ψ_R behave differently under the gauge group. In particular, all ψ_R are singlets and all ψ_L are doublets in the minimal SM (MSM). Thus, mass terms for fermions (of the form $\bar{\psi}_L\psi_R + \text{h.c.}$) are forbidden in the symmetric limit. Fermion masses are introduced, together with $W^\pm$ and Z masses, by the mechanism of symmetry breaking. The covariant derivatives $D_\mu\psi_{L,R}$ are explicitly given by

$$D_\mu\psi_{L,R} = \left[\partial_\mu + ig\sum_{A=1}^{3} t^A_{L,R} W^A_\mu + ig'\frac{1}{2}Y_{L,R}B_\mu\right]\psi_{L,R} \qquad [10]$$

where $t^A_{L,R}$ and $1/2Y_{L,R}$ are the SU(2) and U(1) generators, respectively, in the reducible representations $\psi_{L,R}$. The commutation relations of the SU(2) generators are given by

$$[t^A_L, t^B_L] = i\epsilon_{ABC}t^C_L \quad \text{and} \quad [t^A_R, t^B_R] = i\epsilon_{ABC}t^C_R \qquad [11]$$

We use the normalization $\text{tr}[t^A t^B] = 1/2\delta^{AB}$ in the fundamental representation of SU(2). The electric

charge generator Q (in units of e, the positron charge) is given by

$$Q = t_L^3 + 1/2 Y_L = t_R^3 + 1/2 Y_R \qquad [12]$$

All fermion couplings to the gauge bosons can be derived directly from eqns [7] and [10]. The charged-current (CC) couplings are the simplest. From

$$\begin{aligned} g\left(t^1 W_\mu^1 + t^2 W_\mu^2\right) &= g\left\{\left[\left(t^1 + it^2\right)/\sqrt{2}\right]\right. \\ &\quad \left.\times\left[\left(W_\mu^1 - iW_\mu^2\right)/\sqrt{2}\right] + \text{h.c.}\right\} \\ &= g\left\{\left[\left(t^+ W_\mu^-\right)/\sqrt{2}\right] + \text{h.c.}\right\} \end{aligned} \qquad [13]$$

where $t^\pm = t^1 \pm it^2$ and $W^\pm = (W^1 \pm iW^2)/\sqrt{2}$, we obtain the vertex

$$\begin{aligned} V_{\bar{\psi}\psi W} = g\bar{\psi}\gamma_\mu\Big[&\left(t_L^+/\sqrt{2}\right)(1-\gamma_5)/2 + \left(t_R^+/\sqrt{2}\right) \\ &\times (1+\gamma_5)/2\Big]\psi W_\mu^- + \text{h.c.} \end{aligned} \qquad [14]$$

In the neutral-current (NC) sector, the photon A_μ and the mediator Z_μ of the weak NC are orthogonal and normalized linear combinations of B_μ and W_μ^3:

$$\begin{aligned} A_\mu &= \cos\theta_W B_\mu + \sin\theta_W W_\mu^3 \\ Z_\mu &= -\sin\theta_W B_\mu + \cos\theta_W W_\mu^3 \end{aligned} \qquad [15]$$

Equations [15] define the weak mixing angle θ_W. The photon is characterized by equal couplings to left and right fermions with a strength equal to the electric charge. Recalling eqn [12] for the charge matrix Q, we immediately obtain

$$g\sin\theta_W = g'\cos\theta_W = e \qquad [16]$$

or, equivalently,

$$\tan\theta_W = g'/g \qquad [17]$$

Once θ_W has been fixed by the photon couplings, it is a simple matter of algebra to derive the Z couplings, with the result

$$\begin{aligned} \Gamma_{\bar{\psi}\psi Z} = g/(2\cos\theta_W)\bar{\psi}\gamma_\mu\big[&t_L^3(1-\gamma_5) + t_R^3(1+\gamma_5) \\ &-2Q\sin^2\theta_W\big]\psi Z^\mu \end{aligned} \qquad [18]$$

where $\Gamma_{\bar{\psi}\psi Z}$ is a notation for the vertex. In the MSM, $t_R^3 = 0$ and $t_L^3 = \pm 1/2$. Note that the CC and NC weak couplings do not conserve P (parity) and C (charge conjugation).

In order to derive the effective four-fermion interactions that are equivalent, at low energies, to the CC and NC couplings given in eqns [14] and [18], we anticipate that large masses, as experimentally observed, are provided for $W^\pm$ and Z by $\mathcal{L}_{\text{Higgs}}$. For left–left CC couplings, when the momentum transfer squared can be neglected with respect to m_W^2 in the propagator of Born diagrams with single W exchange, from eqn [14], we can write

$$\begin{aligned} \mathcal{L}_{\text{eff}}^{\text{CC}} \simeq &\left(g^2/8m_W^2\right)\left[\bar{\psi}\gamma_\mu(1-\gamma_5)t_L^+\psi\right] \\ &\times\left[\bar{\psi}\gamma^\mu(1-\gamma_5)t_L^-\psi\right] \end{aligned} \qquad [19]$$

By specializing further in the case of doublet fields such as $\nu_e - e^-$ or $\nu_\mu - \mu^-$, we obtain the tree-level relation of g with the Fermi coupling constant G_F measured from μ decay ($G_F = 1.16639(2)\times 10^{-5}\,\text{GeV}^{-2}$):

$$G_F/\sqrt{2} = g^2/8m_W^2 \qquad [20]$$

By recalling that $g\sin\theta_W = e$, we can also cast this relation in the form

$$m_W = \mu_{\text{Born}}/\sin\theta_W \qquad [21]$$

with

$$\mu_{\text{Born}} = \left(\pi\alpha/\sqrt{2}G_F\right)^{1/2} \simeq 37.2802\,\text{GeV} \qquad [22]$$

where α is the fine-structure constant of QED ($\alpha \equiv e^2/4\pi = 1/137.036$).

In the same way, for neutral currents we obtain, in Born approximation, from eqn [18], the effective four-fermion interaction given by

$$\mathcal{L}_{\text{eff}}^{\text{NC}} \simeq \sqrt{2}G_F\rho_0\bar{\psi}\gamma_\mu[\ldots]\psi\bar{\psi}\gamma^\mu[\ldots]\psi \qquad [23]$$

where

$$[\ldots] \equiv t_L^3(1-\gamma_5) + t_R^3(1+\gamma_5) - 2Q\sin^2\theta_W \qquad [24]$$

and

$$\rho_0 = m_W^2/m_Z^2\cos^2\theta_W \qquad [25]$$

All couplings given in this section are obtained at tree level and are modified in higher orders of perturbation theory. In particular, the relations between m_W and $\sin\theta_W$ (eqns [21] and [22]) and the observed values of ρ ($\rho = \rho_0$ at tree level) in different NC processes are altered by computable small electroweak radiative corrections.

The gauge-boson self-interactions can be derived from the $F_{\mu\nu}$ term in $\mathcal{L}_{\text{symm}}$, by using eqn [15] and $W^\pm = (W^1 \pm iW^2)/\sqrt{2}$. For the three-gauge-boson vertex W^+W^-V with $V = Z, \gamma$, we obtain

$$\begin{aligned} \Gamma_{W^-W^+V} = ig_{W^-W^+V}\,&[g_{\mu\nu}(q-p)_\lambda + g_{\mu\lambda}(p-r)_\nu \\ &+ g_{\nu\lambda}(r-q)_\mu] \end{aligned} \qquad [26]$$

with

$$\begin{aligned} g_{W^-W^+\gamma} &= g\sin\theta_W = e \quad \text{and} \\ g_{W^-W^+Z} &= g\cos\theta_W \end{aligned} \qquad [27]$$

This form of the triple gauge vertex is very special: in general, there could be departures from the above SM expression, even restricting us to SU(2) ⊗ U(1) gauge symmetric and C and P invariant couplings. In fact, some small corrections are already induced by the radiative corrections. The SM form of the triple gauge vertex has been experimentally confirmed by measuring the cross section $e^+e^- \to W^+W^-$ at LEP.

We now turn to the Higgs sector of the electroweak Lagrangian. The Higgs Lagrangian is specified by the gauge principle and the requirement of renormalizability to be

$$\mathcal{L}_{\text{Higgs}} = (D_\mu\phi)^\dagger(D^\mu\phi) - V(\phi^\dagger\phi) - \bar{\psi}_L\Gamma\psi_R\phi - \bar{\psi}_R\Gamma^\dagger\psi_L\phi^\dagger \quad [28]$$

where ϕ is a column vector including all Higgs scalar fields; it transforms as a reducible representation of the gauge group. The quantities Γ (which include all coupling constants) are matrices that make the Yukawa couplings invariant under the Lorentz and gauge groups. The potential $V(\phi^\dagger\phi)$, symmetric under SU(2) ⊗ U(1), contains, at most, quartic terms in ϕ so that the theory is renormalizable:

$$V(\phi^\dagger\phi) = -\tfrac{1}{2}\mu^2\phi^\dagger\phi + \tfrac{1}{4}\lambda(\phi^\dagger\phi)^2 \quad [29]$$

Spontaneous symmetry breaking is induced if the minimum of V, which is the classical analog of the quantum-mechanical vacuum state (both are the states of minimum energy) is obtained for nonvanishing ϕ values. This occurs because we have taken μ^2 and λ positive in V (note the "wrong" sign of the mass term). Precisely, we denote the vacuum expectation value (VEV) of ϕ, that is, the position of the minimum, by v:

$$\langle 0|\phi(x)|0\rangle = v \neq 0 \quad [30]$$

The fermion mass matrix is obtained from the Yukawa couplings by replacing $\phi(x)$ by v:

$$M = \bar{\psi}_L\mathcal{M}\psi_R + \bar{\psi}_R\mathcal{M}^\dagger\psi_L \quad [31]$$

with

$$\mathcal{M} = \Gamma \cdot v \quad [32]$$

In the SM, where all left fermions, ψ_L, are doublets and all right fermions, ψ_R, are singlets, only Higgs doublets can contribute to fermion masses. There are enough free couplings in Γ, so that one single complex Higgs doublet is indeed sufficient to generate the most general fermion mass matrix. It is important to observe that by a suitable change of basis we can always make the matrix $\mathcal{M}$ Hermitian, γ_5-free and diagonal. In fact, we can make separate unitary transformations on ψ_L and ψ_R according to

$$\psi'_L = U\psi_L, \qquad \psi'_R = V\psi_R \quad [33]$$

and consequently

$$\mathcal{M} \to \mathcal{M}' = U^\dagger\mathcal{M}V \quad [34]$$

This transformation does not alter the general structure of the fermion couplings in $\mathcal{L}_{\text{symm}}$.

If only one Higgs doublet is present, the change of basis that makes $\mathcal{M}$ diagonal will at the same time diagonalize also the fermion–Higgs Yukawa couplings. Thus, in this case, no flavor-changing neutral Higgs exchanges are present. This is not true, in general, when there are several Higgs doublets. But one Higgs doublet for each electric charge sector, that is, one doublet coupled only to u-type quarks, one doublet to d-type quarks, one doublet to charged leptons would also be satisfactory, because the mass matrices of fermions with different charges are diagonalized separately. In fact, at the moment, the simplest model with only one Higgs doublet seems adequate for describing all observed phenomena.

Weak charged currents are the only tree-level interactions in the SM that change flavor: by emission of a W, a u-type quark is turned into a d-type quark, or a ν_l neutrino is turned into an l^- charged lepton (all fermions are left-handed). If we start from a u-type quark that is a mass eigenstate, emission of a W turns it into a d-type quark state d' (the weak isospin partner of u) that in general is not a mass eigenstate. In general, the mass eigenstates and the weak eigenstates do not coincide and a unitary transformation connects the two sets:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V \begin{pmatrix} d \\ s \\ b \end{pmatrix} \quad [35]$$

or, in shorthand, $D' = VD$, where V is the Cabibbo–Kobayashi–Maskawa (CKM) matrix. Thus, in terms of mass eigenstates the charged weak current of quarks is of the form

$$J^+_\mu \propto \bar{u}\gamma_\mu(1-\gamma_5)VD \quad [36]$$

Since V is unitary (i.e., $VV^\dagger = V^\dagger V = 1$) and commutes with T^2, T_3, and Q (because all d-type quarks have the same isospin and charge) the neutral current couplings are diagonal both in the primed and unprimed basis (if the Z d-type quark current is abbreviated as $\bar{D}'\Gamma D'$ then by changing basis we get $\bar{D}V^\dagger\Gamma VD$ and V and Γ commute because, as seen from eqn [24], Γ is made of Dirac matrices and T_3 and Q generator matrices). It follows that $\bar{D}'\Gamma D' = \bar{D}\Gamma D$. This is the Glashow–Iliopoulos–Maiani (GIM)

mechanism that ensures natural flavor conservation of the neutral current couplings at the tree level. For three generations of quarks, the CKM matrix depends on four physical parameters: three mixing angles and one phase. This phase is the unique source of CP violation in the SM.

We now consider the gauge-boson masses and their couplings to the Higgs. These effects are induced by the $(D_\mu\phi)^\dagger(D^\mu\phi)$ term in $\mathcal{L}_{\text{Higgs}}$ (eqn [28]), where

$$D_\mu\phi = \left[\partial_\mu + \mathrm{i}g\sum_{A=1}^{3} t^A W_\mu^A + \mathrm{i}g'(Y/2)B_\mu\right]\phi \qquad [37]$$

Here t^A and $1/2Y$ are the $\mathrm{SU}(2)\otimes\mathrm{U}(1)$ generators in the reducible representation spanned by ϕ. Not only doublets but all non-singlet Higgs representations can contribute to gauge-boson masses. The condition that the photon remains massless is equivalent to the condition that the vacuum is electrically neutral:

$$Q|v\rangle = (t^3 + \tfrac{1}{2}Y)|v\rangle = 0 \qquad [38]$$

The charged W mass is given by the quadratic terms in the W field arising from $\mathcal{L}_{\text{Higgs}}$, when $\phi(x)$ is replaced by v. We obtain

$$m_{\mathrm{W}}^2 W_\mu^+ W^{-\mu} = g^2\left|\left(t^+ v/\sqrt{2}\right)\right|^2 W_\mu^+ W^{-\mu} \qquad [39]$$

whilst for the Z mass we get (recalling eqn [15])

$$\tfrac{1}{2}m_{\mathrm{Z}}^2 Z_\mu Z^\mu = \left|\left[g\cos\theta_{\mathrm{W}} t^3 - g'\sin\theta_{\mathrm{W}}(Y/2)\right]v\right|^2 Z_\mu Z^\mu \qquad [40]$$

where the factor of 1/2 on the left-hand side is the correct normalization for the definition of the mass of a neutral field. For Higgs doublets

$$\phi = \begin{pmatrix}\phi^+\\ \phi^0\end{pmatrix}, \qquad v = \begin{pmatrix}0\\ v\end{pmatrix} \qquad [41]$$

we obtain

$$m_{\mathrm{W}}^2 = 1/2g^2v^2, \qquad m_{\mathrm{Z}}^2 = 1/2g^2v^2/\cos^2\theta_{\mathrm{W}} \qquad [42]$$

Note that by using eqn [20] we obtain

$$v = 2^{-3/4}G_{\mathrm{F}}^{-1/2} = 174.1\,\mathrm{GeV} \qquad [43]$$

It is also evident that for Higgs doublets

$$\rho_0 = m_{\mathrm{W}}^2/m_{\mathrm{Z}}^2\cos^2\theta_{\mathrm{W}} = 1 \qquad [44]$$

This relation is typical of one or more Higgs doublets and would be spoiled by the existence of, for example, Higgs triplets. This result is valid at the tree level and is modified by calculable small electroweak radiative corrections. The ρ_0 parameter has been measured from the intensity of NC interactions (recall eqn [25]) and confirmed to be close to unity at a few per milli level.

In MSM only one Higgs doublet is present. Then the fermion–Higgs couplings are in proportion to the fermion masses. In fact, from the Yukawa couplings $g_{\phi\bar{f}f}(\bar{f}_{\mathrm{L}}\phi f_{\mathrm{R}} + \text{h.c.})$, the mass m_f is obtained by replacing ϕ by v, so that $m_f = g_{\phi\bar{f}f}v$. In MSM, three out of the four Hermitian fields are removed from the physical spectrum by the Higgs mechanism and become the longitudinal modes of W^+, W^-, and Z which acquire a mass. The fourth neutral Higgs is physical and should be found. If more doublets are present, two more charged and two more neutral Higgs scalars should be around for each additional doublet.

The couplings of the physical Higgs H to the gauge bosons can be simply obtained from $\mathcal{L}_{\text{Higgs}}$, by the replacement

$$\phi(x) = \begin{pmatrix}\phi^+(x)\\ \phi^0(x)\end{pmatrix} \to \begin{pmatrix}0\\ v + (H/\sqrt{2})\end{pmatrix} \qquad [45]$$

(so that $(D_\mu\phi)^\dagger(D^\mu\phi) = 1/2(\partial_\mu H)^2 + \cdots$), with the result

$$\begin{aligned}\mathcal{L}[H,W,Z] &= g^2(v/\sqrt{2})W_\mu^+ W^{-\mu}H + (g^2/4)W_\mu^+ W^{-\mu}H^2\\ &+ \left[(g^2 v Z_\mu Z^\mu)/\left(2\sqrt{2}\cos^2\theta_{\mathrm{W}}\right)\right]H\\ &+ \left[g^2/(8\cos^2\theta_{\mathrm{W}})\right]Z_\mu Z^\mu H^2\end{aligned}$$

In MSM, the Higgs mass $m_{\mathrm{H}}^2 \sim \lambda v^2$ is of order of the weak scale v but cannot be predicted because the value of λ is not fixed. The dominant decay mode of the Higgs is in the $b\bar{b}$ channel below the WW threshold, while the W^+W^- channel is dominant for sufficiently large m_{H}. The width is small below the WW threshold, not exceeding a few MeV, but increases steeply beyond the threshold, reaching the asymptotic value of $\Gamma \sim 1/2m_{\mathrm{H}}^3$ at large m_{H}, where all energies and masses are in TeV.

A central role in the experimental verification of the standard electroweak theory has been played by CERN, the European Laboratory for Particle Physics, located near Geneva, between France and Switzerland. The indirect effects of the Z_0, that is, the occurrence of weak processes induced by the neutral current, were first observed in 1974 at CERN by the Collaboration Gargamelle (the name of the bubble chamber used in the experiment). Later, in 1982, the $W^\pm$ and the Z_0 were, for the first time, directly produced and observed in proton–antiproton collisions by the UA1 and UA2 collaborations and then further studied with the same technique both at CERN and subsequently at the Tevatron of Fermilab near Chicago. Starting from 1989 LEP, the large e^+e^- collider was functioning at CERN till 2000. In the LEP circular ring of circumference ~27 km, electrons and

positrons were accelerated in opposite directions to an equal energy in the range between 45 and 103 GeV. The beams were made to cross and collide in correspondence of four experimental areas where the ALEPH, DELPHI, L3, and OPAL detectors were located to study the final states produced in the collisions. In its first phase, called LEP1, from 1989 to 1995 the LEP operation had been completely dedicated to a precise study of the Z_0 properties, mass, lifetime, and decay modes in order to accurately test the predictions of the SM. The main lessons of the precision tests of the standard electroweak theory can be summarized as follows. It has been checked that the couplings of quarks and leptons to the weak gauge bosons $W^{\pm}$ and Z are indeed precisely those prescribed by the gauge symmetry. The accuracy of a few tenths of 1% for these tests implies that, not only the tree level, but also the structure of quantum corrections has been verified. Then, since the end of 1995, the energy of LEP was increased and the phase of LEP2 was started. The total energy was gradually increased up to 206 GeV. The main physics goals of LEP2 were the search for the Higgs and for possible new particles, the precise measurement of m_W and the experimental study of the triple gauge vertices $WW\gamma$ and WWZ_0. The Higgs particle of the SM could in principle be produced at LEP2 in the reaction $e + e^- \rightarrow Z_0 H$, which proceeds by Z_0 exchange. The nonobservation of the Higgs particle at LEP2 has allowed to establish a lower limit on its mass: $m_H \gtrsim 114\,\text{GeV}$. Indirect indications on the Higgs mass were also obtained from the precision tests of the SM, as the radiative effects depend logarithmically on m_H. The indication is that the Higgs mass cannot be too heavy if the SM is valid: $m_H \lesssim 219\,\text{GeV}$ at 95% c.l. In 2001, LEP was dismantled and, in its tunnel, a new double ring of superconducting magnets is being installed. The new accelerator, the LHC (Large Hadron Collider), will be a proton–proton collider of total center-of-mass energy 14 TeV. Two large experiments ATLAS and CMS will continue to search for the Higgs starting in the year 2007. The sensitivity of LHC experiments to the SM Higgs will go up to masses m_H of $\sim$1 TeV.

See also: Effective Field Theories; Electric–Magnetic Duality; Electroweak Theory; General Relativity: Experimental Tests; Noncommutative Geometry and the Standard Model; Perturbative Renormalization Theory and BRST; Quantum Chromodynamics; Quantum Electrodynamics and its Precision Tests; Quantum Field Theory: a Brief Introduction; Relativistic Wave Equations Including Higher Spin Fields; Renormalization: General Theory; Supersymmetric Particle Models.

Further Reading

Altarelli G (2000) *The Standard Electroweak Theory and Beyond*, Proceedings of the Summer School on Phenomenology of Gauge Interactions. Zuoz, Switzerland hep-ph/0011078.

Altarelli G (2001) *A QCD Primer*, Proceedings of the 2001 European School of High Energy Physics. Beatenberg, Switzerland, hep-ph/0204179.

Close F, Marten, and Sutton MC (2002) *The Particle Odyssey*. Oxford: Oxford University Press.

Fraser G (ed.) (1998) *The Particle Century*. London: The Institute of Physics.

Martin BR and Shaw G (1997) *Particle Physics*, 2nd edn. Chichester: Wiley.

Particle Data Group (2004) Review of particle physics. *Physics Letters B* 592: 1.

Perkins DH (2000) *Introduction to High Energy Physics*. Reading: Addison-Wesley.

Supergravity

K S Stelle, Imperial College, London, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction: Minimal $D=4$ Supergravity

The essential idea of supersymmetry is an extension of the relativistic structure group of spacetime, which in ordinary four-dimensional physics in the absence of gravity is the Poincaré group ISO(3, 1). In a minimal supersymmetric theory in flat $D=4$ spacetime, the minimal supersymmetry algebra (the "graded Poincaré algebra") adds spinorial generators Q_α to the Lorentz generators M_{mn} and the translational generators (momenta) P_m, where $m=0, 1, 2, 3$. The core relation is the "anticommutator" of two Q_α:

$$\{Q_\alpha, \bar{Q}_\beta\} = -2\gamma^m_{\alpha\beta} P_m \qquad [1]$$

where $\bar{Q}=Q^\dagger\gamma^0$ and the γ^m are the Dirac gamma matrices. In the minimal $D=4$ supersymmetry algebra, the spinor generator Q_α is taken to be Majorana: $Q=C(\bar{Q})^{\mathrm{T}}$, where C is the charge-conjugation matrix and A^{T} denotes the transpose of the matrix A. The full supersymmetry algebra adjoins to the anticommutation relation [1] the usual commutation relations among the Lorentz generators and the commutators of the Lorentz generators with the momenta and the spinors Q_α; the latter express respectively the vectorial and spinorial characters of P_m and Q_α:

$$\mathrm{i}[M_{mn}, M_{pq}] = \eta_{np}M_{mq} - \eta_{mp}M_{nq} \qquad [2]$$

$$\mathrm{i}[M_{mn}, P_q] = \eta_{nq}P_m - \eta_{mq}P_n \qquad [3]$$

$$\mathrm{i}[M_{mn}, Q_\alpha] = \tfrac{1}{2}(\gamma_{mn}Q)_\alpha \qquad [4]$$

where $\gamma_{mn}=(1/2)(\gamma_m\gamma_n-\gamma_n\gamma_m)$ and $\eta_{mn}=\mathrm{diag}(-1, 1, 1, 1)$ is the Minkowski metric. The final relation in the supersymmetry algebra expresses the flatness of Minkowski space:

$$[P_m, P_n] = 0 \qquad [5]$$

This algebra has been considered as an extension of the symmetry algebra of particle physics since the work of Gol'fand and Likhtman in 1971, and especially since the linearly realized supersymmetric model of Wess and Zumino in 1974. That model contains a pair of $D=4$ scalar fields and a $D=4$ Majorana spinor, so the numbers of bosonic and fermionic degrees of freedom are equal; this is a fundamental characteristic of supersymmetric theories.

The work of Wess and Zumino led to an explosion of interest in supersymmetry, especially once it was realized that renormalizable supersymmetric models display a cancellation of some of the divergences that have plagued relativistic quantum field theory since its inception in the 1930s. In particular, in renormalizable flat-space field theory models, divergences quadratic in a high-momentum cutoff vanish as a result of cancellations between virtual bosonic and fermionic particles. This is a very attractive feature for control of the "hierarchy problem" in particle physics, especially for the instability inherent in having vastly different scales within the same theory, for example, the TeV scale of ordinary electroweak physics and the 10^{16} GeV scale where unification with the strong interactions might come in.

When one includes gravity, the stability problems of particle physics become much more severe. Einstein's theory of general relativity is itself non-renormalizable, that is, its ultraviolet divergences are of different forms from the terms present in the original "classical" action and there is no acceptable finite set of correction terms that can be added to it to remove this defect. Moreover, when otherwise tolerably behaved matter field theories that are renormalizable in a flat-spacetime context are coupled to general relativity, the gravitational couplings pollute the matter theories with non-renormalizable divergences. This is a key aspect of the great difficulty that has been encountered in interpreting gravity as a quantum theory.

Supersymmetry, with its divergence-canceling powers, was thus a very attractive option in the struggle to formulate a quantum theory of gravity, and the creation of a supergravity theory was thus a very high priority task. This was achieved in 1976 by Freedman, Ferrara, and Van Nieuwenhuizen using the technique of iterative Noether coupling to build up this nonlinear theory order-by-order in powers of the fermionic fields. The fermionic partner of the massless spin-2 "graviton" field is a massless fermionic spin-3/2 field that has come to be called the "gravitino."

A second 1976 paper by Deser and Zumino soon followed, emphasizing how supergravity manages to circumvent the well-known problems of coupling spins higher than 1 to gravity. A key point in achieving this result is the role played by the local version of the supersymmetry algebra [1]–[5]. As one can see from the translations occurring on the right-hand side of [1], when one replaces translation symmetry by local general coordinate invariance in a gravitational context, the supersymmetry transformations must themselves become local as well. Local symmetries allow for transformation parameters

that are local in the spacetime coordinates x^m, and in interacting theories they require coupling of the corresponding "gauge field" to a conserved current. In the case of supergravity, the gravitino field $\psi_{m\alpha}$ plays this gauge-field role, and its coupling to the conserved current of supersymmetry is the key to allowing a consistent coupling between the spin-2 graviton and the spin-3/2 gravitino.

The Minimal Supergravity Action

The action for minimal supergravity in $D=4$ dimensions can be written, using the vierbein formalism where the metric is expressed as a quadratic expression in a nonsymmetric 4×4 vierbein matrix e^a_m, $g_{mn}=e^a_m e^b_n \eta_{ab}$, as

$$I=\frac{1}{2\kappa^2}\int \mathrm{d}^4x \det(e)R(e,\omega(e)+K(\psi)) - \frac{\mathrm{i}}{2}\int \mathrm{d}^4x \epsilon^{mnpq}\bar{\psi}_m\gamma_5\gamma_n D_p(e,\omega(e)+K(\psi))\psi_q \quad [6]$$

where $\kappa=\sqrt{8\pi G}$ is the gravitational coupling constant,

$$\omega^{ab}_m(e)=\tfrac{1}{2}e^{na}\left(e^b_{n,m}-e^b_{m,n}\right)-\tfrac{1}{2}e^{nb}\left(e^a_{n,m}-e^a_{m,n}\right)+\tfrac{1}{2}e^{na}e^{rb}(e_{nc,r}-e_{rc,n})e^c_m \quad [7]$$

is the usual vierbein formalism spin connection (in which $e^b_{n,m}=\partial_m e^b_n$ and e^{ma} is the matrix inverse of e_{ma}), and

$$K^{ab}_m(\psi)=\frac{\mathrm{i}\kappa^2}{4}(\bar{\psi}_m\gamma^a\psi^b+\bar{\psi}^a\gamma_m\psi^b-\bar{\psi}_m\gamma^b\psi^a) \quad [8]$$

is the fermionic contorsion, an additional part of the covariant derivative $D_m(e+K(\psi))$ appearing in the action [6]. (Indices m,n are taken to be "world" indices while indices a,b are "tangent space" indices; one can convert from one type to another using the vierbein e^a_m and its inverse, e.g., $\psi_{a\alpha}=e^m_a\psi_{m\alpha}$.)

Keeping the terms in the action grouped as above using the nonstandard covariant derivative $e^{ab}_m+K^{ab}_m$ is what has been called "1.5 order formalism": this greatly simplifies the writing and analysis of the supergravity action [6]. In the action [6], one has the Ricci scalar $R(e,\omega(e)+K(\psi))$ written in terms of this generalized torsional spin connection. One may of course expand out all the $\omega^{ab}_m+K^{ab}_m$ combinations and write the nonlinear fermionic terms separately. Doing this produces a quartic term

$$L_4=\frac{\kappa^2}{32}[\bar{\psi}^b\gamma^a\psi^c(\bar{\psi}_b\gamma_a\psi_c+2\bar{\psi}_a\gamma_b\psi_c) - 4(\bar{\psi}^a\gamma^b\psi^c)(\bar{\psi}_a\gamma_b\psi_c)]$$

showing the highly nonlinear nature of supergravity theory – when expanded out, the theory becomes much more cumbersome to study. The 1.5 order formalism trick is one of a large number of algebraic simplifications that had to be developed in order to master the technical aspects of supergravity. It also reveals a characteristic physical feature: this theory naturally involves a connection with torsion built from the fermionic fields.

In terms of the torsional covariant derivative $D_m\epsilon(x)=(\partial_m+(1/4)(\omega^{ab}_m(e)+K^{ab}_m(\psi))\gamma_{ab})\epsilon(x)$ of the infinitesimal supersymmetry parameter $\epsilon(x)$, the local supersymmetry transformations which leave the action [6] invariant (up to the integral of a total derivative) are

$$\delta e^a_m=\mathrm{i}\bar{\epsilon}\gamma_a\psi_m \quad [9]$$

$$\delta\psi_m=2\kappa^{-1}D_m\epsilon \quad [10]$$

The inhomogeneous part $2\kappa^{-1}\partial_m\epsilon$ in the gravitino transformation [10] demonstrates the gauge-field nature of the gravitino field. For a distribution of "supermatter" fields (e.g., Wess–Zumino model scalars and spinors), the integrated "charge" that one would get from a Gauss's law surface integral at spatial infinity using the gravitino gauge field is the total supercharge Q_α, which in turn plays the role of the supersymmetry generator in the original matter-sector supersymmetry algebra [1].

Both the gravitational field and the gravitino field are thus effectively gauge fields, albeit not of a standard Yang–Mills type. The local algebra is a deformation of the rigid supersymmetry algebra [1]–[5], generalizing the relation between general covariance and flat-space Poincaré symmetry. Some basic consequences of the flat-space algebra are preserved, however. An extremely important instance of this is energy positivity. As one can see by multiplying [1] by γ^0 and then contracting on the spinor index,

$$E=P^0=\frac{1}{2}\sum_\alpha\{Q_\alpha,Q^\dagger_\alpha\}$$

The right-hand side is manifestly non-negative provided the theory is quantized in a positive-metric Hilbert space. One can see this even more explicitly in a Majorana spinor basis, where $Q^\dagger_\alpha=Q_\alpha$. Accordingly, for flat-space supersymmetric theories, one obtains directly the result that energy is non-negative. This carries over to the local algebra of supergravity, where the total energy is obtained from a Gauss's law integral over the sphere at spatial infinity.

In general relativity, an integrated energy can be defined with respect to an asymptotic timelike Killing vector at spatial infinity. Showing that this

energy is non-negative remained for decades a famously unsolved problem in gravitational physics; it was ultimately proven in Yau's positive-energy theorem. The algebraic structure of supergravity makes energy positivity much more transparent, however. Since pure general relativity can be obtained by setting the gravitino field to zero, this result is inherited by pure Einstein theory as a consequence of its being embeddable into supergravity. Energy positivity can thus be proved even at the classical level using ideas taken from supergravity, as was done by Witten and later streamlined by Nester, in an argument much simpler than Yau's proof. This argument writes the energy as an integral over a positive-semidefinite expression quadratic in a commuting spinor field which is analogous to the (anticommuting) spinor parameter of supergravity in the transformations [9] and [10].

Auxiliary Fields and Superspace

Supergravity shares with flat-space supersymmetric theories a curious technical feature that gives a hint of a new underlying geometry. Standard counting of the gauge-invariant continuous degrees freedom of the graviton and the gravitino in momentum space yield the same result per momentum value: two bosonic degrees of freedom and two fermionic degrees of freedom. This accords with the general requirement in supersymmetric theories that the numbers of bosonic and fermionic degrees of freedom match. This count follows from the Einstein and spin-3/2 equations of motion, or "on-shell." If one compares the count of nongauge degrees of freedom without using the equations of motion (i.e., "off-shell"), one obtains an imbalance, however: six nongauge graviton versus 12 nongauge fermion fields. This is directly related to another puzzling feature of the supergravity realization of local supersymmetry: the local supersymmetry algebra closes onto a finite set of transformations only when the equations of motion are imposed.

As in flat-space supersymmetry, the cure for this problem is to add nondynamical "auxiliary" fields to the action. In the supergravity case, the imbalance in the off-shell bose–fermi field count indicates that an additional six bosonic fields are needed. In the minimal set of auxiliary fields, these organize into a vector b_m and a scalar-pseudoscalar pair M, N; the additional terms in the action [6] are simply

$$\int \mathrm{d}^4x \det(e)\left(-\tfrac{1}{3}M^2 - \tfrac{1}{3}N^2 + \tfrac{1}{3}b_m b^m\right)$$

while the local supersymmetry transformations are changed to include the auxiliary fields, e.g., the gravitino transformation becomes

$$\delta\psi_m = 2\kappa^{-1}D_m(\omega, K)\epsilon + \gamma_5\left(b_m - \tfrac{1}{3}\gamma_m\gamma^n b_n\right)\epsilon - \tfrac{1}{3}\gamma_m(M + \gamma_5 N)\epsilon)$$

while the auxiliary fields transform into expressions that vanish on-shell. Since the field equations for the auxiliary fields are algebraic in character and since for source-free supergravity they have the simple solution $b_m = M = N = 0$, one can directly regain the on-shell formalism by algebraically eliminating the auxiliary fields.

The inclusion of auxiliary fields is not an empty trick, however. The local supersymmetry transformations including the auxiliary fields form a closed set without the use of equations of motion ("off-shell closure"). This standardizes the form of the supersymmetry transformations so that they remain the same even when supermatter is coupled to supergravity instead of needing a case-by-case Noether construction as in the case without the auxiliary fields. In this way, a standard set of coupling rules can be drawn up, known as the "tensor calculus." This tensor calculus is of great importance as it allows for the construction of general models of supergravity coupled to supermatter (Wess–Zumino multiplets and super Yang–Mills multiplets consisting of spin-1 gauge fields and spin-1/2 "gaugino" fields). These general couplings form the basis for essentially all supersymmetric phenomenology, and in particular for the formulation of the Minimal Supersymmetric Standard Model. Since supersymmetry is not directly observed in low-energy physics, it must be spontaneously broken, like many other gauge symmetries. As it happens, the physically realistic mechanisms of supersymmetry breaking all originate from supergravity couplings derived using the tensor calculus.

Given the regular set of tensor calculus rules for coupling supergravity to supermatter, one is led to suspect that a geometrical structure lies in the background. This is indeed the case; the corresponding construction is known as "superspace."

The basic idea of superspace is a generalization of the coset space construction of Minkowski space as the coset space given by the Poincaré group divided by the Lorentz group: $\mathcal{M}_4(x^m) = \mathrm{ISO}(3,1)/\mathrm{SO}(3,1)$. For supersymmetric theories, one analogously constructs $\mathrm{Superspace}(x^m, \theta^\alpha) = \mathrm{Graded\ Poincaré}/\mathrm{SO}(3,1)$. The basic ideas of superspace were introduced by Akulov and Volkov in 1972, while the idea of expanding in "functions" on this space, thus yielding "superfield," was introduced by Salam and Strathdee

in 1974. This led to a formulation of the Wess–Zumino model in terms of a chiral superfield $\phi(x,\theta)$, which is subjected to a covariant superspace constraint.

In order to manage the formalism of superspace more efficiently, it is convenient to use a two-component spinor formalism corresponding to the Weyl basis for the Dirac gamma matrices, in which the Majorana spinor coordinate θ is represented as

$$\theta = \begin{pmatrix} \theta_\alpha \\ \bar{\theta}^{\dot{\alpha}} \end{pmatrix}$$

where two-component indices $\alpha, \dot{\alpha} = 1, 2$ are raised and lowered with the covariant two-index antisymmetric tensors $\epsilon^{\alpha\beta}, \epsilon^{\dot{\alpha}\dot{\beta}}$, which both take the numerical value $i\sigma_2$. The flat-space fermionic covariant derivatives are then

$$\begin{aligned} D_\alpha &= \frac{\partial}{\partial\theta^\alpha} + i\sigma^m_{\alpha\dot{\beta}}\bar{\theta}^{\dot{\beta}}\partial_m \\ \bar{D}_{\dot{\alpha}} &= -\frac{\partial}{\partial\bar{\theta}^{\dot{\alpha}}} + i\bar{\theta}^\beta\sigma^m_{\beta\dot{\alpha}}\partial_m \end{aligned} \qquad [11]$$

where the $\sigma^m_{\alpha\dot{\beta}} = (1, \sigma_i)$ for $m = (0, i)$ (where σ_i are the Pauli matrices) are the Van der Waerden matrices which establish the mapping between vector indices and (chiral, antichiral) spinor index pairs. The Wess–Zumino multiplet is then described by a complex chiral superfield satisfying the constraint $\bar{D}_{\dot{\alpha}}\phi = 0$. Unlike the situation in Minkowski space, where the only Lorentz-covariant solution to a constraint that sets to zero the $\partial/\partial x^m$ derivatives is a constant, superspace has a reducible set of coordinates $(x^m, \theta^\alpha, \bar{\theta}^{\dot{\alpha}})$ and, as a result, requiring ϕ to be annihilated by $\bar{D}_{\dot{\alpha}}$ does not require the whole superfield to be a constant.

Since the fermionic coordinates of superspace $\theta^\alpha, \bar{\theta}^{\dot{\alpha}}$ are anticommuting (i.e., they are elements of a Grassman algebra), and since $\alpha, \dot{\alpha} = 1, 2$ have an index range of two, powers of them higher than the second order necessarily vanish. As a result, superfields like ϕ can be expanded into sets of component fields, each of which is an ordinary field in Minkowski space. In this way, a chiral superfield expands into $(A(x), B(x), \chi_\alpha(x), \bar{\chi}_{\dot{\alpha}}(x), F(x), G(x))$, where the fields A, B, χ, and $\bar{\chi}$ are the physical fields of the Wess–Zumino model, while F and G are dimension-2 auxiliary fields. In this way, the auxiliary fields of supersymmetry naturally fit into a superspace formalism as higher components in a superfield expansion. It is in this sense that they point toward the superspace formulations of supersymmetric theories.

For supergravity, there are a number of different approaches to realizing the theory in superspace, and these correspond naturally to the various possible choices of auxiliary-field sets. With the minimal set, the supergravity multiplet is described by a superfield carrying a vector index $H_m(x, \theta, \bar{\theta})$; this superfield is called the prepotential of supergravity. Note the fact that since the divisor group in the coset-space construction of superspace is the Lorentz group, superfields may carry indices corresponding to any Lorentz representation. The component-field expansion of the H_m superfield yields the physical $e^a_m, \psi_{m\alpha}, \bar{\psi}_{m\dot{\alpha}}$ and auxiliary fields (b_m, M, N) together with a number of other components of dimension lower than those of the physical fields. This is not, however, all that surprising: even the physical fields $e^a_m, \psi_{m\alpha}, \bar{\psi}_{m\dot{\alpha}}$ contain components that are not directly related to the physical modes because we are dealing with a gauge theory. What occurs in superspace is a redundant expression of the supergravity multiplet with the presence of various component gauge fields.

The full expression of local supersymmetry in superspace can be given in a number of different formalisms. Suffice it here to indicate the transformation of the linearized theory expanded in small fluctuations about empty flat superspace. Converting the vector index of H_m into a (chiral, antichiral) spinor index pair via $H_{\alpha,\dot{\beta}} = \sigma^m_{\alpha\dot{\beta}}H_m$, the linearized local symmetry transformation of the supergravity multiplet is

$$\delta H_{\alpha\dot{\beta}} = D_\alpha \bar{L}_{\dot{\beta}} - \bar{D}_{\dot{\beta}} L_\alpha \qquad [12]$$

where the transformation parameter superfield L_α carrying a spinor index is antichiral: $D_\alpha \bar{L}_{\dot{\beta}} = 0$ (while the conjugate parameter superfield $\bar{L}_\alpha$ is chiral). Expanding in component fields and comparing with the expansion of H_m, one sees that the chiral spinor superfield contains precisely the components needed to provide the standard gauge symmetries of e^a_m and $\psi_{m\alpha}, \psi_{m\dot{\alpha}}$ and also to transform the other gauge components of H_m as well. One can then make various gauge choices according to taste in a given context.

One frequently encountered superspace gauge choice sets to zero all the fields in H_m except for the physical and auxiliary fields $(e^a_m, \psi_{m\alpha}, \bar{\psi}_{m\dot{\alpha}}, b_m, M, N)$. This is called a Wess–Zumino gauge following the analogy to a similar construction for super Maxwell theory (containing spins 1 and 1/2). Wess–Zumino gauge choices are not, however, supersymmetrically covariant. This shows up when one works out the supersymmetry algebra in such a gauge: the presence of auxiliary fields gives closure, as required, without use of the equations of motion, but the anticommutator of two supersymmetry

transformations when acting on a gauge field such as the Maxwell field or the vierbein gives a combination of the anticipated translation with an admixture of a gauge transformation with a field-dependent parameter.

The prepotential superfield of minimal supergravity can itself be fit into larger formalisms in superspace that are analogous to standard differential geometry, with supervielbeins, superspin connections and so forth. An unavoidable feature of these more seemingly geometric constructions, however, is their high degree of redundancy: superspace vielbeins and spin connections carrying Lorentz indices have many component fields in addition to those found in the prepotential. This redundancy is then cut down in turn by imposing superspace constraints on the geometrical superfields, for example, on the components of the torsion tensor in superspace.

Extended Supergravities and Supergravities in Higher Dimensions

The possible graded extensions of the Poincaré algebra allow for more than one spinorial generator. Thus, one can have N supersymmetry generators $Q^i_\alpha, \bar{Q}_{\dot{\beta}j}, i,j=1,\ldots N$, with basic anticommutators (in Lorentz two-component notation)

$$\{Q^i_\alpha, \bar{Q}_{\dot{\beta}i}\} = 2\delta^i_j \sigma^m_{\alpha\dot{\beta}} P_m \qquad [13]$$

$$\{Q^i_\alpha, Q^j_\beta\} = 2\epsilon_{\alpha\beta} a^{\ell ij} Z_\ell \qquad [14]$$

$$\{\bar{Q}_{\dot{\alpha}i}, \bar{Q}_{\dot{\beta}j}\} = 2\epsilon_{\dot{\alpha}\dot{\beta}} \bar{a}^\ell_{ij} Z_\ell \qquad [15]$$

The right-hand sides of [14] and [15] allow for the possibility of nonvanishing commutators between supersymmetry generators of the same chirality. As one can see from the overall symmetry in pairs of indices $(\alpha i, \beta j)$, the coefficients $a^{\ell ij}$ must be antisymmetric in the i, j indices, so such nonvanishing same-chirality anticommutators cannot occur for $N=1$. The corresponding abelian generators Z_ℓ are called central charges since they must commute with all the other $(Q^i_\alpha, \bar{Q}_{\dot{\beta}],j}, P_m)$ elements of the algebra.

The i,j indices may be endowed with a symmetry meaning as well, although this is not obligatory in every model. When the central charges are absent, $Z_\ell = 0$, one has U(N) (or SU(N)) as the maximal such external automorphism; the choice of index placement on Q^i_α and $\bar{Q}_{\dot{\beta}j}$ anticipates this. If such a symmetry is realized in a given model, the fact that the $Q^i_\alpha, \bar{Q}_{\dot{\beta}j}$ carry representations both for that symmetry and for the spacetime Poincaré symmetry demonstrates how supersymmetry evades the no-go theorem barring unified spacetime and internal symmetries. This theorem (the Coleman–Mandula theorem) can be evaded, since at the time it was written, graded Lie symmetry algebras were not yet considered. For nonzero central charges, the external automorphism algebra becomes a subalgebra of U(N) determined by the requirement that invariant antisymmetric tensors $a^{\ell ij}$ exist.

The representations of the algebra [13]–[14] span an increasing range of spins as the number N of $D=4$ supersymmetries increases. For massive representations without central charges, the spins of the smallest supersymmetry representation extend from states of spin 0 (scalars) up to spin $N/2$; with central charges, the spin range can be shortened down to a minimum range of $N/4$. For massless representations, the range of helicities in a PCT (parity–change–time reversal) symmetric multiplet is from $-N/4$ to $N/4$. This spin range has an important implication for the maximal extension of supersymmetry that can be realized in an interacting supersymmetric field theory, because no interacting theories with a finite set of spins exist for spins >2. Accordingly, the maximal extension of supersymmetry is $N=8$ for massless theories, and in order to have massive states with spins that do not exceed spin 2 in an $N=8$ theory, the central charges have to be active for maximal multiplet shortening.

The $N=8$ supergravity theory, found by Cremmer and Julia in 1978, is thus the largest possible supergravity in $D=4$ dimensions. It contains the following "spin" range (allowing for a certain imprecision of expression: for massless fields one should really speak only of helicities)

$N=8$ supergravity spins

Spin	2	$\frac{3}{2}$	1	$\frac{1}{2}$	0
Multiplicity	1	8	28	56	70

In order to realize the automorphism SU(8) symmetry, one has to consider the field strengths for the 28 spin-1 fields, separated into complex self-dual and anti-self-dual parts in their antisymmetric Lorentz indices. These complex field strengths can then be endowed with a complex 28-dimensional representation of SU(8). The 70 scalars, on the other hand, fit precisely into the four-index antisymmetric self-dual representation of SU(8), $\phi^{i_1 i_2 i_3 i_4} = 1/(4!)\epsilon^{i_1 i_2 i_3 i_4 j_1 j_2 j_3 j_4} \bar{\phi}_{j_1 j_2 j_3 j_4}$. It is the use of the eight-index epsilon tensor here that restricts the automorphism group to SU(8) instead of U(8).

The SU(8) automorphism symmetry of $N=8$ supergravity theory is linearly realized. It plays an important role in another symmetry of this theory which is highly nonlinear. This theory has a

remarkable nonlinear E_7 symmetry. In fact, the 70 scalars form a nonlinear sigma model with the fields taking their values in the coset space $E_7/SU(8)$ (of dimension $133-63=70$), where the SU(8) divisor is the linearly realized automorphism group discussed above.

The extended supergravities point to another aspect of supergravity theory: the existence of higher-dimensional supergravities, from which the extended theories in $D=4$ spacetime can be derived by Kaluza–Klein dimensional reduction. If one considers a D' dimensional massless theory in a spacetime where d dimensions form a compact d-torus, then the theory can be viewed as a $D=D'-d$ dimensional theory in which the discrete Fourier modes arising from the periodicity requirements on the d-torus give rise to towers of equally spaced massive Kaluza–Klein states, plus a massless sector in $D'-d$ dimensions corresponding to the modes with no dependence on the d-torus coordinates.

Importantly, $N=8$ supergravity in four-dimensional spacetime can be obtained in this way from a supergravity theory that exists in 11 spacetime dimensions. Upon dimensional reduction on a 7-torus to four dimensions, one obtains $N=8, D=4$ supergravity at the massless level, plus an infinite tower of massive $N=8$ supermultiplets with central charges so that their spin range extends only up to spin 2. This $D=11$ supergravity was in fact found before the $N=8$ theory by Cremmer, Julia, and Scherk, with the details of the more complicated $N=8, D=4$ theory being worked out via the techniques of Kaluza–Klein dimensional reduction. The fields of the $D=11$ theory include an exotic field type not encountered in $D=4$ theories: the bosonic fields of the theory comprise the graviton e_M^A plus a three-index antisymmetric tensor gauge field C_{MNP}. Counting the number of propagating modes of these fields for a given momentum value gives $44+84=128$ bosonic degrees of freedom. This precisely balances the 128 fermionic degrees of freedom coming from the $D=11$ gravitino $\psi_{M\alpha}$.

Supergravity Effective Theories, Strings and Branes

The hope for a cancellation of the ultraviolet divergences in a supersymmetric theory of gravity turned out to be ephemeral, although there is in fact a postponement of the divergence onset until a higher order in quantum field loops. There is agreement that the nonmaximal supergravities diverge at the three-loop order. For the $N=8, D=4$ theory, the situation remains unclear, but divergences are nonetheless expected to occur at some finite loop order.

This persistence of nonrenormalizability in $D=4$ supergravity theories is no longer seen as a disaster, however, because these theories are now seen as effective theories for the massless modes arising from a deeper microscopic quantum theory. In addition, the theories that are most directly connected to this underlying quantum theory are, surprisingly, the maximal supergravities in spacetime dimensions 10 and 11. $D=11$ supergravity can be dimensionally reduced on a 1-torus (i.e., a circle) to $D=10$ where the massless sector yields type IIA supergravity theory. This theory is the effective theory for a consistent quantum theory of type IIA superstrings in $D=10$. Theories of relativistic strings (i.e., one-dimensional extended objects) have strikingly different properties from theories of point particles. In particular, the spread-out nature of the interactions leads to a damping out of the quantum field theory divergences, while the underlying supersymmetry causes a cancellation of other infinities that could have arisen owing to the two-dimensional nature of the string world sheets. This gives, for the first time, a perturbatively well-defined quantum theory including gravity.

In addition to the type IIA theory, there are four other consistent superstring theories in $D=10$, and these are in turn related to various $D=10$ supergravity effective theories for the massless modes: type IIB, $E_8\times E_8$ heterotic, SO(32) heterotic, and SO(32) type I. Remarkably, the maximal $D=11$ supergravity enters into this picture as well, as a consequence of a pattern of duality symmetries that have been found among the superstring theories.

The dualities of string theory are directly related to the nonlinear symmetries of the dimensionally reduced supergravities in $D=4$. The string quantum corrections do not respect the E_7 symmetry of the classical $N=8$ theory, but they do respect a discrete subgroup of this symmetry in which the E_7 group elements are required to take integer values: $E_7(\mathbb{Z})$.

This quantum-level restriction to a discrete subgroup can be seen from another phenomenon characteristic of superstring theories: the existence of "electric" and "magnetic" brane solutions. The antisymmetric-tensor (or "form") fields of the higher-dimensional supergravities naturally give rise to solitonic solutions in which $p+1$ dimensions form a flat Poincaré invariant subspace. This can be interpreted as the world volume of an infinite p-brane extended object. In the $D=11$ supergravity theory, the branes that emerge in this way are a 2-brane and a 5-brane. The three-dimensional world volume of the 2-brane naturally couples to the

3-form field C_{MNP}, just as an ordinary Maxwell vector field couples to the one-dimensional world line of a point particle (or 0-brane). The 2-brane is thus naturally electrically charged with respect to the 3-form field; its charge can be obtained, in a direct generalization of the Maxwell case, from a Gauss' law integral of the field strength $H_{[4]} = dC_{[3]}$ over a 7-sphere at spatial infinity in the eight directions transverse to the brane worldvolume. The 5-brane, on the other hand, has a magnetic type charge; it is the 7-form dual to $H_{[4]}$ that is integrated to give its charge. In addition to these static infinite p-branes, the theory contains dynamical finite-extent branes as well, although for these one generally does not have explicit solutions.

As one reduces a higher-dimensional supergravity to lower and lower dimensions, there is a proliferation of solitonic brane solutions of varying dimensionality, and of both electric and magnetic charge types. In a quantum theory context, these electrically and magnetically charged branes pair up in ways that must satisfy a generalization of the Dirac quantization condition for $D = 4$ electric and magnetic point particles. This ends up requiring all the supergravity solitonic brane charges to lie on a charge lattice. It is the requirement that this discrete brane-charge lattice be respected that restricts the classical supergravity nonlinear symmetry groups to discrete duality subgroups.

The dualities relate brane solutions within a given theory and also between different string theories. They include transformations that invert the radii of compactifying tori, giving a large–small compactification scale duality. They also include transformations that invert the string coupling constant, thus interchanging strong and weak coupling. The type IIB theory, for example, is self-dual under strong–weak coupling duality. In the case of the type IIA theory, however, something remarkable happens. The strong coupling limit of this theory turns out to be related by duality, not to another string theory, but to the maximal $D = 11$ supergravity. The role of the Kaluza–Klein massive modes for the 11 to 10 reduction is played by an infinite tower of extremal charged black holes.

Thus, even $D = 11$ supergravity theory has a role to play in the effective theory of the underlying quantum dynamics. This underlying theory has been dubbed "M-theory." It is still only partially understood, but many of its most important properties are presaged by the remarkable nonlinear structure of the classical supergravities.

See also: Brane Construction of Gauge Theories; Brane Worlds; Branes and Black Hole Statistical Mechanics; Random Algebraic Geometry, Attractors and Flux Vacua; Renormalization: General Theory; Spinors and Spin Coefficients; Stability of Minkowski Space; Supermanifolds; Superstring Theories; Supersymmetric Particle Models; Symmetries and Conservation Laws; Symmetries in Quantum Field Theory: Algebraic Aspects.

Further Reading

Buchbinder JL and Kuzenko SM (1998) *Ideas and Methods of Supersymmetry and Supergravity*. Bristol: IoP Publishing Ltd.

Stelle KS (1998) BPS branes in supergravity, Trieste 1987 School of High-Energy Physics and Cosmology, arXiv:hep-th/9803116.

Van Nieuwenhuizen P (1981) Supergravity. *Physics Reports* 68: 189–398.

Wess J and Bagger J (1983) *Supersymmetry and Supergravity*. Princeton: Princeton University Press.

Supersymmetric Particle Models

S Pokorski, Warsaw University, Warsaw, Poland

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Supersymmetric quantum field theories (*see* Supergravity) are characterized by the existence of one ($N=1$ supersymmetry) or several ($N>1$ extended supersymmetry) conserved Noether-like charges Q_A $A=1,\ldots,N$, which establish symmetry links between particle states of different spin. Supersymmetry ensures equal numbers of bosonic and fermionic particle states. If it is exact, bosons and fermions related by supersymmetry transformations have equal masses. Moreover, supersymmetry imposes stringent relations between interactions which involve particles of different spin. This gives rise to a special ultraviolet behavior of supersymmetric theories. Their ultraviolet divergences are much softer than in nonsupersymmetric theories. In particular, $N=4$ supersymmetric quantum field theories are finite and for any N they are free from quadratic divergences plaguing ordinary theories with elementary scalars. $N>4$ supersymmetric theories necessarily involve particles of spin higher than 1 and are not renormalizable. Supersymmetry promoted to a local symmetry includes gravity.

Only $N=1$ supersymmetric theories allow for chiral fermions which are the fundamental objects in elementary particle interactions (*see* Standard Model of Particle Physics). This is because parity and

Notes:

charge conjugation symmetries are violated in weak interactions. Therefore, $N > 1$ theories may not be of immediate phenomenological relevance. However, they may be useful for constructing supersymmetric theories in more than four dimensions (more than three spatial dimensions). Chiral (effective) theory in four dimensions can be then obtained after compactification of extra dimensions. For instance, $N=2$ theory in five dimensions (x_μ, y) compactified on a circle with reflection symmetry $y \to -y$ (orbifold compactification) gives chiral $N=1$ theory in four dimensions.

Absence of quadratic divergences in supersymmetric theories is the main argument supporting the belief that fundamental interactions of elementary particles at energies not higher that $\mathcal{O}(1\,\text{TeV})$ should be described by an (approximately) $N=1$ supersymmetric extension of the standard model (SM). Indeed, supersymmetric models elegantly solve the so-called hierarchy problem of the SM. At present, supersymmetry remains a theoretical hypothesis. No experimental evidence for it has been found yet (for experimental lower bounds on the masses of supersymmetric particles see Eidelman *et al.* (2004)). Supersymmetric models will be tested experimentally at the Large Linear Collider at CERN (Geneva), after the completion of its construction in 2007. Supergravity theories may be physically relevant as an intermediate step in constructing phenomenologically viable models from superstring theories.

The essence of the hierarchy problem of the standard model (SM) – the successful $\text{SU}(3)_c \times \text{SU}(2)_L \times \text{U}(1)_Y$ gauge theory of interactions of quarks and leptons at energies up to about 100 GeV – is the following. By itself, the SM does not explain the value of the Fermi scale v of the electroweak $\text{SU}(2)_L \times \text{U}(1)_Y$ symmetry breaking ($v \sim G_F^{-1/2}$ where G_F is the Fermi constant determined by the life time of the muon). Indeed, in the SM, the electroweak symmetry breaking is realized by an elementary Higgs field H (an SU(2) doublet) with a potential

$$V = m^2 H^\dagger H + \frac{\lambda}{2}(H^\dagger H)^2 \qquad [1]$$

where m and λ are free parameters of the SM. When $m^2 < 0$ is chosen, the minimum of the potential occurs when

$$\langle H^\dagger H \rangle = -\frac{m^2}{\lambda} \equiv \frac{v^2}{2} \qquad [2]$$

that is, the Higgs doublet acquires $\text{SU}(2) \times \text{U}(1)_Y$ breaking vacuum expectation value v which is just the Fermi scale. The masses of the intermediate vector bosons $W^\pm$ and Z^0 are proportional to v and depend also on the gauge couplings. Within the SM understood as a theory with the momentum cut-off Λ_{SM}, quantum corrections to the mass parameter m^2 in eqn [1] are quadratically divergent:

$$\delta m^2 = \frac{3}{64\pi^2}(3g_2^2 + g_1^2 + \lambda - 8y_t^2)\Lambda_{\text{SM}}^2 + \cdots \qquad [3]$$

Here, g_1, g_2, and y_t are the gauge couplings of the groups $\text{U}(1)_Y, \text{SU}(2)_L$, and the top-quark Yukawa coupling, respectively. This means that if, above the energy scale Λ_{SM}, the SM is replaced by some more fundamental theory, in which there are particles of masses $M \gtrsim \Lambda_{\text{SM}}$, the quantum corrections to m^2 are quadratically dependent on the new mass scale M. For $M \gg v$, this is very unnatural even if the original parameter m^2 remains a free parameter of this underlying theory and particularly difficult to accept if in the underlying theory m^2 is fixed by some more fundamental considerations. If the SM was the correct theory up to, for example, the mass scale suggested by the see-saw mechanism for the neutrino masses, $\Lambda_{\text{SM}} \sim 10^{15}\,\text{GeV}$

$$|\delta m^2| \sim 10^{28}\,\text{GeV}^2 \sim 10^{24} v^2!$$

Clearly, this excludes the possibility of understanding the magnitude of the Fermi scale v in any sensible way. Thus, for naturalness of the Higgs mechanism in the SM there should exist a new mass scale $M \gtrsim v$, say only one order of magnitude higher than v and the theory describing the physics above that scale should be free of quadratic divergences. (Approximate) supersymmetry is at present the most elegant and theoretically most complete solution to the hierarchy problem of the SM.

Supersymmetric Extensions of the SM

In supersymmetry, the gauge fields A_μ^a are promoted to vector superfields $\hat{V}^a = (A_\mu^a, \lambda^a, D^a)$, one for each gauge symmetry group generator, where λ^a's are Weyl fermions (called gauginos) and D^a's are nondynamical auxiliary fields. A renormalizable supersymmetric gauge theory is completely defined (see, e.g., Sohnius (1985) and Wess and Bagger (1992)) by specifying the gauge group, the set of chiral supermultiplets $\hat{\Phi}_i = (\phi_i, \psi_i, F_i)$ representing matter fields, and the superpotential – a holomorphic polynomial function of at most third order in the chiral superfields which determines Yukawa couplings of the fermions ψ_i and scalars ϕ_i. Auxiliary fields D^a and F_i can be eliminated via their (algebraic) equations of motion.

The so-called minimal supersymmetric SM (MSSM) encodes the main features of any supersymmetric extension of the SM. Its gauge group is

$SU(3) \times SU(2) \times U(1)$ – the same as in the SM – and the chiral superfields are associated to each of the SM quark and lepton fields. Thus, quarks and leptons get scalar spin zero superpartners, the squarks and sleptons, carrying the same quantum numbers as their corresponding fermions and the vector superfields provide spin 1/2 superpartners for the gauge fields – the gluinos, the winos, and the bino. The SM Higgs doublet with weak hypercharge $Y=1/2$ becomes a scalar component of a chiral superfield $\hat{H}_u$ which contains in addition one doublet of Weyl fermions – the Higgsinos. The chiral anomaly cancelation condition requires that there be also a second Higgs chiral superfield $\hat{H}_d$ with $Y=-1/2$. Such a superfield is also required for giving masses to all flavors of quarks; because of the holomorphicity of the superpotential the same Higgs doublet cannot couple simultaneously to all quarks.

With the MSSM superfield content, the most general renormalizable superpotential consistent with the gauge symmetry has the form

$$W = Y_u \hat{U}^c \hat{Q} \hat{H}_u + Y_d \hat{D}^c \hat{Q} \hat{H}_d + Y_l \hat{E}^c \hat{L} \hat{H}_d + \mu \hat{H}_d \hat{H}_u + \lambda_1 \hat{D}^c \hat{Q} \hat{L} + \lambda_2 \hat{E}^c \hat{L} \hat{L} + \lambda_3 \hat{U}^c \hat{D}^c \hat{D}^c + \lambda_4 \hat{L} \hat{H}_u \quad [4]$$

(flavor indices are suppressed) where the superfield $\hat{Q}$ contains the SU(2) quark doublet Q and its scalar superpartner $\tilde{Q}$ and similarly for the lepton doublet $\hat{L}$, quark singlets $\hat{U}, \hat{D}$, and lepton singlet $\hat{E}$ superfields. The three first terms in [4] give the SM-like Yukawa couplings of quarks and leptons to the Higgs fields together with Yukawa couplings of the corresponding superpartners. The fourth term has no SM analogy; it gives supersymmetric masses to the Higgs scalar and Higgsinos. The interactions in the second line do not conserve baryon and lepton numbers, respectively B and L, and should be forbidden (or strongly suppressed) by some additional symmetry of the theory as they would lead to rapid proton decay. A discrete symmetry, called R-parity $R=(-1)^{2S+3(B-L)}$, where S is the spin of the field, is an interesting possibility. R-parity acts differently on the different components of the superfields: it is even for all SM particles and odd for their superpartners. Its conservation implies that superpartners must appear in pairs in any interaction vertex. Thus, with R-parity imposed, the lightest supersymmetric particle is stable and it is an excellent candidate for the dark matter in the universe.

Supersymmetry cannot be an exact symmetry of nature because there do not exist elementary fermions and bosons degenerate in mass. The superpotential [4] does not break supersymmetry spontaneously but even if it did the elementary fermions and bosons would on average have equal masses (they would satisfy some mass sum rule) which is also contradicted by the experimental data. Therefore, in the MSSM, supersymmetry has to be broken explicitly but in such a way that the soft ultraviolet behavior remains intact. Remarkably, the supersymmetry breaking terms which can be added to the MSSM Lagrangian without reintroducing quadratic divergences make heavy just those fields which are opposite statistics superpartners of the SM gauge bosons and fermions. These so-called soft terms are:

$$\begin{aligned}\mathcal{L}_{\text{soft}} = &-\tfrac{1}{2}\tilde{G}\tilde{G}^a\tilde{G}^a - \tfrac{1}{2}\tilde{W}\tilde{W}^a\tilde{W}^a - \tfrac{1}{2}\tilde{B}\tilde{B}\tilde{B} \\ &- m_Q^2|\tilde{Q}|^2 - m_U^2|\tilde{U}^c|^2 - m_D^2|\tilde{D}^c|^2 \\ &- m_L^2|\tilde{L}|^2 - m_E^2|\tilde{E}^c|^2 - m_{H_u}^2|H^u|^2 \\ &- m_{H_d}^2|H^d|^2 - m_3^2(H^u H^d + \text{c.c.}) \\ &+ A_U \tilde{U}^c \tilde{Q} H_u + A_D \tilde{D}^c \tilde{Q} H_d + A_E \tilde{E}^c \tilde{L} H_d \end{aligned} \quad [5]$$

and yield gaugino (gluino $\tilde{G}$, wino $\tilde{W}$, and bino $\tilde{B}$) and scalar mass terms as well as explicit trilinear couplings between scalars (scalar mass terms and A-terms are 3×3 matrices in the flavor space). As a result, supersymmetry is broken in the mass spectra but not in the dimensionless couplings.

The origin of the soft supersymmetry breaking remains an open issue. Terms [5] are most probably remnants of the spontaneous supersymmetry breaking in the so-called "hidden" sector – a hypothetical set of fields that do not interact directly with the MSSM fields. For example, in the popular scenario, they interact with the MSSM fields only gravitationally and spontaneous supersymmetry breaking in the hidden sector is communicated to the MSSM sector by gravitational interactions giving rise to terms [5]. Several other mechanisms of supersymmetry breaking transmission have also been proposed (gauge mediation, anomaly mediation, etc.).

The mass parameters and A-terms in [5] are free parameters of the low-energy supersymmetric theory and, combined with the interactions like $Q\tilde{Q}\tilde{G}$ originating from supersymmetric kinetic terms, may be a new, troublesome, source of flavor changing neutral currents and of CP violation.

Higgs Sector of the MSSM

The MSSM Higgs potential reads

$$V = m_1^2|H_d|^2 + m_2^2|H_u|^2 + m_3^2(H_u H_d + \text{c.c.}) + \frac{g_1^2 + g_2^2}{8}\left(|H_d|^2 - |H_u|^2\right)^2 \quad [6]$$

Its quartic part is uniquely determined by the structure of the supersymmetric gauge theory. The parameters m_1^2, m_2^2, and m_3^2 are determined by

the soft supersymmetry breaking Higgs boson masses [5] and the μ parameter in [4]. The potential [6] is bounded from below for $m_1^2 + m_2^2 > 2m_3^4$, and for $m_1^2 m_2^2 - m_3^2 < 0$ it has the electroweak symmetry breaking minimum at $v_u = \langle H_u^0 \rangle \neq 0, v_u = \langle H_d^0 \rangle \neq 0$. The ratio $v_u/v_d \equiv \tan\beta$ is then phenomenologically a very important parameter.

Quantum corrections to the mass parameters in [6] are controlled by the mass scale M_{soft} of the supersymmetry breaking terms [5]; at the one-loop level instead of [3], one finds

$$\delta m_{1,2}^2 \sim \frac{1}{16\pi^2}\left(3g_2^2 + g_1^2 - 12y_{b,t}^2\right) M_{\text{soft}}^2 \ln\frac{\Lambda_{\text{NEW}}^2}{M_{\text{soft}}^2} \quad [7]$$

where y_b and y_t are the bottom- and top-quark Yukawa couplings, respectively and Λ_{NEW} is the scale at which the soft supersymmetry breaking terms are generated by the putative supersymmetry breaking transmission mechanism. In gravity mediation scenarios, $\Lambda_{\text{NEW}} \sim M_{\text{Pl}}$. In gauge mediation scenarios, Λ_{NEW} is low but it is a new scale, introduced by hand.

In the softly broken supersymmetric models, the hierarchy problem is solved for $M_{\text{soft}} \lesssim \mathcal{O}(10)v$. Moreover, eqn [7] shows that via quantum corrections the large top-quark Yukawa coupling y_t drives the mass parameter m_2^2 to a negative value, inducing the electroweak symmetry breaking. This means that in supersymmetric models the electroweak scale is calculable in terms of the known coupling constants and the (unknown) scales M_{soft} and cutoff scale Λ_{NEW} to the MSSM. If $M_{\text{soft}} \lesssim \mathcal{O}(10)v$, the correct electroweak scale is obtained for $\Lambda_{\text{NEW}} \sim M_{\text{GUT}}$. This nicely fits with unification of the gauge couplings.

In supersymmetric models, the quartic couplings in the Higgs potential are restricted. This typically leads to a strong upper bound on the mass of the lightest Higgs particle. In the minimal model with the potential [6], at the tree level

$$M_{\text{Higgs}} < M_Z \approx 91\,\text{GeV} \quad [8]$$

This bound is substantially modified by quantum corrections. They depend quadratically on the top-quark mass and logarithmically on the stop mass scale $M_{\tilde{t}} \sim M_{\text{soft}}$:

$$M_{\text{Higgs}}^2 < \lambda v^2 \quad [9]$$

where λ is given by

$$\lambda = \tfrac{1}{8}(g_2^2 + g_1^2)\cos^2 2\beta + \Delta\lambda$$

$$\text{with } \Delta\lambda = \frac{3g_2^2}{8\pi^2}\frac{m_t^4}{v^2 M_W^2}\ln\frac{M_{\tilde{t}}^2}{m_t^2} \quad [10]$$

For $M_{\tilde{t}} < 1\,\text{TeV}, M_{\text{Higgs}} < 130\,\text{GeV}$.

The minimal-model bound on the Higgs mass can be relaxed in models with extended Higgs sector. For instance, if an additional gauge group singlet chiral superfield couples to the Higgs doublets, the Higgs self-coupling λ in [9] receives additional contributions. Explicit calculations show that in such and other models, with $M_{\text{soft}} \lesssim 1\,\text{TeV}$, the bound on the Higgs mass cannot be raised above $\sim$150 GeV if one wants to preserve perturbative gauge coupling unification.

Supersymmetric Grand Unified Theories

There are two striking aspects of the matter spectrum in the SM. One is the chiral anomalies cancelation (Weinberg 1996–2000, Pokorski 2000), which is necessary for a unitary (and renormalizable) theory, and occurs thanks to certain conspiracy between quarks and leptons suggesting a deeper link between them. The second one is that the spectrum fits into simple representations of the SU(5) and SO(10) groups (Ross 1985). Indeed, each generation of the SM matter fills $5^* + 10 + 1$ (if the right-handed neutrino is included into the spectrum) representations of SU(5) and for SO(10), $16 = 5^* + 10 + 1$. The assignment of fermions to the SU(5) or SO(10) representations fixes the normalization of the $U(1)_Y$ generator. Both facts suggest unification of strong and electroweak elementary forces in a grand unified theory with some bigger gauge symmetry group. Such unification implies that all the SM gauge forces become of equal strength at some unification scale. Their strength is measured by the running gauge couplings $\alpha_i = g_i^2/4\pi$, $i = 1, 2, 3$, of the three group factors $SU(3)_c \times SU(2)_L \times U(1)_Y$. The energy scale dependence of α_i is governed by the renormalization group equations. In the first nontrivial approximation, they read:

$$\frac{1}{\alpha_i(Q)} = \frac{1}{\alpha_i(M_Z)} - \frac{b_0^{(i)}}{2\pi}\ln\left(\frac{Q}{M_Z}\right) \quad [11]$$

Here, $1/\alpha_i(M_Z) = (58.98 \pm 0.04, 29.57 \pm 0.03, 8.40 \pm 0.14)$ are the experimental values of the gauge couplings at the Fermi scale and $b_0^{(i)}$ are the coefficients which depend on the matter content of the theory. They are

$$b_0 = \left(\tfrac{1}{10} + \tfrac{4}{3}N_g, -\tfrac{43}{6} + \tfrac{4}{3}N_g, -11 + \tfrac{4}{3}N_g\right)$$

in the SM and

$$b_0 = \left(\tfrac{3}{5} + 2N_g, -5 + 2N_g, -9 + 2N_g\right)$$

in the MSSM, where N_g is the number of fermion generations. In the SM, the running gauge couplings

approach each other at high scale of order 10^{13} GeV but never unify.

In the MSSM, with sparticle spectrum characterized by $M_{\text{soft}} \approx 1$ TeV and for the initial Fermi scale values given above, the three gauge couplings unify with high precision at the scale $M_{\text{GUT}} \sim 10^{16}$ GeV. Therefore, the MSSM can be embedded into supersymmetric grand unified theories with no hierarchy problem for the Fermi scale (it is stable with respect to radiative corrections generated by particles with masses $\sim M_{\text{GUT}}$) and no conflict with the measured values of the gauge couplings.

In the SM, the baryon number is (perturbatively) conserved since there are no renormalizable couplings violating this symmetry. Experimental search for proton decay, for example, $p \to e^+\pi^0$, $p \to K^+\nu$, is one of the most fundamental tests for particle physics. The present limit on the proton life time is $\tau_p > 10^{33}$ yr. In grand unified theories, baryon number conservation is violated by interactions mediated by the heavy gauge bosons corresponding to the enlarged gauge symmetry (e.g., SU(5)), spontaneously broken at M_{GUT} to the SM gauge symmetry. Such interactions manifest themselves at low energy as additional, nonrenormalizable interactions added to the SM Lagrangian. Proton decay is then induced by the set of dimension-6 operators of the form

$$O_i^{(6)} = \frac{c_i^{(6)}}{M_{(6)}^2} qqql \qquad [12]$$

where q, l denote quarks and leptons, respectively. For $c_i^{(6)} \sim \alpha_{\text{GUT}} \approx 1/25$, the experimental limit on τ_p requires $M_{(6)} \gtrsim 10^{15}$ GeV, consistently with $M_{\text{GUT}} = 10^{16}$ GeV in supersymmetric GUTs. However, in supersymmetric GUTs, there is still another, genuinely supersymmetric, source of contributions to the proton decay amplitudes. These are the dimension-5 operators

$$O_i^{(5)} = \frac{c_i^{(5)}}{M_{(5)}^2} qq\tilde{q}\tilde{l} \qquad [13]$$

where $\tilde{q}, \tilde{l}$ denote squarks and sleptons, respectively. Such operators originate from the exchange of the color triplet scalars present in the Higgs boson GUT multiplets, with $M_{(5)} \sim M_{\text{GUT}} \sim 10^{16}$ GeV, and $c^{(5)} \gtrsim 10^{-7}$ is given by the Yukawa couplings. Inserted into diagrams with gaugino exchanges they give rise to dimension-6 operators of the form [12]. One then gets $c^{(6)} = \alpha_{\text{GUT}} c^{(5)}$, $M_{(6)}^2 = M_{(5)} M_{\text{SUSY}}$. Given various uncertainties, for example, in the unknown squark, gaugino, and heavy Higgs boson mass spectrum, such contributions in supersymmetric GUT models predict the proton life time to be consistent with but close to the present experimental limits.

Summary

Supersymmetry is distinct in several very important points from all other proposed solutions to the hierarchy problem. First of all, it provides a general theoretical framework which allows one to address many physical questions. Supersymmetric models, like the MSSM or its simple extensions, satisfy a very important criterion of "perturbative calculability." In particular, they are easily consistent with the precision electroweak data. The SM is their low-energy approximation in the sense of the Appelquist–Carazzone decoupling, so most of the successful structure of the SM is built into supersymmetric models. The quadratically divergent quantum corrections to the Higgs mass parameter (the origin of the hierarchy problem in the SM) are absent in any order of perturbation theory. Therefore, the cutoff to a supersymmetric theory can be as high as the Planck scale, and "small" scale of the electroweak breaking is still natural. Supersymmetry is not only consistent with grand unification of elementary forces but, in fact, makes it very successful. And, finally, supersymmetry is needed for string theory.

However, there are also some problems to be solved: the hierarchy problem of the electroweak scale is solved but the origin of the soft supersymmetry breaking scale M_{soft} remains an open question: spontaneous supersymmetry breaking and its transmission to the visible sector is a difficult problem and a fully satisfactory mechanism which would yield M_{soft} hierarchically smaller than the Planck (string) scale has not yet been found. On the phenomenological side, there are new potential sources of flavor-changing neutral current transitions and of CP violation, and baryon and lepton numbers are not automatically conserved by the renormalizable couplings. But even those problems can at least be discussed in a concrete quantitative way.

See also: Brane Construction of Gauge Theories; Perturbation Theory and its Techniques; Seiberg–Witten Theory; Standard Model of Particle Physics; Supergravity; Supermanifolds.

Further Reading

Eidelman S *et al.* (The Particle Data Group) (2004) Review of particle physics. *Physics Letters B* 592: 1.

Kane GL (ed.) (1998a) *Perspectives in Supersymmetry*. Singapore: World Scientific.

Kane GL (ed.) (1998b) *Perspectives on Higgs Physics II*. Singapore: World Scientific.

Nilles H-P (1984) *Physics Reports* C 110: 1.

Pokorski S (2000) *Gauge Field Theories*, Cambridge Monographs on Mathematical Physics, 2nd edn. Cambridge: Cambridge University Press.

Ross GG (1985) *Grand Unified Theories*. Redwood City, CA: Addison-Wesley.

Sohnius M-F (1985) *Physics Reports* C 128: 39.

Weinberg S (1996–2000) *The Quantum Theory of Fields*, vols. I–III. Cambridge: Cambridge University Press.

Wess J and Bagger J (1992) *Supersymmetry and Supergravity*, Princeton Series in Physics, 2nd edn. Princeton: Princeton University Press.



Symmetry Breaking in Field Theory

T W B Kibble, Imperial College, London, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Spontaneous symmetry breaking in its simplest form occurs when there is a symmetry of a dynamical system that is not manifest in its ground state or equilibrium state. It is a common feature of many classical and quantum systems. In quantum field theories, in the infinite-volume limit, there are new features, the appearance of unitarily inequivalent representations of the canonical commutation relations, and the possibility of a true phase transition – a point in the phase space where the thermodynamic free energy is nonanalytic. The spontaneous breaking of a continuous global symmetry implies the existence of massless particles, the Goldstone bosons, while in the local-symmetry case some or all of these may be eliminated by the Higgs mechanism. Spontaneous symmetry breaking in gauge theories is however a more elusive concept.

Breaking of Global Symmetries

In a quantum-mechanical system a (time-independent) symmetry is represented by a unitary operator $\hat{U}$ acting on the Hilbert space of quantum states which commutes with the Hamiltonian $\hat{H}$. If the ground state $|0\rangle$ of the system in not invariant under $\hat{U}$, then $|0'\rangle = \hat{U}|0\rangle \neq c|0\rangle$ is also a ground state. In other words, the ground state is degenerate.

For a system with a finite number of degrees of freedom, whose states are represented by vectors in a separable Hilbert space $\mathcal{H}$, symmetry breaking of an abelian symmetry group G is impossible, unless there are additional accidental symmetries. Consider, for example, a particle in a double-well potential

$$V = \frac{m\omega^2}{4a^2}(x^2 - a^2)^2 \qquad [1]$$

which has the discrete symmetry group $G = \mathbb{Z}_2$; the inversion symmetry operator $\hat{U}$ satisfies $\hat{U}^2 = \hat{1}$. There are then two approximate ground states $|0\rangle$ and $|0'\rangle = \hat{U}|0\rangle$, with wave functions proportional to $\exp[-(1/2)\, m\omega(x \mp a)^2]$. However, there is an overlap between these, and the off-diagonal matrix element $\langle 0|\hat{H}|0'\rangle$ is nonzero, although exponentially small, so the true energy eigenstates are, approximately, $|0_\pm\rangle = (1/\sqrt{2})(|0\rangle \pm |0'\rangle)$. (More accurate energy eigenfunctions and eigenvalues may be found by using the WKB approximation.)

Of course, if the symmetry group is nonabelian, and the ground state belongs to a nontrivial representation, then degeneracy is unavoidable. For example, if G is the rotation group SO(3) (or SU(2))

Notes:

and the ground state has angular momentum $j \neq 0$, then it is $(2j+1)$-fold degenerate.

The situation is different, however, in a quantum field theory. In the infinite-volume limit, even abelian symmetries can be spontaneously broken. Take, for example, a real scalar field with Lagrangian

$$\mathcal{L} = \tfrac{1}{2}\partial_\mu\phi\partial^\mu\phi - V = \tfrac{1}{2}\dot{\phi}^2 - \tfrac{1}{2}(\nabla\phi)^2 - V \qquad [2]$$

(where we set $c = \hbar = 1$), again with a double-well potential

$$V = \tfrac{1}{8}\lambda(\phi^2 - \eta^2)^2 \qquad [3]$$

exhibiting a $\mathbb{Z}_2$ symmetry under which $\phi(x) \mapsto -\phi(x)$.

At least in the semiclassical or tree approximation, there are two degenerate vacuum states $|0\rangle$ and $|0'\rangle$, with

$$\langle 0|\hat{\phi}(x)|0\rangle \approx \eta \quad \text{and} \quad \langle 0'|\hat{\phi}(x)|0'\rangle \approx -\eta \qquad [4]$$

If we quantize the system in a box of finite volume $\mathcal{V}$, then, as earlier, there is an off-diagonal matrix element of the Hamiltonian connecting the two states, so the true ground state is (approximately) $(1/\sqrt{2})(|0\rangle + |0'\rangle)$. However, this matrix element goes to zero exponentially as $\mathcal{V} \to 0$. Even for large but finite volume, the rate of transitions from $|0\rangle$ to $|0'\rangle$ is exponentially slow.

Similarly, we can consider a complex scalar field theory with a sombrero potential:

$$\begin{aligned} \mathcal{L} &= |\dot{\phi}|^2 - |\nabla\phi|^2 - V \\ V &= \tfrac{1}{2}\lambda\left(|\phi|^2 - \tfrac{1}{2}\eta^2\right)^2 \end{aligned} \qquad [5]$$

This model is invariant under the U(1) group of phase transformations, $\phi(x) \mapsto \phi(x)e^{i\alpha}$, so we now have a continuously infinite set of degenerate vacuum states $|0_\alpha\rangle$ labeled by an angle α, and satisfying

$$\langle 0_\alpha|\hat{\phi}(x)|0_\alpha\rangle \approx \frac{1}{\sqrt{2}}\eta e^{i\alpha} \qquad [6]$$

Once again, one finds that in the infinite-volume limit there are no matrix elements connecting the different vacuum states. Moreover, in this limit no polynomial formed from the field operators $\hat{\phi}(x)$ in a finite volume can have nonzero matrix elements between $|0_\alpha\rangle$ and $|0_\beta\rangle$ for $\alpha \neq \beta$. Applying the operators $\hat{\phi}(x)$ to any one of these vacuum states $|0_\alpha\rangle$, we can construct a Fock space $\mathcal{H}_\alpha$, and the representations of the canonical commutation relations on these separate Hilbert spaces are unitarily inequivalent. Formally, we can introduce operators $\hat{U}_\alpha$ that perform the symmetry transformations:

$$\hat{U}_\alpha\hat{\phi}(x)\hat{U}_\alpha^{-1} = \hat{\phi}(x)e^{i\alpha} \qquad [7]$$

However, these are not unitary operators on the spaces $\mathcal{H}_\beta$, but rather maps from one space to another: $\hat{U}_\alpha : \mathcal{H}_\beta \to \mathcal{H}_{\alpha+\beta}$ – or, alternatively, operators on the nonseparable Hilbert space $\mathcal{H} = \bigoplus_\alpha \mathcal{H}_\alpha$.

So far, our discussion has been restricted to the tree approximation. For a full quantum treatment, $V(\phi)$ must be replaced by the effective potential $V_{\text{eff}}(\phi)$, which may be defined as the minimum value of the mean energy density in all states in which the field $\hat{\phi}$ has the uniform expectation value $\langle\hat{\phi}(x)\rangle = \phi$. V_{eff} may be computed by summing vacuum loop diagrams.

A point to note is that although the degenerate vacua $|0_\alpha\rangle$ are mathematically distinct, in the absence of any external definition of phase, they are physically identical. There is no internal observational test that will distinguish them.

Symmetry-Breaking Phase Transitions

Spontaneous symmetry breaking often occurs in the context of a phase transition. At high temperature, $T \gg \eta$, there are large fluctuations in ϕ and the central hump of the potential is unimportant. Then the equilibrium state is symmetric, with $\langle\hat{\phi}\rangle = 0$. However, as the temperature falls, it becomes less probable that the field will fluctuate over the top of the hump. It will tend to fall into the trough, and acquire a nonzero average value $\langle\hat{\phi}\rangle$ – the order parameter for the phase transition – thus breaking the symmetry. The direction of symmetry breaking (e.g., the phase of ϕ in the U(1) model) is random, determined in practice by small preexisting fluctuations or interactions with the environment.

One way of studying this process is to compute the temperature-dependent effective potential $V_{\text{eff}}(\phi, T)$. In the one-loop approximation, at high temperature, the leading corrections to the zero-temperature effective potential $V_{\text{eff}}(\phi, T)$ are of the form

$$\begin{aligned} V_{\text{eff}}(\phi, T) = V_{\text{eff}}(\phi, 0) &- \frac{\pi^2}{90}N_* T^4 \\ &+ \frac{1}{24}M_*^2(\phi)T^2 + \mathcal{O}(T) \end{aligned} \qquad [8]$$

where N_* is the total number of helicity states of light particles (those with masses $\ll T$), and M_*^2, which depends on ϕ, is the sum of their squared masses. (Fermions if present contribute to N_* with a factor of 7/8 and to M_*^2 with a factor of 1/2.) In the simplest case, where we have only a multiplet $\boldsymbol{\phi} = (\phi_a)_{a=1,\ldots,N}$ of real scalar fields, $N_* = N$ and $M_*^2 = M_{aa}^2$

(summation over a implied), where the mass-squared matrix is

$$M_{ab}^2 = \frac{\partial^2 V}{\partial\phi_a \partial\phi_b} \qquad [9]$$

For example, in an O(N) theory, with $V = (1/8)\lambda(\boldsymbol{\phi}^2 - \eta^2)^2$, where $\boldsymbol{\phi}^2 = \phi_a\phi_a$, one has

$$M_{ab}^2 = \tfrac{1}{2}\lambda(\boldsymbol{\phi}^2 - \eta^2)\delta_{ab} + \lambda\phi_a\phi_b \qquad [10]$$

whence

$$V_{\text{eff}}(\boldsymbol{\phi}, T) \approx \frac{1}{8}\lambda(\boldsymbol{\phi}^2 - \eta^2)^2 - \frac{\pi^2}{90}NT^4 + \frac{1}{48}\lambda T^2[(N+2)\boldsymbol{\phi}^2 - N\eta^2] \qquad [11]$$

It is then easy to see that the minimum occurs at $\boldsymbol{\phi} = 0$ for $T > T_c$, where in this approximation $T_c{}^2 = 12\eta^2/(N+2)$, while below the critical temperature the minimum is at

$$\boldsymbol{\phi}^2 = \phi_{\text{eq}}^2(T) \approx \eta^2 - \frac{N+2}{12}T^2 \qquad [12]$$

As $T \to 0$, the equilibrium state approaches one of the vacuum states $|0_{\boldsymbol{n}}\rangle$, labeled by an N-dimensional unit vector $\boldsymbol{n}$, such that $\langle 0_{\boldsymbol{n}}|\hat{\boldsymbol{\phi}}|0_{\boldsymbol{n}}\rangle = \eta\boldsymbol{n}$.

It is often convenient to introduce a classical symmetry-breaking potential. For example, in the O(N) model, we may take $V_{\text{sb}} = -\boldsymbol{j}\cdot\boldsymbol{\phi}(x)$, where $\boldsymbol{j}$ is a constant N-vector. This has the effect of tilting the potential, thus removing the degeneracy. A characteristic of spontaneous symmetry breaking is that the limits $\boldsymbol{j} \to 0$ and $\mathcal{V} \to \infty$ do not commute. If (for $T < T_c$) we take the infinite-volume limit first, and then let $\boldsymbol{j} \to 0$, we get different equilibrium states, depending on the direction from which $\boldsymbol{j}$ approaches zero; if we fix $\boldsymbol{n}$ and let $\boldsymbol{j} = j\boldsymbol{n}$, $j \to 0$, then we find

$$\lim_{j\to 0}\lim_{\mathcal{V}\to\infty}\langle\hat{\boldsymbol{\phi}}(x)\rangle_{j\boldsymbol{n}} = \phi_{\text{eq}}(T)\boldsymbol{n} \qquad [13]$$

We may also regard $\boldsymbol{j}$ as representing an interaction with the external environment (e.g., other fields). If such a term is present during the cooling of the system through the phase transition, it will constrain the direction of the spontaneous symmetry breaking. Note that one always arrives in this way at one of the degenerate vacua $|0_{\boldsymbol{n}}\rangle$, not a linear combination of them.

Goldstone Bosons

The Goldstone theorem states that spontaneous breaking of any continuous global symmetry leads inevitably (except, as we discuss later, in the presence of long-range forces) to the appearance of massless modes – the Goldstone bosons.

The proof is straightforward. Associated with any continuous symmetry there is a Noether current satisfying the continuity equation $\partial_\mu\hat{j}^\mu = 0$ and such that infinitesimal symmetry transformations are generated by the spatial integral of $\hat{j}^0$. The fact that the symmetry is broken means that there is some scalar field $\hat{\phi}(x)$ whose vacuum expectation value $\langle 0|\hat{\phi}(0)|0\rangle$ is not invariant under the symmetry transformation. Hence,

$$\lim_{\mathcal{V}\to 0} \mathrm{i}\int_{\mathcal{V}} \mathrm{d}^3x\langle 0|[\hat{j}^0(x), \hat{\phi}(0)]|0\rangle|_{x^0=0} \neq 0 \qquad [14]$$

Moreover, the time derivative of this integral is

$$\lim_{\mathcal{V}\to 0} \mathrm{i}\int_{\mathcal{V}} \mathrm{d}^3x\langle 0|[\partial_0\hat{j}^0(x), \hat{\phi}(0)]|0\rangle|_{x^0=0} = -\lim_{\mathcal{V}\to 0} \mathrm{i}\int_{\partial\mathcal{V}} \mathrm{d}S_k\langle 0|[\hat{j}^k(x), \hat{\phi}(0)]|0\rangle|_{x^0=0} = 0 \qquad [15]$$

where $\partial\mathcal{V}$ is the bounding surface of $\mathcal{V}$. This vanishes because the surface integral is zero – in a relativistic theory, because the commutator vanishes at spacelike separation, and more generally in the absence of long-range interactions because it tends rapidly to zero at large spatial separation.

Now, inserting a complete set of momentum eigenstates $|n, \boldsymbol{p}\rangle$ in [14], we can see that there must exist states such that $\langle n, \boldsymbol{p}|\hat{\phi}(0)|0\rangle \neq 0$, with $p^0 \to 0$ in the limit $|\boldsymbol{p}| \to 0$, that is, massless modes.

One can see this more directly in the U(1) model above. Consider a vacuum state $|0\rangle$ such that $\langle 0|\hat{\phi}|0\rangle = \eta/\sqrt{2}$ is real. Then it is useful to shift the origin of ϕ by writing

$$\phi(x) = \frac{1}{\sqrt{2}}[\eta + \varphi_1(x) + \mathrm{i}\varphi_2(x)] \qquad [16]$$

where φ_1 and φ_2 are real. Then the Lagrangian becomes

$$\mathcal{L} = \tfrac{1}{2}\Big[\dot{\varphi}_1^2 - (\nabla\varphi_1)^2 + \dot{\varphi}_2^2 - (\nabla\varphi_2)^2 - \lambda\eta^2\varphi_1^2 - \lambda\eta\varphi_1(\varphi_1^2 + \varphi_2^2) - \tfrac{1}{4}\lambda(\varphi_1^2 + \varphi_2^2)^2\Big] \qquad [17]$$

Evidently, the field φ_1, corresponding to radial oscillations in ϕ, is massive, with mass $\sqrt{\lambda}\eta$. But there is no term in φ_2^2, so φ_2 is massless.

In the case of spontaneous symmetry breaking of nonabelian symmetries, there may be several Goldstone bosons, one for each broken component of the continuous symmetry. In our theory with symmetry group $G = \mathrm{O}(N)$, the possible values of the vacuum expectation value at $T = 0$ are $\langle 0_{\boldsymbol{n}}|\hat{\phi}(0)|0_{\boldsymbol{n}}\rangle \approx \eta\boldsymbol{n}$,

where $\boldsymbol{n}$ is an arbitrary unit vector. In this case, for given $\boldsymbol{n}$, there is an unbroken symmetry subgroup

$$H = \{R \in \mathrm{O}(N) : R\boldsymbol{n} = \boldsymbol{n}\} = \mathrm{O}(N-1) \quad [18]$$

and the number of broken symmetries is

$$\dim G - \dim H = N - 1 \quad [19]$$

Thus, the radial component of ϕ is massive, and there are $N-1$ Goldstone bosons, the $N-1$ transverse components.

Spontaneously Broken Gauge Theories

As we shall see, symmetry breaking in gauge theories is a more problematic concept but, for the moment, these complications are ignored and the present discussion will continue with an approach similar to that used above.

The simplest local gauge symmetry theory is a U(1) Higgs model, a model of a complex scalar field $\phi(x)$ interacting with a gauge potential $A_\mu(x)$, described by the Lagrangian

$$\mathcal{L} = D_\mu\phi^* D^\mu\phi - \tfrac{1}{4}F_{\mu\nu}F^{\mu\nu} - V(|\phi|) \quad [20]$$

where V is a sombrero potential as in [5], while the covariant derivative $D_\mu\phi$ and gauge field $F_{\mu\nu}$ are given by

$$D_\mu\phi = \partial_\mu\phi + ieA_\mu\phi, \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \quad [21]$$

The model is invariant under the local U(1) gauge transformations

$$\begin{aligned} \phi(x) &\mapsto \phi(x)\mathrm{e}^{\mathrm{i}\alpha(x)} \\ A_\mu(x) &\mapsto A_\mu(x) - \frac{1}{e}\partial_\mu\alpha(x) \end{aligned} \quad [22]$$

The Goldstone theorem does not apply to local-symmetry theories. The problem is that to have a Hilbert space containing only physical states one must eliminate the gauge freedom by choosing a gauge condition (e.g., in the U(1) case the Coulomb gauge $\partial_k A^k(x) = 0$, which has the effect of restricting the number of polarization states of photons to two). This necessarily breaks manifest Lorentz invariance, although the theory is, of course, still fully Lorentz invariant. The proof of the theorem fails because the current is no longer local; the long-range Coulomb interaction makes the commutator fall off only like $1/r^2$, so the surface integral no longer vanishes in the infinite-volume limit. (The theorem also fails for nonrelativistic models with long-range forces.)

Again, consider a vacuum state $|0\rangle$ in which $\langle 0|\hat{\phi}|0\rangle = \eta/\sqrt{2}$, and make the same decomposition, [16]. Then, if we set

$$A'_\mu = A_\mu + \frac{1}{e\eta}\partial_\mu\varphi_2 \quad [23]$$

we find that the kinetic term for φ_2 has been absorbed into a mass term $(1/2)e^2\eta^2 A'_\mu A'^\mu$ for the vector field. We have a model with only massive fields: the "Higgs field" φ_1 with mass $\sqrt{\lambda}\eta$ and the gauge field A'_μ with mass $e\eta$. The Goldstone bosons have been "eaten up" by the vector field to provide its longitudinal mode. This is the Higgs mechanism, first noted by Anderson in the context of the photon in a plasma becoming a massive plasmon.

A more elegant way of seeing this is to note that we can always make a gauge transformation to ensure that ϕ is real (at least so long as $\phi \neq 0$; where it is zero, there may be problems). This means that $\phi(x) = (1/\sqrt{2})(\eta + \varphi_1)$; φ_2 disappears altogether, and its kinetic term reduces to $(1/2)e^2 A_\mu A^\mu(\eta + \varphi_1)^2$, which includes the mass term for A_μ as well as cubic and quartic interaction terms.

As before, the discussion can be generalized to nonabelian theories, although there are additional problems to be discussed later. If we have a local symmetry group G that breaks spontaneously to leave an unbroken subgroup H, then the gauge fields associated with H remain massless. Each of the (dim G − dim H) complementary fields "eats up" one of the Goldstone bosons, becoming massive in the process. We are left only with other, "radial" components of $\boldsymbol{\phi}$, the massive Higgs fields.

Consider, for example, a local SO(3) model, with scalar fields $\boldsymbol{\phi} = (\phi_a)_{a=1,2,3}$ and gauge potentials $\boldsymbol{A}_\mu = (A_{a\mu})$. The infinitesimal gauge transformations are

$$\delta\boldsymbol{\phi} = \delta\boldsymbol{\omega} \times \boldsymbol{\phi}, \qquad \delta\boldsymbol{A}_\mu = \delta\boldsymbol{\omega} \times \boldsymbol{A} - \frac{1}{e}\partial_\mu\delta\boldsymbol{\omega} \quad [24]$$

where $\delta\boldsymbol{\omega}$ is the gauge parameter. The Lagrangian is

$$\mathcal{L} = \tfrac{1}{2}D_\mu\boldsymbol{\phi} \cdot D^\mu\boldsymbol{\phi} - \tfrac{1}{4}\boldsymbol{F}_{\mu\nu} \cdot \boldsymbol{F}^{\mu\nu} - \tfrac{1}{8}\lambda(\boldsymbol{\phi}^2 - \eta^2)^2 \quad [25]$$

where the covariant derivative and gauge field are

$$\begin{aligned} D_\mu\boldsymbol{\phi} &= \partial_\mu\boldsymbol{\phi} + e\boldsymbol{A}_\mu \times \boldsymbol{\phi} \\ \boldsymbol{F}_{\mu\nu} &= \partial_\mu\boldsymbol{A}_\nu - \partial_\nu\boldsymbol{A}_\mu + e\boldsymbol{A}_\mu \times \boldsymbol{A}_\nu \end{aligned} \quad [26]$$

If we take $\langle\hat{\boldsymbol{\phi}}\rangle$ in the 3-direction, the fields $A_{1\mu}$ and $A_{2\mu}$ absorb the Goldstone fields ϕ_1, ϕ_2 to become massive. As in the abelian case, we can use the local SO(3) invariance to rotate $\boldsymbol{\phi}$ everywhere to the 3-direction, and write $\boldsymbol{\phi} = (0, 0, \eta + \varphi_3)$. In this gauge the kinetic term $(1/2)(e\boldsymbol{A}_\mu \times \boldsymbol{\phi})^2$ gives a mass $e\eta$ to the fields $A_{1\mu}, A_{2\mu}$ while $A_{3\mu}$ remains massless, and the Higgs field φ_3 again has mass $\sqrt{\lambda}\eta$.

Elitzur's Theorem; the Role of Gauge Fixing

The concept of spontaneous symmetry breaking in the context of a local symmetry requires further discussion, in particular because of Elitzur's theorem, proved in 1975, which states in essence that "spontaneous breaking of a local symmetry is impossible." In the light of this theorem, it may seem that a "spontaneously broken gauge theory" is an oxymoron. In fact, it means something rather different, although even that is not unproblematic.

The theorem was proved in the context of lattice gauge theory, where the spatial continuum is replaced by a discrete lattice. The scalar field is then represented by values $\boldsymbol{\phi}_x$ at each lattice site, and the gauge potential by values $A_{x,\mu}$ on the links of the lattice. This is significant because on the lattice one can use a manifestly gauge-invariant formalism. Expectation values of gauge-invariant physical variables can be found, for example, by a Monte Carlo algorithm that effectively averages over all possible gauges. In this context, it is possible to show that the expectation value of any gauge-noninvariant operator (such as $\hat{\boldsymbol{\phi}}_x$) necessarily vanishes identically.

To be more specific, suppose we incorporate a symmetry-breaking term of the form $-\boldsymbol{j} \cdot \sum_x \boldsymbol{\phi}_x$, and consider the limits $\mathcal{V} \to \infty$ followed by $j \to 0$. In the global-symmetry case, as we noted earlier, this yields the nonzero result [13]. However, in the case of a local gauge symmetry, one can show rigorously that

$$\lim_{j \to 0} \lim_{\mathcal{V} \to \infty} \langle \hat{\boldsymbol{\phi}}_x \rangle_{j\mathcal{V}} = 0 \qquad [27]$$

The essential reason for this is that we can make a gauge transformation in the neighborhood of the point x to make $\boldsymbol{\phi}_x$ have any value we like without changing the energy by more than a very small amount that goes to zero as $j \to 0$. Within this manifestly gauge-invariant formalism, it is clear that the expectation value of a gauge-noninvariant operator such as $\hat{\boldsymbol{\phi}}$ is not an appropriate order parameter. One must instead look for a gauge-invariant order parameter.

It is important to note, however, that this result applies only in the context of a manifestly gauge-invariant formalism. But, in general, gauge theories cannot be quantized in a manifestly gauge-invariant way. In a path-integral formalism, the action functional, which appears in the exponent, is constant along the orbits of the gauge-group action. Consequently, the integral contains an infinite factor, the volume of the (infinite-dimensional) gauge group. There are corresponding divergences in the perturbation series. As is well known, this problem can be dealt with by introducing a gauge-fixing term, which explicitly breaks the gauge symmetry, and renders Elitzur's theorem inapplicable. But this procedure leaves a global symmetry unbroken, and it is in fact that global symmetry that is broken spontaneously.

One example is the Landau–Ginzburg model of a superconductor, which is essentially just the nonrelativistic limit of the abelian Higgs model, although there is one significant difference: here the field $\hat{\phi}$ annihilates a Cooper pair, a bound pair of electrons with equal and opposite momenta and spins, so e above is replaced by the charge $2e$ of a Cooper pair. The appearance of a condensate of Cooper pairs in the low-temperature superconducting phase corresponds to a state in which $\langle \hat{\phi} \rangle$ is nonzero. This would not be possible without fixing a gauge. In the nonrelativistic context, the obvious gauge to choose is the Coulomb gauge, defined by the condition $\partial_k A^k = 0$. This gauge-fixing condition breaks the local symmetry explicitly, but it leaves unbroken the global symmetry $\phi(x) \to \phi(x)e^{i\alpha}$ with constant α. It is that global symmetry that is spontaneously broken when $\langle \hat{\phi} \rangle \neq 0$.

For a model with nonabelian local symmetry the standard procedure used to derive a perturbation expansion is that of Faddeev and Popov. Consider, for example, the SO(3) gauge theory discussed in the preceding section. To fix the gauge, we can choose a set of functions $F = (F_a)$ of the fields, and introduce into the path integral a gauge-fixing term of the form

$$\mathcal{L}_{\text{gf}} = -\frac{1}{2\xi} \mathbf{F}^2 \qquad [28]$$

where ξ is an arbitrary real constant. However, to ensure that this does not bias the integral, so that the gauge-fixed theory is at least formally equivalent to the original gauge-invariant theory, one must also include the determinant of the Jacobian matrix

$$J_{ab}(x, y) = \frac{\delta F_a(x)}{\delta \omega_b(y)} \qquad [29]$$

The easiest way to do this is to introduce Faddeev–Popov ghost fields $\bar{C}$, C, which are scalar Grassmann variables, and an appropriate term in the Lagrangian

$$\mathcal{L}_{\text{FP}} = \bar{C} \cdot J \cdot C \qquad [30]$$

For the SO(3) model, a convenient choice of gauge is the R_ξ gauge defined by

$$F = \partial_\mu A^\mu - \xi e \eta \boldsymbol{n} \times \boldsymbol{\phi} \qquad [31]$$

where $\boldsymbol{n}$ is an arbitrarily chosen unit vector. It is clear that the full Lagrangian $\mathcal{L}+\mathcal{L}_{\text{gf}}+\mathcal{L}_{\text{FP}}$ is no longer invariant under the full SO(3) gauge group, although there is a residual U(1) gauge invariance corresponding to rotations about $\boldsymbol{n}$. In this gauge, the arbitrary choice of $\boldsymbol{n}$ means that the global SO(3) symmetry is also broken. However, for other choices, such as the Lorentz gauge $F=\partial_\mu A^\mu$ or axial gauge $F=A_3$, the Lagrangian *is* invariant under global SO(3) rotations of all the fields. This global symmetry is then spontaneously broken, with $\hat{\boldsymbol{\phi}}$ acquiring as before a nonzero expectation value of the form $\langle\hat{\boldsymbol{\phi}}(x)\rangle=\eta\boldsymbol{n}$.

It is interesting to look again at the particle content of this model. By setting $\boldsymbol{\phi}(x)=\eta\boldsymbol{n}+\boldsymbol{\varphi}(x)$ with $\boldsymbol{n}=(0,0,1)$, one finds that in the quadratic part of the Lagrangian, the cross-terms between A_μ and $\boldsymbol{\varphi}$ combine to form a total divergence which can be dropped. As before, φ_3 is the Higgs field, with $m^2=\lambda\eta^2$, $A_{3\mu}$ is the massless gauge field corresponding to the unbroken gauge symmetry, and the three transverse components of $A_{1\mu}$ and $A_{2\mu}$ represent the massive vector fields, with $m^2=e^2\eta^2$. There are, however, also unphysical fields with ξ-dependent masses: $\varphi_{1,2}, C_{1,2}, \bar{C}_{1,2}$, and the longitudinal components $\partial_\mu A^\mu_{1,2}$ all have $m^2=\xi e^2\eta^2$. We can now compute the effective potential $V_{\text{eff}}(T,\boldsymbol{\varphi})$. One point that should be noted in performing this calculation is that the ghost fields $C,\bar{C}$ contribute negatively. Obviously, V_{eff}, being ξ-dependent, is not itself physically meaningful. Nevertheless, it can be shown that the stationary points of V_{eff} are physical, and correspond to the possible equilibrium states of the theory. Moreover, the extremal values of V_{eff} are independent of ξ and give correctly the thermodynamic potential in the corresponding equilibrium states. The negative contributions from the ghost fields to N_* and M_*^2 ensure that the ξ dependence cancels out, and we find as expected $N_*=9$ and $M_*^2=(\lambda+6e^2)\eta^2$.

Phase Transitions and Crossovers

Our discussion so far has for the most part been restricted to a semiclassical or mean-field approximation. It is important to bear in mind, however, that this approximation does not suffice to determine whether a phase transition (where the thermodynamic free energy is nonanalytic) exists, or what its nature is. Determining the detailed characteristics of phase transitions requires other methods, such as the renormalization group or lattice simulations. In many cases, it is far from trivial to establish the order of the transition, or even whether a true phase transition actually exists.

Gauge theories pose particular problems because of the infrared divergences in the thermal field theory at high temperature, and because in asymptotically free nonabelian theories the coupling becomes large at very low energy. Even when they appear to exhibit spontaneous symmetry breaking, they do not necessarily undergo a true phase transition. Lattice gauge theory calculations have led to the conclusion that in nonabelian gauge theories with the Higgs field in the fundamental representation, there are values of the coupling constants for which there is no phase transition, only a rapid but smooth crossover from one type of behavior to another, so that the high- and low-temperature phases are analytically connected. If the coupling constant is small, there is a first-order phase transition, and for moderate values the theory exhibits a very rapid crossover that looks quite similar to a symmetry-breaking phase transition. Nevertheless, the analytic connection between the two phases implies that there cannot exist an order parameter that is strictly zero above the transition and nonzero below it.

In particular, it appears that for physical values of the Higgs mass, the electroweak theory does not undergo in fact undergo a true phase transition. It is somewhat ironic that the most famous example of a spontaneously broken gauge theory probably does not, strictly speaking, exhibit a symmetry-breaking phase transition!

Conclusions

We have discussed the main features of spontaneous symmetry breaking in both the global- and local-symmetry cases, especially the appearance of Goldstone bosons when a continuous global symmetry breaks, and their elimination in the local-symmetry case by the Higgs mechanism, as well as the problems attaching to the concept of spontaneous symmetry breaking in gauge theories.

See also: Abelian Higgs Vortices; Effective Field Theories; Electroweak Theory; Finite Group Symmetry Breaking; Lattice Gauge Theory; Noncommutative Geometry and the Standard Model; Phase Transitions in Continuous Systems; Quantum Central Limit Theorems; Quantum Spin Systems; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Topological Defects and their Homotopy Classification.

Further Reading

Anderson PW (1963) Plasmons, gauge invariance and mass. *Physical Review* 130: 439–442.

Coleman S (1985) *Aspects of Symmetry*, ch. 5. Cambridge: Cambridge University Press.

Elitzur S (1975) Impossibility of spontaneously breaking local symmetries. *Physical Review D* **12**: 3978–3982.

Englert F and Brout R (1964) Broken symmetry and the mass of gauge vector bosons. *Physical Review Letters* **13**: 321–323.

Fradkin E and Shenker SH (1979) Phase diagrams of lattice gauge theories with Higgs fields. *Physical Review D* **19**: 3682–3697.

Goldstone J, Salam A, and Weinberg S (1962) Broken symmetries. *Physical Review* **127**: 965–970.

Guralnik GS, Hagen CR, and Kibble TWB (1964) Global conservation laws and massless particles. *Physical Review Letters* **13**: 585–587.

Higgs PW (1964) Broken symmetries, massless particles and gauge fields. *Physics Letters* **12**: 132–133.

Kibble TWB (1967) Symmetry-breaking in non-abelian gauge theories. *Physical Review* **155**: 1554–1561.

Weinberg S (1996) *The Quantum Theory of Fields, Vol. II. Modern Applications*, chs. 19 and 21. Cambridge: Cambridge University Press.

Twistor Theory: Some Applications

L Mason, University of Oxford, Oxford, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Roger Penrose introduced twistor theory as a geometrical framework for basic physics in order to unify quantum theory and gravity. This program has had many successes along the way, but the long-term goals of reformulating and superceding the established theories of basic physics are still a long way from being fulfilled. Nevertheless, the successes have had many important applications across mathematics and mathematical physics. This article will concentrate on three areas of application: integrable systems, geometry, and perturbative gauge theory (via twistor-string theory). It is intended to be self-contained as far as possible, but the reader may well find it easier to first read the article Twistors.

Notes:

Twistor Theory

A basic motivation of twistor theory is to bring out the complex (holomorphic) geometry that underlies real spacetime. In general relativity, a spacetime is a 4-manifold with metric g of signature $(1,3)$, and when it is flat, that is, $g = dt^2 - dx^2 - dy^2 - dz^2$, where (t,x,y,z) are coordinates on $\mathbb{R}^4$, it is called Minkowski space. The first appearance of a complex structure arises from the fact that, at a given event, the celestial sphere of light rays (directions of zero length with respect to g) naturally has the structure of the Riemann sphere, $\mathbb{CP}^1$, in such a way that Lorentz transformations (linear transformations of the tangent space preserving the metric) act on this sphere by Möbius transformations. These are the maximal group of complex analytic transformations of $\mathbb{CP}^1$.

Twistor space extends this idea to the whole of Minkowski space. Denoted $\mathbb{PT}$, the twistor space for Minkowski space is complex projective 3-space, $\mathbb{CP}^3$, the space of one-dimensional subspaces of $\mathbb{C}^4$; it is a three-dimensional complex manifold obtained by adding a "plane at infinity" to $\mathbb{C}^3$. Explicitly, we can introduce homogeneous coordinates $Z^\alpha \in \mathbb{C}^4 - \{0\}$ with $\alpha = 0,1,2,3$ but where $Z^\alpha \sim \lambda Z^\alpha$ for $\lambda \in \mathbb{C} - \{0\}$. Affine coordinates on a $\mathbb{C}^3$ chart $Z^3 \neq 0$ can be obtained by setting $(z_1, z_2, \lambda) = (Z^0/Z^3, Z^1/Z^3, Z^2/Z^3)$. Physically, points of twistor space correspond to spinning massless particles in Minkowski space. Mathematically, the correspondence can be understood as the Klein correspondence.

The Klein Correspondence

The correspondence between $\mathbb{PT}$ and Minkowski space can be extended first to complexified Minkowski space so that the coordinates are allowed to take on values in $\mathbb{C}$, and then to its conformal compactification by including the "light cone at infinity." It then coincides with the classical complex Klein correspondence. The Klein correspondence is the one-to-one correspondence between lines in $\mathbb{CP}^3$ and points of a four complex-dimensional quadric, $\mathbb{CM}$, in $\mathbb{CP}^5$. The 4-quadric $\mathbb{CM}$ can be understood as conformally compactified complexified Minkowski space. Introducing affine coordinates (z_1, z_2, λ) on $\mathbb{PT}$ and (t,x,y,z) on $\mathbb{CM}$, we find that a point (t,x,y,z) in $\mathbb{CM}$ corresponds to a line in $\mathbb{PT}$ according to

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} t-z & x+iy \\ x-iy & t+z \end{pmatrix} \begin{pmatrix} 1 \\ \lambda \end{pmatrix}$$

Alternatively, fixing (λ, z_1, z_2) in these equations gives a 2-plane in complex Minkowski space corresponding to all the lines in $\mathbb{PT}$ through (λ, z_1, z_2). Such 2-planes are called "α-planes." They are totally null (i.e., the tangent vectors not only have zero length but are also mutually orthogonal) and also self-dual (under the differential geometer's notion of Hodge duality).

This complex correspondence can also be restricted to give correspondences for $\mathbb{R}^4$ with metrics of positive-definite, Euclidean, signature or ultrahyperbolic, $(2,2)$, signature. A particular simplification in Euclidean signature is that the complex α-planes intersect the real slice in a point. The conformal compactification of Euclidean $\mathbb{R}^4$ is the 4-sphere S^4 given by adding a single point at infinity, and so we have a projection $p: \mathbb{PT} \to S^4$ whose fibers are holomorphically embedded $\mathbb{CP}^1$s. These fibers can be characterized as the lines in $\mathbb{PT}$ that are invariant under a quaternionic complex conjugation which is an antiholomorpic map $\hat{}: \mathbb{PT} \to \mathbb{PT}$ with no fixed points. (Here quaternionic means that on the nonprojective twistor space, $\mathbb{T} = \mathbb{C}^4$, the conjugation has the property $\hat{\hat{Z}} = -Z$ so that it defines a second complex structure anticommuting with the standard one; this is sufficient to express $\mathbb{T} = \mathbb{Q}^2$, where $\mathbb{Q}$ denotes the quaternions. The complex structures i, j, and k of the quaternions are given by identifying i with $\sqrt{-1}$ on $\mathbb{C}^4$ and j with $\hat{}$ and $k = ij$.)

The Penrose Transform

A basic task of twistor theory is to transform solutions to the field equations of mathematical physics into objects on twistor space. This works well for linear massless fields such as the Weyl neutrino equation, Maxwell's equations for electromagnetism and linearized gravity. In its general form, this transform has become known as the Penrose transform. Such fields correspond to freely prescribable holomorphic functions $f(\lambda, z_1, z_2)$ (or, more precisely, analytic cohomology classes) on regions of twistor space. The field can be obtained from this function by means of a contour integral. The simplest of these integral formulas is

$$\phi(x^a) = \oint f(\lambda, t - z + \lambda(x + iy), x - iy + \lambda(t + z))d\lambda$$

and differentiation under the integral sign leads easily to the fact that ϕ satisfies the wave equation

$$\frac{\partial^2\phi}{\partial t^2} - \frac{\partial^2\phi}{\partial x^2} - \frac{\partial^2\phi}{\partial y^2} - \frac{\partial^2\phi}{\partial z^2} = 0$$

This formula was originally discovered by Bateman. Note that f must have singularities on twistor space to yield a nontrivial ϕ and even then, there are many choices of f that yield zero. For a solution ϕ defined

over a region U in spacetime, the function f is correctly understood as a representative of a Cech cohomology class defined on the region U' in twistor space swept out by the lines corresponding to points of U. Furthermore, the function f should be taken globally to be a function of homogeneity -2, $f(\lambda Z^\alpha) = \lambda^{-2} f(Z^\alpha)$. This formula has generalizations to massless fields of all helicities in which a field of helicity s corresponds to a function (Cech cocycle) of homogeneity degree $2s - 2$.

The Penrose transform has found important applications in representation theory and integral geometry. For a review, the reader is referred to Baston and Eastwood (1989), the relevant survey articles in Bailey and Baston (1990), or Mason and Hughston (1990, chapter 1).

Twistor Theory and Nonlinear Equations

The Penrose transform for the Maxwell equations and linearized gravity turns out to be linearizations of correspondences for the nonlinear analogs of these equations: the Einstein vacuum equations and the Yang–Mills equations. However, the constructions only work when these fields are anti-self-dual. This is the condition that the curvature 2-forms satisfy $F^* = -iF$, where $*$ denotes the Hodge dual (which, up to certain factors of $\pm i$, has the effect of interchanging electric and magnetic fields); it is a nonlinear generalization of the right-handed circular polarization condition. Explicitly, in terms of spacetime indices $a, b, \ldots = 0, 1, 2, 3$, $F^*_{ab} = (1/2)\varepsilon_{abcd}F^{cd}$, where $\varepsilon_{0123} = 1$ and $\varepsilon_{abcd} = \varepsilon_{[abcd]}$. In Minkowski signature, the i factor in the anti-self-duality condition implies that real fields cannot be anti-self-dual. Thus, these extensions are not sufficient to fulfill the ambitions of twistor theory to incorporate real classical nonlinear physics in Minkowski space. However, the factor of i is not present in Euclidean and ultrahyperbolic signature, so the anti-self-duality condition is consistent with real fields in these signatures and this is where the main applications of these constructions have been.

The Nonlinear Graviton Construction and Its Generalizations

The first nonlinear twistor construction was due to Penrose (1976), and was inspired by Newman's (1976) construction of "heavens" from the infinities of asymptotically flat spacetimes in general relativity.

The nonlinear graviton construction proceeds from the definition of twistors in flat spacetime as α-planes in complexified Minkowski space. It is natural to ask which complexified metrics admit a full family of α-surfaces, that is, 2-surfaces that are totally null and self-dual. The answer is that a full family of α-surfaces exists iff the conformally invariant part of the curvature tensor, the Weyl tensor, is anti-self-dual. If this is the case, twistor space can be defined to be the (necessarily three-dimensional) space of such α-surfaces.

The remarkable fact is that the twistor space, together with its complex structure, is sufficient to determine the original spacetime. Twistor space is again a three-dimensional complex manifold, and contains holomorphically embedded rational curves, $\mathbb{CP}^1$s, at least one for each point of the spacetime. However, holomorphic rigidity implies that the family of rational curves is precisely four-dimensional over the complex numbers. Furthermore, incidence of a pair of curves can be taken to imply that the corresponding points in spacetime lie on a null geodesic and this yields a conformal structure on spacetime. Further structures on twistor space can be imposed to give the complex spacetime a metric that is vacuum, perhaps with a cosmological constant. The correspondence is stable under small deformations and so the data defining the twistor space is effectively freely prescribable, see Penrose (1976).

In Euclidean signature, again the complex α-planes intersect the real spacetime in a point, so the twistor space again fibers over spacetime. The twistor fibration can be constructed as the projectivized bundle of self-dual spinors or more commonly as the unit sphere bundle in the space of self-dual 2-forms (Atiyah *et al.* 1978). In the latter formulation, the complex structure on the twistor space arises from the direct sum of the naturally defined complex structures on the horizontal and vertical tangent spaces to the bundle; that on the vertical subspace is the standard one on the sphere, and that on the horizontal subspace is a multiple of the self-dual 2-form at the given point of the fiber.

There are now large families of extensions, generalizations, and reductions of this construction. They are all based on the idea of realizing a space with a given complexified geometric structure as the parameter space of a family of holomorphically embedded submanifolds inside a twistor space. In general, the most useful of these constructions are those in which the "spacetime" is obtained as the space of rational curves in a twistor space. This is because the equations that are solved on the corresponding spacetime can be thought of as a completely integrable system in which the integrability condition for the generalized α-surfaces is interpreted as the consistency condition of a Lax

pair or more general linear system. For a more detailed discussion from this point of view, see Mason and Woodhouse (1996, chapter 13).

The Anti-Self-Dual Yang–Mills Equation and Its Twistor Correspondence

The anti-self-dual Yang–Mills equations extend Maxwell's equations for electromagnetism in the right-circularly polarized case. They are a family of equations that depend on a choice of Lie group G, usually taken to be a group of complex matrices; Maxwell's equations arise when $G = \mathrm{U}(1)$.

Introduce coordinates $x^a, a = 0, 1, 2, 3$, on $\mathbb{R}^4$ with metric $\mathrm{d}s^2 = \mathrm{d}x^0 \cdot \mathrm{d}x^3 - \mathrm{d}x^1 \cdot \mathrm{d}x^2$ (this is a metric of ultrahyperbolic signature – Euclidean signature can be obtained by choosing the coordinates to be complex, but with $(x^3, -x^2)$ the complex conjugates of (x^0, x^1)). The dependent variables are the components A_a of a connection $D_a = \partial_a - A_a$, where $\partial_a = \partial/\partial x^a$ and $A_a = A_a(x^b) \in \mathrm{Lie}\, G$, the Lie algebra of G. This connection defines a method of differentiating vector-valued functions s in some representation of G. The freedom in changing bases for the vector bundle induce the gauge transformations $A_a \to g^{-1}A_a g - g^{-1}\partial_a g, g(x) \in G$ on A_a; two connections that are related by a gauge transformation are deemed to be the same. The self-dual Yang–Mills equations are the condition

$$[D_0, D_2] = [D_1, D_3] = [D_0, D_3] - [D_1, D_2] = 0$$

They are the compatibility conditions

$$[D_0 + \lambda D_1,\ D_2 + \lambda D_3] = 0$$

for the linear system of equations

$$(D_0 - \lambda D_1)s = (D_2 - \lambda D_3)s = 0 \qquad [1]$$

where $\lambda \in C$ and s is an n-component column vector. These latter equations form a "Lax pair" for the system.

The Ward (1977) construction provides a one–one correspondence between gauge equivalence classes of solutions of the self-dual Yang–Mills equations and holomorphic vector bundles on regions in twistor space. The key point here is that eqn [1] defines parallel propagation along α-planes. To each point Z in twistor space, we can associate the vector space E_Z of solutions to eqn [1] along the corresponding α-plane. These vector spaces vary holomorphically with Z and that is what one means by a holomorphic vector bundle $E \to \mathbb{PT}$. The remarkable fact is that the anti-self-dual Yang–Mills field can be reconstructed up to gauge from E, and, in effect, for local analytic solutions, E can be represented by freely prescribable "patching" data consisting of local holomorphic matrix-valued functions on twistor space. To construct the solution on spacetime, one must first find a Birkhoff factorization of the patching data on each Riemann sphere in twistor space corresponding to points of the appropriate region in spacetime. On each Riemann sphere, the Birkhoff factorization starts with the given patching function with values in $\mathrm{GL}(n, \mathrm{C})$ on the real axis in the complex plane, and expresses it as a product of functions with values in $\mathrm{GL}(n, \mathrm{C})$ one of which extends over the upper-half plane, and the other over the lower-half complex plane. The anti-self-dual connection can be obtained by differentiating the resulting matrices. See Penrose (1984, 1986), Ward and Wells (1990), or Mason and Woodhouse (1996) for a full discussion, and Atiyah (1979) for the formulation appropriate to Euclidean signature.

Completely Integrable Systems

In effect, the twistor constructions amount to providing a geometric general local solution to the anti-self-duality equations; the twistor data is, for a local solution, freely prescribable. In this sense, they demonstrate complete integrability of the anti-self-duality equations. The reconstruction of a solution on spacetime from twistor data is not a quadrature – it involves, in the anti-self-dual Yang–Mills case, a Birkhoff factorization (also sometimes referred to as the solution to a Riemann–Hilbert problem), and in the case of the anti-self-dual Einstein equations, the construction of a family of rational curves inside a complex manifold. Nevertheless, such constructions are a familiar part of the apparatus of the theory of integrable systems.

In Ward (1985), this connection with integrable systems was developed further, and the anti-self-dual Yang–Mills equations were shown to yield many important integrable systems under symmetry reduction. Ward's list has been extended and now includes many of the most famous examples of integrable systems such as the Painlevé equations, the Korteweg–de Vries (KdV) equation, the nonlinear Schrödinger equation, the n-wave equations, and so on, see Mason and Woodhouse (1996) for a review. There are some notable omissions from the list such as the Kadomtsev–Petviashvili (KP) and Davey–Stewartson equations (at least if one restricts oneself to finite-dimensional gauge groups; reductions using infinite dimensional gauge groups have been obtained).

The list of integrable systems obtainable by symmetry reduction nevertheless remains impressive and provides a route to the classification of at least those integrable systems that can be obtained in this

way. Such systems can be classified by the choice of ingredients required in the symmetry reduction: the gauge group, the group of spacetime symmetries to be reduced by, the choice of Euclidean or ultra-hyperbolic signature, and the choice of certain constants of integration that arise in the reduction.

Another implication is that if an integrable system can be obtained from one of the self-duality equations by symmetry reduction, then it inherits a reduced twistor correspondence because the twistor correspondences share the symmetry groups of the spacetime field equations. These twistor correspondences can be seen to underlie much of the theory of these equations; for example, Backlund transformations of solutions correspond to elementary algebraic operations on the twistor data, similarly the Kac–Moody Lie algebras of hidden symmetries act locally on the twistor data by matrix multiplication of the appropriate loop algebras. Similarly, the inverse-scattering transform for the KdV and nonlinear Schrodinger equations can be seen to arise as particular presentations of the twistor construction.

By and large, although twistor methods have yielded new insight into the geometry and structure of systems in dimensions 1 and 2, they have not necessarily superceded pre-existing techniques for constructing solutions and analyzing the solution space. The systems for which twistor methods have been particularly effective for constructing solutions and characterizing their properties are in $2+1$ or higher dimension. Key examples here are of course the anti-self-dual Yang–Mills and Einstein equations themselves, and their single translation reductions. In the anti-self-dual Yang–Mills case, these reductions lead either to Ward's or Manakov and Zakharov's chiral model in Lorentzian signature, $2+1$, or the Bogomolny equations for monopoles, the reduction from Euclidean signature. In both cases, the twistor construction has played a major role in constructing and studying the solitonic solutions.

See Ward and Wells (1990), Mason and Woodhouse (1996), Ward's article in Huggett *et al.* (1998) and the first few chapters of Mason *et al.* (1995), and Mason *et al.* (2001) for more examples of aspects of the theory of integrable systems arising from twistor correspondences.

Applications to Geometry

These applications are, to a large extent, higher-dimensional analogs of those discussed above; most of the problems in geometry to which twistor theory has been applied are those for which the underlying differential equations are integrable. These start with the Euclidean signature versions of the original Ward construction for anti-self-dual Yang–Mills fields and Penrose's nonlinear graviton construction for Ricci-flat anti-self-dual metrics but, as we will discuss, these constructions have a number of extensions and generalizations.

The first dramatic application of these constructions was the ADHM construction of Yang–Mills instantons. These are absolute minima of the Yang–Mills action, $S[A] = \int \mathrm{tr}(F \wedge F^*)$ on the 4-sphere, S^4, with its round metric. A simple argument shows that the action is bounded below by the second Chern class of the bundle and that this bound is achieved only for anti-self-dual fields. Thus, the problem was to characterize all the anti-self-dual Yang–Mills fields on S^4. In this Euclidean context, twistor space, $\mathbb{CP}^3$, fibers over S^4 and the corresponding Ward vector bundle is a bundle over all of $\mathbb{CP}^3$. It turns out that all such bundles satisfying a certain stability condition had been constructed reasonably explicitly by algebraic geometers. Since the stability condition was implied by the context, this could be turned into an algebraic construction of the general instanton explicit enough to give some insight into both the local and global structure of the solution space. See Atiyah (1979) for a review.

Hitchin used the Euclidean version of the nonlinear graviton to develop the theory of gravitational instantons that are asymptotically locally Euclidean (i.e., asymptotically $\mathbb{R}^4/\Gamma$, where Γ is a finite subgroup of the rotation group). These were finally constructed by Kronheimer who again used twistor theory to identify the appropriate parameter space, see his article in Mason *et al.* (2001) and Dancer's review of hyper-Kähler manifolds in LeBrun and Wang (1999).

Even in four dimensions, there are a number of variants of the nonlinear graviton construction. The basic twistor correspondence produces a twistor space that is a complex 3-manifold $\mathcal{PT}$ for 4-manifolds with conformal structures whose Weyl tensor is anti-self-dual. There are four natural specializations that have attracted study: (1) the Ricci-flat case, (2) the Einstein case (with nonzero cosmological constant), (3) the scalar-flat Kähler case, and (4) the hypercomplex case.

The twistor space in the Ricci-flat case admits the additional structure of a fibration over $\mathbb{CP}^1$ together with a holomorphic Poisson structure on the fibers with values in the pullback of the 1-forms on $\mathbb{CP}^1$ (alternatively, the bundle of holomorphic 3-forms should be the pullback of the square of the bundle of holomorphic 1-forms on $\mathbb{CP}^1$). The Einstein case with nonzero cosmological constant is a variant of this in which the twistor space admits a

nondegenerate holomorphic contact structure, that is, a distribution of 2-plane elements, which are only integrable when the cosmological constant vanishes. It also admits a Kähler form when the scalar curvature is positive (in the negative case the corresponding Kähler form is indefinite). For the case of Kähler metrics with vanishing scalar curvature, the twistor space admits a holomorphic volume form with a double pole. The Ricci-flat case is equivalent to the case of hyper-Kähler metrics, those that are Kähler with respect to three different complex structures I, J, and K satisfying the standard quaternionic relations $IJ = K$, etc. A hypercomplex structure is obtained when one only has the three integrable complex structures satisfying the quaternion relations. Such manifolds admit an underlying conformal structure that is anti-self-dual, and the corresponding twistor space admits a fibration to $\mathbb{CP}^1$.

These constructions have all played a significant role in the general analysis of these geometric structures, and the construction of examples. A striking example of an application of the nonlinear graviton construction to general properties is due to Donaldson and Friedman who show that if two 4-manifolds admit anti-self-dual conformal structures, then their direct sum does also.

In higher dimensions, most generalizations rely on quaternionic geometry and its reductions. The Euclidean signature formulation of the nonlinear graviton construction has natural extensions to quaternionic manifolds in $4k$ dimensions. These are manifolds with metric whose holonomies are contained in $\mathrm{Sp}(k) \times \mathrm{Sp}(1)$. The latter $\mathrm{SP}(1) = \mathrm{SU}(2)$ factor leads to an associated S^2 bundle whose total space is the twistor space $\mathcal{PT}$ and it naturally has the structure of a $(2k+1)$-dimensional complex manifold.

For a series of review articles, the reader is referred to Bailey and Baston (1990, chapters 3 and 4) and also LeBrun and Wang (1999, chapters 2, 5, 6, 10, and 14) which, despite being a book on the distinct subject of Einstein manifolds, is strongly influenced by twistor theory. Other applications along these lines are summarized in Mason *et al.* (2001, chapter 1).

There are a number of applications that go beyond complete integrability. A striking application is the twistor framework of Merkulov for studying arbitrary geometric structures. This has led to a classification of all possible irreducible holonomies of torsion-free affine connections, see Merkulov's article in Huggett *et al.* (1998). Another important area is in the field of conformal invariants in which the local twistor connection plays a prominent role. This is a connection that is naturally defined on any conformal manifold being the spinor representation of the Cartan conformal connection. An impressive application here is the construction of conformally invariant differential operators and other conformal invariants. See the article by Baston and Eastwood in Bailey and Baston (1990).

Beyond Classical Integrability: Twistor-String Theory

Until Witten (2004), there was little indication that twistor theory would have much useful to say about Yang–Mills or gravitational fields that are not anti-self-dual. Furthermore, it was problematic to incorporate quantum field theory into twistor ideas. However, twistor-string theory has transformed the situation and has furthermore had impressive applications to the field of perturbative gauge theory.

The story starts with a formulation by Nair of the remarkable Park–Taylor formulas for the so-called maximal helicity violating (MHV) amplitudes in gauge theory. These are scattering amplitudes at tree level in which helicity conservation is maximally violated; using crossing symmetry to take all the particles to be outgoing, these are amplitudes in which $n-2$ of the particles have helicity -1 and two have helicity $+1$. These amplitudes can be expressed simply as follows. Let the n particles have color t_i in the Lie algebra of the gauge group and null momenta p_i with spinor decompositions $p_i^a = \tilde{\pi}_i^A \pi_i^{A'}$, $i = 1, \ldots, n$ where the $\pi_i^{A'}$ are self-dual spinors and $\tilde{\pi}^A$ are anti-self-dual spinors using the index notation of Spinors and Spin Coefficients, and Twistors. Let $i = r$ and $i = s$ be the two gluons of helicity $+1$. Then the coefficient of the colour term $\mathrm{tr}(t_1 t_2 \cdots t_n)$ is

$$\delta^4\left(\sum_{i=1}^{n} p_i^a\right) \frac{\pi_r \cdot \pi_s}{\Pi_{i=1}^{n} \pi_i \cdot \pi_{i+1}}$$

where $\pi_i \cdot \pi_j = \pi_i^{A'} \pi_{jA'}$ denotes the standard skew-symmetric inner product on chiral spinors and $\pi_{n+1} = \pi_1$. A striking feature is that, except for the delta function, it is holomorphic in the π_is except at the simple poles $\pi_i \cdot \pi_{i+1} = 0$. Nair interprets these poles as those associated to fermion correlators in a current algebra on a $\mathbb{CP}^1$ parametrized by π. Using a supersymmetric formulation adapted to $N = 4$ super Yang–Mills, he formulated the amplitude as arising from an integral over lines in supertwistor space $\mathbb{CP}^{3|4}$.

Witten extends these ideas to give, at least conjecturally, a complete theory. He proposes that full perturbative $N = 4$ super Yang–Mills theory on spacetime is equivalent to a string theory, a topological

B model, on a supersymmetric version of twistor space, $\mathbb{PT}_s = \mathbb{CP}^{3|4}$. This is the space obtained by taking $\mathbb{C}^{4|4}$ with bosonic coordinates $Z^\alpha, \alpha = 0, \ldots, 3$ and fermionic coordinates $\eta^i, i = 1, \ldots, 4$ moduli the equivalence relation $(Z^\alpha, \eta^i) \sim \lambda(Z^\alpha, \eta^i)$ where $\lambda \in \mathbb{C}, \lambda \neq 0$.

The number 4 here plays two crucial but different roles. It is the maximum number of supersymmetries that Yang–Mills can have; it has the effect of incorporating both the positive and negative helicity parts of the gauge field in the same supermultiplet. It is also the only value of N for which $\mathbb{CP}^{3|N}$ is a Calabi–Yau manifold and this is a necessary condition for the topological twisted B model to be anomaly-free. The Calabi–Yau condition is the condition that the manifold admit a global holomorphic volume form which here is

$$\Omega_s = \varepsilon_{\alpha\beta\gamma\delta} Z^\alpha \mathrm{d}Z^\beta \wedge \mathrm{d}Z^\gamma \wedge \mathrm{d}Z^\delta \wedge \mathrm{d}\eta^1 \wedge \mathrm{d}\eta^2 \wedge \mathrm{d}\eta^3 \wedge \mathrm{d}\eta^4$$

This is invariant under $(Z^\alpha, \eta^i) \to (\lambda Z^\alpha, \lambda \eta^i)$ because $\mathrm{d}(\lambda\eta^i) = \lambda^{-1}\mathrm{d}\eta^i, \lambda \in \mathbb{C}$ follows from the Berezinian rule of integration $\int \theta \mathrm{d}\theta = 1$ for anticommuting variables.

Open-string topological twisted B models are known to correspond to holomorphic Chern–Simons theories on their target space. A holomorphic Chern–Simons theory is a theory whose basic variable is a d-bar operator $\bar\partial_{\mathcal{A}} = \bar\partial + \mathcal{A}$ on a complex vector bundle $E \to \mathbb{PT}^{3|4}$, where $\mathcal{A}$ is a Lie algebra valued $(0,1)$-form on the target space and whose action is

$$S[\mathcal{A}] = \int \frac{1}{2}\left(\mathcal{A}\bar\partial\mathcal{A} + \frac{1}{3}\mathcal{A}^3\right) \wedge \Omega_s$$

The field equations are $\bar\partial_{\mathcal{A}}^2 = 0$. The classical solutions therefore consist of holomorphic vector bundles on the target space, here $\mathbb{CP}^{3|4}$. The twistor-space representation of the fields are obtained by expanding $\mathcal{A}$ in the anticommuting variables η^i to obtain

$$\mathcal{A} = a + \eta^i b_i + \eta^i\eta^j c_{ij} + \eta^i\eta^j\eta^k d_{ijk} + \eta^1\eta^2\eta^3\eta^4 g$$

and a has homogeneity zero, but because the homogeneity of η^i is of degree 1, b_i has homogeneity degree -1, and so on down to homogeneity degree -4 for g. Via the Ward construction, the a component corresponds to an anti-self-dual Yang–Mills field on spacetime. The other components of $\mathcal{A}$ can be seen to correspond to spacetime fields with helicities $-1/2$ to $+1$ that are background coupled to the anti-self-dual Yang–Mills field.

As it stands, although this holomorphic Chern–Simons theory gives the correct field content of $N = 4$ super Yang–Mills, the couplings are only those of an anti-self-dual sector and more couplings are needed to obtain full $N = 4$ super Yang–Mills. The remarkable fact is that these can be naturally introduced by coupling in certain D1 instantons. The D1 instantons are algebraic curves C in twistor space and the coupling is via a pair of spinor fields α and β on C with values in E and E^*, respectively with action

$$S[\alpha, \beta, \mathcal{A}] = \int_C \beta\bar\partial_{\mathcal{A}}\alpha$$

This leads to explicit expressions for Yang–Mills scattering amplitudes in terms of integrals of fermion correlators over the moduli spaces of such algebraic curves in supertwistor space. In principle, the integral is over all algebraic curves. However, algebraic curves have two topological invariants, their degree denoted d and genus g. An argument based on a classical scaling symmetry gives that integration over just those of curves of degree d gives the subset of processes for which

$$d = q - 1 + l$$

where q is the number of outgoing particles of helicity $+1$ in the process and l is the number of loops. It is also the case that $g \leq l$.

An elegant formula for the amplitudes is that for the on-shell generating functional for tree-level scattering amplitudes $\mathscr{A}[\mathcal{A}]$, where $\mathcal{A}$ is the on-shell twistor field, being the above-mentioned $(0,1)$-form. The generating functional for processes with $q = d + 1$ external fields of helicity $+1$ is then

$$\mathscr{A}^d[\mathcal{A}] = \int_{C \in \mathscr{M}^d} \det(\bar\partial + \mathcal{A})|_C \mathrm{d}\mu$$

where $\mathrm{d}\mu$ is a natural measure on the moduli space $\mathscr{M}^d$ of connected rational (genus 0) curves in $\mathbb{CP}^{3|4}$ of degree d. This approach has been successfully exploited to obtain implicit algebraic formulas for all tree-level scattering amplitudes.

In an alternative version, the curves of degree d can be taken to be maximally disconnected, being the union of d lines. However, in this approach, we need to also incorporate Chern–Simons propagators which, for tree diagrams, join the lines into a tree. This gives a very flexible calculus for perturbative gauge theory in which scattering processes are obtained by gluing together MHV diagrams. It has been argued that the two formulations are equivalent. On the one hand, the Chern–Simons propagator has a simple pole when the lines meet and the

contour integral over the moduli space can be performed using residues in such a way as to eliminate the Chern–Simons propagators leaving an integral over d intersecting lines. On the other hand, the measure on the space of connected curves has a simple pole where the curve acquires double points and again the contour integral can be performed in such a way as to yield the same integral over d intersecting lines.

It should be mentioned that Berkovits has given an alternative version of twistor-string theory which is a heterotic open-string theory with target supertwistor space in which the strings are taken to have boundary on the real slice $\mathbb{RP}^3$ in $\mathbb{CP}^3$ (this is appropriate to a spacetime with split signature) and the D1-instanton expansions are replaced by expansions in the fundamental modes of the string (this is not a topological theory). This gives rise to the same formulas for scattering amplitudes as Witten's original model.

There have been many applications now of these ideas, perhaps the most striking being the recursion relations of Britto, Cachazo, Feng, and Witten which give, at tree level, on-shell recurrence relations for Yang–Mills scattering amplitudes that suggests a hitherto unsuspected underlying structure for Yang-Mills theory.

Despite all these successes, twistor-string theory is not thought by string theorists to be a good vehicle for basic physics. The most serious problem is that the closed-string sector gives rise to conformal supergravity which is an unphysical theory. This is particularly pernicious from the point of view of analyzing loop diagrams as from the point of view of string theory, loop diagrams will carry supergravity modes. From this point of view, twistor-string theory is another duality, like AdS-CFT etc., that gives insight into some standard physics but is fundamentally limited.

From the point of view of a twistor theorist, however, twistor-string theory has overcome major obstacles to the twistor programme. Hodges has used the BCFW recursion relations to provide all twistor diagrams for gauge theory. In Mason (2005) it is shown how to derive the main generating function formulas from Yang–Mills and conformal gravity spacetime action principles via a twistor space actions for these theories. These twistor actions can in the first instance be expressed purely bosonically and distinctly and the twistor-string generating function formulas are obtained by expanding and re-summing the classical limit of the path integral in a parameter that expands about the anti-self-dual sector. This allows one to decouple the Yang–Mills and conformal gravity modes, and indeed to work purely bosonically – one is not tied to super Yang–Mills. Although there is much work to be done to extend these ideas to provide a consistent approach to the main equations of basic physics, obstacles that seemed insurmountable a few years ago have been overcome.

See also: Chern–Simons Models: Rigorous Results; Einstein Equations: Exact Solutions; General Relativity: Overview; Instantons: Topological Aspects; Integrable Systems and the Inverse Scattering Method; Riemann–Hilbert Methods in Integrable Systems; Spinors and Spin Coefficients; Twistors; Classical Groups and Homogeneous Spaces; Quantum Mechanics: Foundations; Several Complex Variables: Compact Manifolds; Several Complex Variables: Basic Geometric Theory.

Further Reading

Atiyah MF (1979) *Geometry of Yang–Mills Fields: Lezioni Fermiane*. Pisa: Accademia Nazionale dei Lincei Scuola Normale Superiore.

Atiyah MF, Hitchin NJ, and Singer IM (1978) Self-duality in four-dimensional Riemannian geometry. *Proceedings of the Royal Society A* 362: 425.

Bailey TN and Baston R (eds.) (1990) *Twistors in Mathematics and Physics*, LMS Lecture Notes Series, vol. 156. Cambridge: Cambridge University Press.

Baston RJ and Eastwood MG (1989) *The Penrose Transform: Its Interaction with Representation Theory*. Oxford: Oxford University Press.

Cachazo F, and Svrcek P (2005) Lectures on twistor strings and perturbative Yang–Mills theory, arXiv:hep-th/0504194.

Hitchin N (1987) *Monopoles, Minimal Surfaces and Algebraic Curves, Seminaire de Mathematiques supérieures*, vol. 105. NATO Advanced Study Institute. Les Presses de l'Universite De Montreal.

Huggett S, Mason LJ, Tod KP, Tsou TS, and Woodhouse NMJ (eds.) (1998) *The Geometric Universe*. Oxford: Oxford University Press.

LeBrun C and Wang M (eds.) (1999) *Essays on Einstein manifolds*, Surveys in Differential Geometry, vol. VI. Boston, MA: International Press.

Mason LJ (2005) Twistor actions for non-self-dual fields, a derivation of twistor-string theory, hep-th/0507269.

Mason LJ and Hughston LP (eds.) (1990) *Further Advances in Twistor Theory, Volume I: The Penrose Transform and Its Applications*. Pitman Research Notes in Maths, vol. 231. Harlow: Longman.

Mason LJ, Hughston LP, and Kobak PZ (1995) *Further Advances in Twistor Theory, Volume II: Integrable Systems, Conformal Geometry and Gravitation*. Pitman Research Notes in Maths, vol. 232. Harlow: Longman.

Mason LJ, Hughston LP, Kobak PZ, and Pulverer K (eds.) (2001) *Further Advances in Twistor Theory, Volume III: Curved Twistor Spaces*. Pitman Research Notes in Maths, vol. 424. Boca Raton, FL: Chapman and Hall/CRC Press.

Mason LJ and Woodhouse NMJ (1996) *Twistor Theory, Self-Duality and Twistor Theory*. Oxford: Oxford University Press.

Newman ET (1976) Heaven and its properties. *General Relativity and Gravitation* 7(1): 107–111.

Penrose R (1976) Nonlinear gravitons and curved twistor theory. *General Relativity and Gravitation* 7: 31–52.

Penrose R (1984, 1986) *Spinors and Space-Time*, vols. I and II. Cambridge: Cambridge University Press.

Ward RS (1977) On self-dual gauge fields. *Physics Letters A* 61: 81–82.

Ward RS (1985) Integrable and solvable systems and relations among them. *Philosophical Transactions of the Royal Society A* 315: 451–457.

Ward RS and Wells RO (1990) *Twistor Geometry and Field Theory*. Cambridge: Cambridge University Press.

Witten E (2004) Perturbative gauge theory as a string theory in twistor space. *Communications in Mathematical Physics* 252: 189 (arXiv:hep-th/0312171).

Two-Dimensional Models

B Schroer, Freie Universität Berlin, Berlin, Germany

© 2006 Elsevier Ltd. All rights reserved.

History and Motivation

Local quantum physics of systems with infinitely many interacting degrees of freedom leads to situations whose understanding often requires new physical intuition and mathematical concepts beyond that acquired in quantum mechanics and perturbative constructions in quantum field theory. In this situation, two-dimensional soluble models turned out to play an important role. On the one hand, they illustrate new concepts and sometimes remove misconceptions in an area where new physical intuition is still in the process of being formed. On the other hand, rigorously soluble models confirm that the underlying physical postulates are mathematically consistent, a task which for interacting systems with infinite degrees of freedom is mostly beyond the capability of pedestrian methods or brute force application of hard analysis on models whose natural invariances have been mutilated by a cutoff.

In order to underline these points and motivate the interest in two-dimensional QFT, let us briefly look at the history, in particular at the physical significance of the three oldest two-dimensional models of relevance for statistical mechanics and relativistic particle physics, in chronological order: the Lenz–Ising (L–I) model, Jordan's model of bosonization/fermionization, and the Schwinger model (QED_2). (A more detailed account of the changeful history concerning their correct physical interpretation and generalizations to higher dimensions of these models and the increasing conceptual role of low-dimensional models in QFT can be found in Schroer (2005).)

The L–I model was proposed in 1920 by Wihelm Lenz (see Lenz (1925)) as the simplest discrete statistical mechanics model with a chance to go beyond the P Weiss phenomenological ansatz involving long-range forces and instead explain ferromagnetism in terms of nonmagnetic short-range interactions. Its one-dimensional version was solved four years later by his student Ernst Ising. Its changeful history reached a temporary conceptual climax when Onsager succeeded to rigorously establish a second-order phase transition in two dimensions.

Another conceptually rich model which lay dormant for almost two decades as a result of a misleading speculative higher-dimensional generalization by its protagonist is the bosonization/fermionization model first proposed by Jordan (1937). This model establishes a certain equivalence between massless two-dimensional fermions and bosons and is related to Thirring's massless 4-fermion coupling model and also to Luttinger's one-dimensional model of an electron gas (Schroer). One reason why even nowadays hardly anybody knows Jordan's contribution is certainly the ambitious but unfortunate title "the neutrino theory of light" under which he published a series of papers.

Both discoveries demonstrate the usefulness of having controllable low-dimensional models; at the same time, their complicated history also illustrates the danger of rushing to premature "intuitive" conclusions about extensions to higher dimensions.

A review of the early historical benchmarks of conceptual progress through the study of solvable two-dimensional models would be incomplete without mentioning Schwinger's (1962) proposed solution of two-dimensional quantum electrodynamics, afterwards referred to as the Schwinger model. He used this model in order to argue that gauge theories are not necessarily tied to zero-mass vector particles. Some work was necessary (Schroer) to unravel its physical content with the result that the would-be charge of that QED_2 model was "screened" and its apparent chiral symmetry broken; in other words, the model exists only in the so-called Schwinger–Higgs phase with massive free scalar particles accounting for its physical content. Another closely related aspect of

Notes:

this model which also arose in the Lagrangian setting of four-dimensional gauge theories was that of the θ-angle parametrizing, an ambiguity in the quantization.

A coherent and systematic attempt at a mathematical control of two-dimensional models came in the wake of Wightman's first rigorous programmatic formulation of QFT (Schroer 2005). This formulation stayed close to the physical ideas underlying the impressive success of renormalized QED perturbation theory, although it avoided the direct use of Lagrangian quantization. The early attempts towards a "constructive QFT" found their successful realization in two-dimensional QFT (the $P\varphi_2$ models (Glimm and Jaffe 1987)); the restriction to low dimensions is related to the mild short-distance singularity behavior (super-renormalizability) which these methods require. We will focus our main attention on alternative constructive methods which, even though not suffering from such short-distance restrictions, also suffer from a lack of mathematical control in higher spacetime dimensions; the illustration of the constructive power of these new methods comes presently from massless $d=1+1$ conformal and chiral QFT as well as from massive factorizing models.

There are several books and review articles (Furlan *et al.* 1989, Ginsparg 1990, Di Francesco *et al.* 1996) on $d=1+1$ conformal as well as on massive factorizing models (Abdalla *et al.* 1991). To the extent that concepts and mathematical structures are used which permit no extension to higher dimensions (Kac–Moody algebras, loop groups, integrability, presence of an infinite number of conservation laws), this line of approach will not be followed in this article since our primary interest will be the use of two-dimensional models of QFT as "theoretical laboratories" of general QFT. Our aim is twofold; on the one hand, we intend to illustrate known principles of general QFT in a mathematically controllable context and on the other hand, we want to identify new concepts whose adaptation to QFT in $d=1+1$ lead to their solvability (Schroer).

General Concepts and Their Two-Dimensional Manifestation

The general framework of QFT, to which the rich world of controllable two-dimensional models contributes as an important testing ground, exists in two quite different but nevertheless closely related formulations: the 1956 approach in terms of pointlike covariant fields due to Wightman (see Streater and Wightman (1964)) (*see* Axiomatic Quantum Field Theory), and the more algebraic setting which can be traced back to ideas which Haag (1992) developed shortly after and which are based on spacetime-indexed operator algebras and related concepts which developed over a long period of time, with contributions of many other authors to what is now referred to as algebraic QFT (AQFT) or simply local quantum physics (LQP). Whereas the Wightman approach aims directly at the (not necessarily observable) quantum fields, the operator-algebraic setting (*see* Algebraic Approach to Quantum Field Theory) is more ambitious. It starts from physically well-motivated assumptions about the algebraic structure of local observables and aims at the reconstruction of the full field theory (including the operators carrying the superselected charges) in the spirit of a local representation theory of (the assumed structure of the) local observables. This has the advantage that the somewhat mysterious concept of an inner symmetry (as opposed to outer (spacetime) symmetry) can be traced back to its physical roots which is the representation-theoretical structure of the local observable algebra (*see* Symmetries in Quantum Field Theory of Lower Spacetime Dimensions). In the standard Lagrangian quantization approach, the inner symmetry is part of the input (multiplicity indices of field components on which subgroups of $U(n)$ or $O(n)$ act linearly) and hence it is not possible to problematize this fundamental question. When in low-dimensional spacetime dimensions the sharp separation (the Coleman–Mandula theorems) of inner versus outer symmetry becomes blurred as a result of the appearance of braid group statistics, the standard Lagrangian quantization setting of most of the textbooks is inappropriate and even the Wightman framework has to be extended. In that case, the algebraic approach is the most appropriate.

The important physical principles which are shared between the Wightman approach (see Streater and Wightman (1964)) and the operator algebra (AQFT) setting (Haag 1992) are the spacelike locality or Einstein causality (in terms of pointlike fields or algebras localized in causally disjoint regions) and the existence of positive-energy representations of the Poincaré group implementing covariance and the stability of matter. In the algebraic approach, the observable content of the theory is encoded into a family of (weakly closed) operator algebras $\{\mathcal{A}(\mathcal{O})\}_{\mathcal{O}\in\mathcal{K}}$ indexed by a family of convex causally closed spacetime regions $\mathcal{O}$ (with $\mathcal{O}'$ denoting the spacelike complement and $\mathcal{A}'$ the von Neumann commutant) which act in one common Hilbert space. Covariant local fields lose their distinguished role

which they have in the classical setting and which (via Lagrangian quantization) was at least partially inherited by the Wightman approach and, apart in their role as local generators of symmetries (conserved currents), became mere "field coordinatizations" of local algebras. (There is a denumerable set of such pointlike field generators which form a local equivalence (Borchers) class of fields and in the absence of interactions permits a neat description in terms of Wick-ordered free-field polynomials (Haag 1992). Certain properties cannot be naturally formulated in the pointlike field setting (e.g., Haag duality for convex regions $\mathcal{A}(\mathcal{O}') = \mathcal{A}(\mathcal{O})'$), but apart from those properties the two formulations are quite close; in particular for two-dimensional theories there are convincing arguments that one can pass between the two without imposing additional technical requirements. (Haag duality holds for observable algebras in the vacuum sector in the sense that any violation can be explained in terms of a spontaneously broken symmetry; in local theories, it can always be enforced by dualization and the resulting Haag dual algebra has a charge superselection structure associated with the unbroken subgroup.) Haag duality is the statement that the commutant of observables not only contains the algebra of the causal complement that is, $\mathcal{A}(\mathcal{O}') \subset \mathcal{A}(\mathcal{O})'$ (Einstein causality) but is even exhausted by it; it is deeply connected to the measurement process and its violation in the vacuum sector for convex causally complete regions signals spontaneous symmetry breaking in the associated charge-carrying field algebra (Haag 1992). It can always be enforced (assuming that the wedge-localized algebras fulfill [1] below) by symmetry-reducing extension called Haag dualization. Its violation for multilocal region reveals the charge content of the model via charge–anticharge splitting in the neutral observable algebra (Schroer).

Another physically important property which has a natural algebraic formulation is the split property: for regions $\mathcal{O}_i$ separated by a finite spacelike distance, one finds $\mathcal{A}(\mathcal{O}_1 \cup \mathcal{O}_2) \simeq \mathcal{A}(\mathcal{O}_1) \otimes \mathcal{A}(\mathcal{O}_2)$ which can be derived from the Buchholz–Wichmann "nuclearity property" (Haag 1992) (an appropriate adaptation of the "finiteness of phase-space cell" property of QM to QFT). Related to the Haag duality is the local version of the "time slice property" (the QFT counterpart of the classical causal dependency property) sometimes referred to as "strong Einstein causality" $\mathcal{A}(\mathcal{O}'') = \mathcal{A}(\mathcal{O})''$.

One of the most astonishing achievements of the algebraic approach (which justifies its emphasis on properties of "local observables") is the DHR theory of superselection sectors (Doplicher *et al.* 1971), that is, the realization that the structure of charged (nonvacuum) representations (with the superposition principle being valid only within one representation) and the spacetime properties of the generating fields which are the carriers of these generalized charges (including their spacelike commutation relations which lead to the particle statistics and also to their internal symmetry properties) are already encoded in the structure of the Einstein causal observable algebra (Symmetries in Quantum Field Theory: Algebraic Aspects). The intuitive basis of this remarkable result (whose prerequisite is locality) is that one can generate charged sectors by spatially separating charges in the vacuum (neutral) sector and disposing of the unwanted charges at spatial infinity (Haag 1992).

An important concept which especially in $d = 1 + 1$ has considerable constructive clout is "modular localization." It is a consequence of the above algebraic setting if either the net of algebras have pointlike field generators, or if the one-particle masses are separated by spectral gaps so that the formalism of time-dependent scattering can be applied (Schroer 2005); in conformal theories, this property holds automatically in all spacetime dimensions. It rests on the basic observation (Tomita–Takesaki Modular Theory) that a standard pair $(\mathcal{A}, \Omega)$ of a von Neumann operator algebra and a standard vector (standardness means that the operator algebra of the pair $(\mathcal{A}, \Omega)$ acts cyclic and separating on the vector Ω) gives rise to a Tomita operator S through its star-operation whose polar decomposition yield two modular objects, a one-parametric subgroup Δ^{it} of the unitary group of operators in Hilbert space whose Ad-action defines the modular automorphism of $(\mathcal{A}, \Omega)$ whereas the angular part J is the modular conjugation which maps $\mathcal{A}$ into its commutant $\mathcal{A}'$.

$$\begin{aligned} &SA\Omega = A^*\Omega,\ S = J\Delta^{1/2} \\ &J_W = U(j_W) = S_{\text{scat}} J_0, \Delta_W^{it} = U(\Lambda_W(2\pi t)) \qquad [1] \\ &\sigma_W(t) := \text{Ad}\Delta_W^{it} \end{aligned}$$

The standardness assumption is always satisfied for any field-theoretic pair $(A(\mathcal{O}), \Omega)$ of a $\mathcal{O}$-localized algebra and the vacuum state (as long as $\mathcal{O}$ has a nontrivial causal disjoint $\mathcal{O}'$), but it is only for the wedge region W that the modular objects have a physical interpretation in terms of the global symmetry group of the vacuum as specified in the second line of [1]; the modular unitary Δ_W^{it} represents the W-associated boost $\Lambda_W(\chi)$ and the modular conjugation J_W implements the TCP-like reflection along the edge of the wedge (Bisognano

and Wichmann 1975). The third line is the definition of the modular group. The importance of this theory for local quantum physics results from the fact that it leads to the concept of modular localization, an intrinsic new scenario for field-theoretic constructions which is different from the Lagrangian quantization schemes (Schroer 2005).

A special feature of $d = 1 + 1$ Minkowski spacetime is the disconnectedness of the right/left spacelike region leading to a right–left ordering structure. So in addition to the Lorentz-invariant timelike ordering $x \prec y$ (x earlier than y, which is independent of spacetime dimensions), there is an invariant spacelike ordering $x < y$ (x to the left of y) in $d = 1 + 1$ which opens the possibility of more general Lorentz-invariant spacelike commutation relations than those implemented by Bose/Fermi fields (Rehren and Schroer 1987) of fields with a spacelike braid group commutation structure. The appearance of such exotic statistics fields is not compatible with their Fourier transforms being creation/annihilation operators for Wigner particles; rather, the state vectors which they generate from the vacuum contain in addition to the one-particle contribution a vacuum polarization cloud (Schroer 2005). This close connection between new kinematic possibilities and interactions is one of the reasons why, (different from higher dimensions where interactions are prescribed by the recipe of local couplings of free fields) low-dimensional QFT offers a more intrinsic access to the central issue of interactions.

Boson/Fermion Equivalence and Superselection Theory in a Special Model

The simplest and oldest but conceptually still rich model is obtained, as first proposed by Jordan (1937), by using a two-dimensional massless Dirac current and showing that it may be expressed in terms of scalar canonical Bose creation/annihilation operators

$$j_\mu =: \overline{\psi}\gamma_\mu\psi := \partial_\mu\phi,\ \phi$$
$$:= \int_{-\infty}^{+\infty} \{e^{ipx}a^*(p) + h.c.\}\frac{dp}{2|p|} \qquad [2]$$

Although the potential $\phi(x)$ of the current as a result of its infrared divergence is not a field in the standard sense of an operator-valued distribution in the Fock space of the $a(p)^{\#}$ (It becomes an operator after smearing with test functions whose Fourier transform vanishes at $p = 0$), the formal exponential defined as the zero-mass limit of a well-defined exponential free massive field

$$: e^{i\alpha\phi(x)} := \lim_{m\to 0} m^{\alpha^2/2} : e^{i}\alpha\phi_m(x) : \qquad [3]$$

turns out to be a *bona fide* quantum field in a larger Hilbert space (which extends the Fock space generated from applying currents to the vacuum). The power in front is determined by the requirement that all Wightman functions (computed with the help of free-field Wick combinatorics) stay finite in this massless limit; the necessary and sufficient condition for this is the charge conservation rule

$$\left\langle \prod_i : e^{i\alpha_i\phi(x)} : \right\rangle$$
$$= \begin{cases} \prod_{i<j}\left(\dfrac{-1}{(\xi_{+ij})_\varepsilon(\xi_{-ij})_\varepsilon}\right)^{(1/2)\alpha_i\alpha_j}, & \Sigma\alpha_i = 0 \\ 0, & \text{otherwise} \end{cases} \qquad [4]$$

where the resulting correlation function has been factored in terms of light-ray coordinates $\xi_{\pm ij} = x_{\pm i} - x_{\pm j}, x_\pm = t \pm x$, and the ε-prescription stands for taking the standard Wightman boundary value $t \to t + i\varepsilon$, $\lim_{\varepsilon\to 0}$ which insures the positive-energy condition. The finiteness of the limit insures that the resulting zero-mass limiting theory is a *bona fide* quantum field theory that is, its system of Wightman functions permits the construction of an operator theory in a Hilbert space with a distinguished vacuum vector.

The factorization into light-ray components [4] shows that the exponential charge-carrying operators inherit this factorization into two independent chiral components $: \exp i\alpha\phi(x) : = \exp i\alpha\phi_+(x_+) : \exp i\alpha\phi_-(x_-) :$, each one being covariant under scaling $\xi \to \lambda\xi$ if one assigns the scaling dimension $d = \alpha^2/2$ to the chiral exponential field and $d = 1$ to the current. As any Wightman field, this is a singular object which only after smearing with Schwartz test functions yields an (unbounded) operator. But the above form of the correlation function belongs to a class of distributions which admits a much larger test-function space consisting of smooth functions which instead of decreasing rapidly only need to be bounded so that they stay finite on the compactified light-ray line $\dot{R} = S^1$. To make this visible, one uses the Cayley transform (now x denotes either x_+ or x_-)

$$z = \frac{1 + ix}{1 - ix} \in S^1 \qquad [5]$$

This transforms the Schwartz test function into a space of test functions on S^1 which have an infinite order zero at $z = -1$ (corresponding to $x = \pm\infty$) but the rotational transformed fields $j(z), : \exp i\alpha\phi(z) :$ permit the smearing with all smooth functions on S^1, a characteristic feature of all conformal invariant

theories as the present one turns out to be. There is an additional advantage in the use of this compactification. Fourier transforming the circular current actually allows for a quantum-mechanical zero mode whose possible nonzero eigenvalues indicate the presence of additional charge sectors beyond the charge-zero vacuum sector. For the exponential field, this leads to a quantum-mechanical pre-exponential factor which automatically insures the charge selection rules so that unrestricted (by charge conservation) Wick contraction rules can be applied. In this approach, the original chiral Dirac fermion $\psi(x)$ (from which the current was formed as the $:\psi\bar{\psi}:$ composite) reappears as a charge-carrying exponential field for $\alpha = 1$ and thus illustrates the meaning of bosonization/fermionization. (It is interesting to note that Jordan's (1937) original treatment of fermionization had such a pre-exponential quantum-mechanical factor.) Naturally, this terminology has to be taken with a grain of salt in view of the fact that the bosonic current algebra only generates a superselected subspace into which the charge-carrying exponential field does not fit. Only in the case of massive two-dimensional QFT fermions can be incorporated into a Fock space of bosons (see last section). At this point, it should however be clear to the reader that the physical content of Jordan's paper had nothing to do with its misleading title "neutrino theory of light" but rather was an early illustration about charge superselection rules in two-dimensional QFT.

A systematic and rigorous approach consists in solving the problem of positive-energy representation theory for the Weyl algebra on the circle (which is the rigorous operator-algebraic formulation of the abelian current algebra). (The Weyl algebra originated in quantum mechanics around 1927; its use in QFT only appeared after the cited Jordan paper. By representation we mean here a regular representation in which the exponentials can be differentiated in order to obtain (unbounded) smeared current operators.) It is the operator algebra generated by the exponential of a smeared chiral current (always with real test functions) with the following relation between the generators

$$\begin{aligned} W(f) &= \mathrm{e}^{\mathrm{i}j(f)} \\ j(f) &= \int \frac{\mathrm{d}z}{2\pi\mathrm{i}} j(z)f(z), \quad [j(z), j(z')] \\ &= -\delta'(z - z') \end{aligned} \tag{6a}$$

$$\begin{aligned} W(f)W(g) &= \mathrm{e}^{-(1/2)s(f,g)} W(f+g) \\ W^*(f) &= W(-f) \end{aligned} \tag{6b}$$

$$\begin{aligned} \mathcal{A}(S^1) &= \mathrm{alg}\{W(f), f \in C_\infty(S^1)\} \\ \mathcal{A}(I) &= \mathrm{alg}\{W(f), \mathrm{supp} f \subset I\} \end{aligned} \tag{6c}$$

where

$$s(.,.) = \int \frac{\mathrm{d}z}{2\pi\mathrm{i}} f'(z)g(z)$$

is the symplectic form which characterizes the Weyl algebra structure and [6c] denotes the unique C^* algebra generated by the unitary objects $W(f)$. A particular representation of this algebra is given by assigning the vacuum state to the generators $\langle W(f)\rangle_0 = e^{-(1/2)\|f\|_0^2}$, $\|f\|_0^2 = \sum_{n\geq 1} n|f_n|^2$. Starting with the vacuum Hilbert space representation $\mathcal{A}(S^1)_0 = \pi_0(\mathcal{A}(S^1))$, one easily checks that the formula

$$\langle W(f)\rangle_\alpha := \mathrm{e}^{\mathrm{i}\alpha f_0} \langle W(f)\rangle_0 \tag{7a}$$

$$\pi_\alpha(W(f)) = \mathrm{e}^{\mathrm{i}\alpha f_0} \pi_0(W(f)) \tag{7b}$$

defines a state with positive energy, that is, one whose GNS representation for $\alpha \neq 0$ is unitarily inequivalent to the vacuum representation. Its incorporation into the vacuum Hilbert space [7b] is part of the DHR formalism. It is convenient to view this change as the result of an application of an automorphism γ_α on the C^*-Weyl algebra $\mathcal{A}(S^1)$ which is implemented by a unitary charge-generating operator Γ_α in a larger (nonseparable) Hilbert space which contains all charge sectors $H_\alpha = \Gamma_\alpha H_0$, $H_0 \equiv H_{\mathrm{vac}} = \overline{\mathcal{A}(S^1)\Omega}$:

$$\begin{aligned} \langle W(f)\rangle_\alpha &= \langle \gamma_\alpha(W(f))\rangle_0 \\ \gamma_\alpha(W(f)) &= \Gamma_\alpha W(f) \Gamma_\alpha^* \end{aligned} \tag{8}$$

$\Gamma_\alpha \Omega = \Omega_\alpha$ describes a state with a rotational homogeneous charge distribution; arbitrary charge distributions ρ_α of total charge α that is, $\int (\mathrm{d}z/2\pi\mathrm{i})\rho_\alpha = \alpha$ are obtained in the form

$$\psi^\zeta_{\rho_\alpha} = \eta(\rho_\alpha) W(\hat{\rho}^\zeta_\alpha) \Gamma_\alpha \tag{9}$$

where $\eta(\rho_\alpha)$ is a numerical phase factor and the net effect of the Weyl operator is to change the rotational homogeneous charge distribution into ρ_α. The necessary charge-neutral compensating function ρ^ζ_α in the Weyl cocycle $W(\rho^\zeta_\alpha)$ is uniquely determined in terms of ρ_α up to the choice of one point $\zeta \in S^1$ (the determining equation involves the $\ln z$ function which needs the specification of a branch cut (Schroer 2005)). From this formula, one derives the commutation relations $\psi^\zeta_{\rho_\alpha} \psi^\zeta_{\rho_\beta} = \mathrm{e}^{\pm \mathrm{i}\pi\alpha\beta} \psi^\zeta_{\rho_\beta} \psi^\zeta_{\rho_\alpha}$ for spacelike separations of the ρ supports; hence, these fields are relatively local (bosonic) for $\alpha\beta = 2\mathbb{Z}$. In particular, if only one

type of charge is present, the generating charge is $\alpha_{\text{gen}} = \sqrt{2N}$ and the composite charges are multiples, that is, $\alpha_{\text{gen}}\mathbb{Z}$. This locality condition providing bosonic commutation relations does not yet ensure the ζ-independence. Since the equation which controls the ζ-change turns out to be

$$\psi_{\rho_\alpha}^{\zeta_1}\left(\psi_{\rho_\alpha}^{\zeta_2}\right)^* = e^{\pm i\pi\alpha\beta}e^{2\pi iQ\alpha} \qquad [10]$$

one achieves ζ-independence by restricting the Hilbert space charges to be "dual" to that of the operators, that is,

$$Q = \left\{\frac{1}{\sqrt{2N}}\mathbb{Z}\right\}$$

The localized $\psi_{\rho_\alpha}^{\zeta_1}$ operators acting on the restricted separable Hilbert space H_{res} generate a ζ-independent extended observable algebra $\mathcal{A}_N(S^1)$ (Schroer) and it is not difficult to see that its representation in H_{res} is reducible and that it decomposes into $2N$ charge sectors

$$\left\{\frac{1}{\sqrt{2N}}n, n = 0, 1, \ldots, N-1\right\}$$

Hence, the process of extension has led to a charge quantization with a finite ("rational") number of charges relative to the new observable algebra which is neutral in the new charge counting

$$\frac{1}{\alpha_{\text{gen}}}\mathbb{Z}/\alpha_{\text{gen}}\mathbb{Z} = \mathbb{Z}/\alpha_{\text{gen}}^2 \simeq \mathbb{Z}_{2N}$$

The charge-carrying fields in the new setting are also of the above form [9], but now the generating field carries the charge

$$\int \frac{dz}{2\pi i}\rho_{\text{gen}} = Q_{\text{gen}}$$

which is a $(1/2N)$ fraction of the old α_{gen}. Their commutation relations for disjoint charge supports are "braidal" (or better "plektonic" which is more on par with being bosonic/fermionic). (In the abelian case like the present, the terminology "anyonic" enjoys widespread popularity, but in the present context the "any" does not go well with charge quantization.) These objects considered as operators localized on S^1 do depend on the cut ζ, but using an appropriate finite covering of S^1 this dependence is removed (Schroer 2005). So the field algebra $\mathcal{F}_{\mathbb{Z}2N}$ generated by the charge-carrying fields (as opposed to the bosonic observable algebra $\mathcal{A}_N$) has its unique localization structure on a finite covering of S^1. An equivalent description which gets rid of ζ consists in dealing with operator-valued sections on S^1. The extension $\mathcal{A} \rightarrow \mathcal{A}_N$, which renders the Hilbert space separable and quantizes the charges, seems to be characteristic for abelian current algebra; in all other models which have been constructed up to now the number of sectors is at least denumerable and in the more interesting ones even finite (rational models). An extension is called maximal if there exists no further extension which maintains the bosonic commutation relation. For the case at hand, this would require the presence of another generating field of the same kind as above, which belongs to an integer N' is relatively local to the first one. This is only possible if N is divisible by a square.

In passing, it is interesting to mention a somewhat unexpected relation between the Schwinger model, whose charges are screened, and the Jordan model. Since the Lagrangian formulation of the Schwinger model is a gauge theory, the analog of the four-dimensional "asymptotic freedom" wisdom would suggest the possibility of "charge liberation" in the short-distance limit of this model. This seems to contradict the statement that the intrinsic content of the Schwinger model (QED_2 with massless Fermions) (after removing a classical degree of freedom) is the QFT of a free massive Bose field and such a simple free field is at first sight not expected to contain subtle information about asymptotic charge liberation. (In its original gauge-theoretical form, the Schwinger model has an infinite vacuum degeneracy. The removal of this degeneracy (restoration of the cluster property) with the help of the "θ-angle formalism" leaves a massive free Bose field (the Schwinger–Higgs mechanism). As expected in $d = 1 + 1$ the model only possesses this phase.) Well, as we have seen above, the massless limit really does have liberated charges and the short-distance limit of the massive free field is the massless model (Schroer).

As a result of the peculiar bosonization/fermionization aspect of the zero-mass limit of the derivative of the massive free field, Jordan's model is also closely related to the massless Thirring model (and the related Luttinger model for an interacting one-dimensional electron gas) whose massive version is in the class of factorizing models (see later section). (Another structural consequence of this aspect leads to Coleman's theorem (Schroer 2005) which connects the Mermin–Wagner no-go theorem for two-dimensional spontaneous continuous symmetry breaking with these zero-mass peculiarities.) The Thirring model is a special case in a vast class of "generalized" multicoupling multicomponent Thirring models, that is, models with 4-fermion interactions. Under this name they were studied in the early 1970s (Schroer) with the aim to identify massless subtheories for which the currents form chiral current algebras.

The counterpart of the potential of the conserved Dirac current in the massive Thirring model is the sine-Gordon field, that is, a composite field which in the attractive regime of the Thirring coupling again obeys the so-called sine-Gordon equation of motion. Coleman gave a supportive argument (Schroer 2005) but some fine points about the range of its validity in terms of the coupling strength remained open. (It was noticed that the current potential of the free massive Dirac Fermion ($g = 0$) does not obey the sine-Gordon equation (Schroer 2005).) A rigorous confirmation of these facts was recently given in the bootstrap form-factor setting (Schroer 2005). Massive models which have a continuous or discrete internal symmetry have "disorder" fields which implement a "half-space" symmetry on the charge-carrying field (acting as the identity in the other half-axis) and together with the basic pointlike field form composites which have exotic commutation relations (see the last section).

The Conformal Setting, Structural Results

Chiral theories play a special role within the setting of conformal quantum fields. General conformal theories have observable algebras which live on compactified Minkowski space (S^1 in the case of chiral models) and fulfill the Huygens principle, which in an even number of spacetime dimension means that the commutator is only nonvanishing for lightlike separation of the fields. The fact that this classically expected behavior breaks down for nonobservable conformal fields (e.g., the massless Thirring field) was noticed at the beginning of the 1970s and considered paradoxical at that time ("reverberation" in the timelike (Huygens) region). Its resolution around 1974–75 confirmed that such fields are genuine conformal covariant objects but that some fine points about their causality needed to be addressed. The upshot was the proposal of two different but basically equivalent concepts about globally causal fields. They are connected by the following global decomposition formula:

$$A(x_{\mathrm{cov}}) = \sum A_{\alpha,\beta}(x),\ A_{\alpha,\beta}(x) = P_\alpha A(x) P_\beta,\ Z = \sum \mathrm{e}^{\mathrm{i}d_\alpha} P_\alpha \qquad [11]$$

On the left-hand side, the spacetime point of the field is a point on the universal covering of the conformal compactified Minkowski space. These are fields (Lüescher and Mack 1975) (Schroer 2005) which "live" in the sense of quantum (modular) localization on the universal covering spacetime (or on a finite covering, depending on the "rationality" of the model) and fulfill the global causality condition previously discovered by I Segal (Schroer 2005). They are generally highly reducible with respect to the center of the covering group. The family of fields on the right-hand side, on the other hand, are fields which were introduced (Schroer and Swieca 1974; Schroer *et al.* 1975) with the aim to have objects which live on the projection $x(x_{\mathrm{cov}})$, that is, on the spacetime of the physics laboratory instead of the "hells and heavens" of the covering (Schroer 2005). They are operator-distributional valued sections in the compactification of ordinary Minkowski spacetime. The connection is given by the above decomposition formula into irreducible conformal blocks with respect to the center $\mathbf{Z}$ of the noncompact covering group $\widetilde{\mathrm{SO}(2,n)}$ where α, β are labels for the eigenspaces of the generating unitary Z of the abelian center $\mathbf{Z}$. The decomposition [11] is minimal in the sense that in general there generally will be a refinement due to the presence of additional charge superselection rules (and internal group symmetries). The component fields are not Wightman fields since they annihilate the vacuum if the right-hand projection differs from $P_0 = P_{\mathrm{vac}}$.

Note that the Huygens (timelike) region in Minkowski spacetime has a timelike ordering structure $x \prec y$ or $x \succ y$ (earlier or later). In $d = 1+1$, the topology allows in addition a spacelike left–right ordering $x \lessgtr y$. In fact, it is precisely the presence of these two orderings in conjunction with the factorization of the vacuum symmetry group $\widetilde{\mathrm{SO}(2,2)} \simeq \widetilde{\mathrm{PSL}(2R)}_l \otimes \widetilde{\mathrm{PSL}(2,R)}_r$, in particular $\mathbf{Z} = \mathbf{Z}_l \otimes \mathbf{Z}_r$, which is at the root of a significant simplification. This situation suggested a tensor factorization into chiral components and led to an extremely rich and successful construction program of two-dimensional conformal QFT as a two-step process: the classification of chiral observable algebras on the light ray and the amalgamation of left–right chiral theories to two-dimensional local conformal QFT. The action on the circular coordinates z is through fractional SU(1, 1) transformations

$$g(z) = \frac{\alpha z + \beta}{\bar{\beta} z + \bar{\alpha}}$$

whereas the covering group acts on the Mack–Luescher covering coordinates.

The presence of an ordering structure permits the appearance of more general commutation relations for the above $A_{\alpha\beta}$ component fields namely

$$A_{\alpha,\beta}(x) B_{\beta,\gamma}(y) = \sum_{\beta'} R^{\alpha,\gamma}_{\beta,\beta'} B_{\alpha,\beta'}(y) A_{\beta',\gamma}(x), \quad x > y \qquad [12]$$

with numerical R-coefficients which, as a result of associativity and relative commutativity with respect to observable fields, have to obey certain structure relations; in this way, Artin braid relations emerge as a new manifestation of the Einstein causality principle for observables in low-dimensional QFT (Rehren and Schroer 1989) (see Schroer 2005). Indeed, the DHR method to interpret charged fields as charge superselection carriers (tied by local representation theory to the bosonic local structure of observable algebras) leads precisely to such a plektonic statistics structure (Fredenhagen *et al.* 1992, Gabbiani and Froehlich 1993) for systems in low spacetime dimension (*see* Symmetries in Quantum Field Theory of Lower Spacetime Dimensions). With an appropriately formulated adjustment to observables fulfilling the Huygens commutativity, this plektonic structure (but now disconnected from particle/field statistics) is also a possible manifestation of causality for the higher-dimensional timelike structure (Schroer 2005).

The only examples known up to the appearance of the seminal BPZ work (Belavin *et al.* 1984) were the abelian current models of the previous section which furnish a rather poor man's illustration of the richness of the decomposition theory. The floodgates of conformal QFT were only opened after the BPZ discovery of "minimal models," which was preceded by the observation (Friedan *et al.* 1984) that the algebra of the stress–energy tensor came with a new representation structure which was not compatible with an underlying internal group symmetry (*see* Symmetries in Quantum Field Theory: Algebraic Aspects).

An important step in the structural study of chiral models was the recognition that the energy–momentum tensor has the commutation structure of a Lie field (Schroer 2005); in the next section, its algebraic structure and its representation theory will be presented.

Chiral Fields and Two-Dimensional Conformal Models

Let us start with a family which generalizes the abelian model of the previous section. Instead of a one-component abelian current we now take n independent copies. The resulting multicomponent Weyl algebra has the previous form except that the current is n-component and the real function space underlying the Weyl algebra consists of functions with values in an n-component real vector space $f \in LV$ with the standard Euclidean inner product denoted by $(,)$. The local extension now leads to $(\alpha, \beta) \in 2\mathbb{Z}$, that is, an even-integer lattice $\mathcal{L}$ in V, whereas the restricted Hilbert subspace $H_{\mathcal{L}^*}$ which ensures ζ-independence is associated with the dual lattice L^*: $(\lambda_i, \alpha_k) = \delta_{ik}$ which contains $\mathcal{L}$. The resulting superselection structure (i.e., the Q-spectrum) corresponds to the finite factor group $\mathcal{L}^*/\mathcal{L}$. For self-dual lattices $\mathcal{L}^* = \mathcal{L}$ (which only can occur if dim V is a multiple of 8), the resulting observable algebra has only the vacuum sector; the most famous case is the Leech lattice Λ_{24} in $\dim V = 24$, also called the "moonshine" model. The observation that the root lattices of the Lie algebras of types A, B, or E (e.g., su(n) corresponding to A_{n-1}) also appear among the even-integral lattices suggests that the nonabelian current algebras associated to those Lie algebras can also be implemented. This turns out to be indeed true as far as the level-1 representations are concerned which brings us to the second family: the nonabelian current algebras of level k associated to those Lie algebras; they are characterized by the commutation relation

$$[J_\alpha(z), J_\beta(z')] = \mathrm{i} f^\gamma_{\alpha\beta} j_\gamma(z)\delta(z - z') - \tfrac{1}{2} k g_{\alpha\beta}\delta'(z - z') \qquad [13]$$

where $f^\gamma_{\alpha\beta}$ are the structure constants of the underlying Lie algebra, g their Cartan–Killing form, and k, the level of the algebra, must be an integer in order that the current algebra can be globalized to a loop group algebra. The Fourier decomposition of the current leads to the so-called affine Lie algebras, a special family of Kac–Moody algebras. For $k = 1$, these currents can be constructed as bilinears in terms of the multicomponent chiral Dirac field; there exists also the mentioned possibility to obtain them by constructing their maximal Cartan currents within the above abelian setting and representing the remaining nondiagonal currents as certain charge-carrying ("vertex" algebra) operators. Level-k algebras can be constructed from reducing tensor products of k level-1 currents or directly via the representation theory of infinite-dimensional affine Lie algebras. (The global exponentiated algebras (the analogs to the Weyl algebra) are called loop group algebras.) Either way one finds that, for example, the SU(2) current algebra of level k has (together with the vacuum sector) $k + 1$ sectors (inequivalent representations). The different sectors are already distinguished by the structure of their ground states of the conformal Hamiltonian L_0. Although the computation of higher point correlation functions for $k > 1$, there is no problem in securing the existence of the algebraic nets which define these chiral models as well as their $k + 1$

representation sectors and to identify their generating charge-carrying fields (primary fields) including their R-matrices appearing in their plektonic commutation relations. It is customary to use the notation $SU(2)_k$ for the abstract operator algebras associated with the current generators [13] and we will denote their $k+1$ equivalence classes of representations by $\{\mathcal{A}_{SU(2)_k,n}|n=0,\ldots,k$, whereas representations of current algebras for higher rank groups require a more complicated labeling (in terms of Weyl chambers).

The third family of models are the so-called minimal models which are associated with the Lie-field commutation structure of the chiral stress–energy tensor which results from the chiral decomposition of a conformally covariant two-dimensional stress–energy tensor

$$[T(z),T(z')] = \mathrm{i}(T(z)+T(z'))\delta'(z-z') + \frac{\mathrm{i}c}{24\pi}\delta'''(z-z') \qquad [14]$$

whose Fourier decomposition yields the Witt–Virasoro algebra, that is, a central extension of the Lie algebra of the Diff(S^1). (The presence of the central term in the context of QFT (the analog of the Schwinger term) was noticed later; however, the terminology Witt–Virasoro algebra in the physics literature came to mean the Lie algebra of diffeomorphisms of the circle including the central extension.) The first two coefficients are determined by the physical role of $T(z)$ as the generating field density for the Lie algebra of the Poincaré group whereas the central extension parameter $c>0$ (positivity of the two-point function) for the connection with the generation of the Moebius transformations and the undetermined parameter $c>0$ (the central extension parameter) is easily identified with the strength of the two-point function. Although the structure of the T-correlation functions resembles that of free fields (in the sense that is an algebraically computable unique set of correlation functions once one has specified the two-point function), the realization that c is subject to a discrete quantization if $c<1$ came as a surprise. As already mentioned, the observation that the superselection sectors (the positive-energy representation structure) of this algebra did not at all follow the logic of a representation theory of an inner symmetry group generated a lot of attention and stimulated a flurry of publications on symmetry concepts beyond groups (quantum groups). A concept of fundamental importance is the DHR theory of localized endomorphisms of operator algebras and the concept of operator-algebraic inclusions (in particular, inclusions with conditional expectations – V Jones inclusions).

The $SU(2)_k$ current coset construction (Goddard *et al.* 1985) revealed that the proof of existence and the actual construction of the minimal models is related to that of the $SU(2)_k$ current algebras. Constructing a chiral model does not necessarily mean the explicit determination of the n-point Wightman functions of their generating fields (which for most chiral models remains a prohibitively complicated task) but rather a proof of their existence by demonstrating that these models are obtained from free fields by a series of computational complicated but mathematically controlled operator-algebraic steps as reduction of tensor products, formation of orbifolds under group actions, coset constructions, and a special kind of extensions. The generating fields of the models are nontrivial in the sense of not obeying free-field equations (i.e., not being "on-shell"). The cases where one can write down explicit n-point functions of generating fields are very rare; in the case of the minimal family this is limited to the field theory of the Ising model (Schroer 2005).

To show the power of inclusion theory for the determination of the charge content of theory, let us look at a simple illustration in the context of the above multicomponent abelian current algebra. The vacuum representation of the corresponding Weyl algebra is generated from smooth V-valued functions on the circle modulo constant functions (i.e., functions with vanishing total integral) $f\in LV_0$. These functions equipped with the aforementioned complex structure and scalar product yield a Hilbert space. The I-localized subalgebra is generated by the Weyl image of I-supported functions (class functions whose representing functions are constant in the complement I')

$$\begin{aligned}\mathcal{A}(I) &:= \mathrm{alg}\{W(f)|f\in K(I)\}\\ K(I) &= \{f\in LV_0|f=\text{const.in } I'\}\end{aligned} \qquad [15]$$

The one-interval Haag duality $\mathcal{A}(I)'=\mathcal{A}(I')$ (the commutant algebra equals the algebra localized in the complement) is simply a consequence of the fact that the symplectic complement $K(I)'$ in terms of $\mathrm{Im}(f,g)$ consists of real functions in that space which are localized in the complement, that is, $K(I)'=K(I')$. The answer to the same question for a double interval $I=I_1\cup I_3$ (think of the first and third quadrant on the circle) does not lead to duality but rather to a genuine inclusion

$$\begin{aligned}K((I_1\cup I_3)') = K(I_2\cup I_4) &\subset K(I_1\cup I_3)'\\ K(I_1\cup I_3) &\subset K((I_1\cup I_3)')'\end{aligned} \qquad [16]$$

The meaning of the left-hand side is clear; these are functions which are constant in $I_1 \cup I_3$ with the same constant in the two intervals whereas the functions on the right-hand side are less restrictive in that the constants can be different. The conversion of real subspaces into von Neumann algebras by the Weyl functor leads to the algebraic inclusion $\mathcal{A}(I_1 \cup I_3) \subset \mathcal{A}((I_1 \cup I_3)')'$. In physical terms, the enlargement results from the fact that within the charge neutral vacuum algebra a charge split with one charge in I_1 and the compensating charge in I_2 for all values of the (unquantized) charge occurs. A more realistic picture is obtained if one allows a charge split to be subjected to a charge quantization implemented by a lattice condition $f(I_2) - f(I_4) \in 2\pi L$ which relates the two multicomponent constant functions (where $f(I)$ denotes the constant value f takes in I). As in the previous one-component case, the choice of even lattices corresponds to the local (bosonic) extensions. Although imposing such a lattice structure destroys the linearity of the K, the functions still define Weyl operators which generated operator algebras $\mathcal{A}_L(I_1 \cup I_2)$. (The linearity structure is recovered on the level of the operator algebra.) But now the inclusion involves the dual lattice L^* (which of course contains the original lattice),

$$\mathcal{A}_L(I_1 \cup I_2) \subset \mathcal{A}_{L^*}(I_1 \cup I_2)$$
$$\mathrm{ind}\{\mathcal{A}_L(I_1 \cup I_2) \subset \mathcal{A}_L((I_1 \cup I_2)')'\} = |G|$$
$$\mathcal{A}_L(I_1 \cup I_2) = \mathrm{inv}_G \mathcal{A}_{L^*}(I_1 \cup I_2)$$

This time the possible charge splits correspond to the factor group $G = L^*/L$, that is, the number of possibilities is $|G|$ which measures the relative size of the bigger algebra in terms of the smaller. This is a special case of the general concept of the so-called Jones index of an inclusion which is a numerical measure of its depth. A prerequisite is that the inclusion permits a conditional expectation which is a generalization of the averaging under the "gauge group" G on $\mathcal{A}_{L^*}(I_1 \cup I_2)$ in the third equation above, which identifies the invariant smaller algebra with the fix-point algebra (the invariant part) under the action of G. In fact, using the conceptual framework of Jones, one can show that the two-interval inclusion is independent of the position of the disjoint intervals characterized by the group G.

There exists another form of this inclusion which is more suitable for generalizations. One starts from the charge quantized extended local algebra $\mathcal{A}_L^{\mathrm{ext}} \supset \mathcal{A}$ described earlier in terms of an even-integer lattice L (which lives in the separable Hilbert space H_{L^*}) as our observable algebra. Again the Haag duality is violated and converted into an inclusion $\mathcal{A}_L^{\mathrm{ext}}(I_1 \cup I_2) \subset \mathcal{A}_L^{\mathrm{ext}}((I_1 \cup I_2)')'$ which turns out to have the same $G = L^*/L$ charge structure (it is in fact isomorphic to the previous inclusion). In the general setting (current algebras, minimal model algebras, ...), this double interval inclusion is particularly interesting if the associated Jones index is finite. One finds Kawahigashi *et al.* (2001) (Schroer 2005).

Theorem 1 *A chiral theory with finite Jones index* $\mu = \mathrm{ind}\{\mathcal{A}((I_1 \cup I_2)')' : \mathcal{A}(I_1 \cup I_2)\}$ *for the double interval inclusion (always assuming that* $A(S^1)$ *is strongly additive and split) is a rational theory and the statistical dimensions* d_ρ *of its charge sectors are related to* μ *through the formula*

$$\mu = \sum_\rho d_\rho^2 \qquad [17]$$

Instead of presenting more constructed chiral models, it may be more informative to mention some of the algebraic methods by which they are constructed and explored. The already mentioned DHR theory provides the conceptual basis for converting the notion of positive-energy representation sectors of the chiral model observable algebras $\mathcal{A}$ (equivalence classes of unitary representations) into localized endomorphisms ρ of this algebra. This is an important step because contrary to group representations which have a natural tensor product composition structure, representations of operator algebras generally do not come with a natural composition structure. The DHR endomorphisms theory of $\mathcal{A}$ leads to fusion laws and an intrinsic notion of generalized statistics (for chiral theories: plektonic in addition to bosonic/fermionic). The chiral statistics parameters are complex numbers (Haag 1992) whose phase is related to a generalized concept of spin via a spin-statistics theorem and whose absolute value (the statistics dimension) generalized the notion of multiplicities of fields known from the description of inner symmetries in higher-dimensional standard QFTs. The different sectors may be united into one bigger algebra called the exchange algebra $\mathcal{F}_{\mathrm{red}}$ in the chiral context (the "reduced field bundle" of DHR) in which every sector occurs by definition with multiplicity 1 and the statistics data are encoded into exchange (commutation) relations of charge-carrying operators or generating fields ("exchange algebra fields") (Schroer 2005). Even though this algebra is useful in that all properties concerning fusion and statistics are nicely encoded, it lacks some cherished properties of standard field theory

namely there is no unique state–field relation, that is, no Reeh–Schlieder property (a field $A_{\alpha\beta}$ whose source projection P_β does not coalesce with the vacuum projection annihilates the vacuum); in operator-algebraic terms, the local algebras are not factors. This poses the question of how to manufacture from the set of all sectors natural (not necessarily local) extensions with these desired properties. It was found that this problem can be characterized in operator-algebraic terms by the existence of the so-called DHR triples (Schroer). In case of rational theories, the number of such extensions is finite and in the aforementioned "classical" current algebra and minimal models they all have been constructed by this method (thus confirming existing results completing the minimal family by adding some missing models). The same method adapted to the chiral tensor product structure of $d=1+1$ conformal observables classifies and constructs all two-dimensional local (bosonic/ fermionic) conformal QFT $\mathcal{B}_2$ which can be associated with the observable chiral input. It turns out that this approach leads to another of those pivotal numerical matrices which encode structural properties of QFT: the coupling matrix Z,

$$\mathcal{A}\otimes\mathcal{A}\subset\mathcal{B}_2$$
$$\sum_{\rho\sigma} Z_{\rho,\sigma}\rho(\mathcal{A})\otimes\sigma(\mathcal{A})\subset\mathcal{A}\otimes\mathcal{A} \qquad [18]$$

where the second line is an inclusion solely expressed in terms of observable algebras from which the desired (isomorphic) inclusion in the first line follows by a canonical construction, the so-called Jones basic construction. The numerical matrix Z is an invariant closely related to the so-called "statistics character matrix" (Schroer 2005) and in case of rational models it is even a modular invariant with respect to the modular $SL(2,Z)$ group transformations (which are closely related to the matrix S in the final section).

Integrability, the Bootstrap Form-Factor Program

Integrability in QFT and the closely associated bootstrap form-factor construction of a very rich class of massive two-dimensional QFTs can be traced back to two observations made during the 1960s and 1970s ideas. On the one hand, there was the time-honored idea to bypass the "off-shell" field-theoretic approach to particle physics in favor of a pure on-shell S-matrix setting which (in particular recommended for strong interactions), as a result of the elimination of short distances via the mass-shell restriction, would be free of ultraviolet divergencies. This idea was enriched in the 1960s by the crossing property which in turn led to the bootstrap idea, a highly nonlinear seemingly self-consistent proposal for the determination of the S-matrix. However, the protagonists of this S-matrix bootstrap program placed themselves into a totally antagonistic fruitless position with respect to QFT so that the strong return of QFT in the form of gauge theory undermined their credibility. On the other hand, there were rather convincing quasiclassical calculations in certain two-dimensional massive QFTs as, for example, the sine-Gordon model which indicated that the obtained quasiclassical mass spectrum is exact and hence suggested that the associated QFTs are integrable (Dashen *et al.* 1975) and have no real particle creation. These provocative observations asked for a structural explanation beyond quasiclassical approximations, and it soon became clear that the natural setting for obtaining such mass formulas was that of the "fusion" of boundstate poles of unitary crossing-symmetric purely elastic S-matrices; first in the special context of the sine-Gordon model (Schroer *et al.* 1976) and later as a classification program from which factorizing S-matrices can be determined by solving well-defined equations for the elastic two-particle S-matrix (Karowski *et al.* 1977). (It was incorrectly believed that the "nontrivial elastic scattering implies particle creation" statement of Aks (Aks, 1963) is also valid for low-dimensional QFTs.) Some equations in this bootstrap approach resembled mathematical structures which appeared in C N Yang's work on nonrelativistic δ-function particle interactions as well as relations for Boltzmann weights in Baxter's work on solvable lattice models; hence, they were referred to as Yang–Baxter relations. These results suggested that the old bootstrap idea, once liberated from its ideological dead freight (in particular from the claim that the bootstrap leads to a unique "theory of everything" (minus gravity)), generates a useful setting for the classification and construction of factorizing two-dimensional relativistic S-matrices. Adapting certain known relations between two-particle form factors of field operators and the S-matrix to the case at hand (Karowski and Weisz 1978), and extending this with hindsight to generalized (multiparticle) form factors, one arrived at the axiomatized recipes of the bootstrap form-factor program of $d=1+1$ factorizable models (Smirnov 1992). Although this approach can be formulated within the

setting of the LSZ scattering formalism, the use of a certain algebraic structure (Zamolodchikov and Zamolodchikov 1979) which in the simplest version reads

$$Z(\theta)Z^*(\theta') = S^{(2)}(\theta-\theta')Z^*(\theta')Z(\theta) + \delta(\theta-\theta')$$
$$Z(\theta)Z(\theta') = S^{(2)}(\theta'-\theta)Z(\theta')Z(\theta) \qquad [19]$$

(the δ-term Faddeev is due to Faddeev) brought significant simplifications. In the general case, the $Z's$ are vector valued and the $S^{(2)}$-structure function is matrix valued. (The identification of the Z–F structure coefficients with the elastic two-particle S-matrix $S^{(2)}$ (which is prenempted by our notation) can be shown to follow from the physical interpretation of the Z-F structure in terms of localization.) In that case the associativity of the Z–F algebra is equivalent to the Yang–Baxter equations. Recently, it became clear that this algebraic relation has a deep physical interpretation; it is the simplest algebraic structure which can be associated with generators of nontrivial wedge-localized operator algebras (see the next section).

Conceptually as well as computationally it is much simpler to identify the intrinsic meaning of integrability in QFT with the factorization of its S-matrix or a certain property of wedge-localized algebras (see next section) than to establish integrability (*see* Integrability and Quantum Field Theory).

The first step of the bootstrap form-factor program namely the classification and construction of model S-matrices follows a combination of two patterns: prescribing particle multiplets transforming according to group symmetries and/or specifying structural properties of the particle spectrum. The simplest illustration for the latter strategy is supplied by the $\mathbb{Z}_N$ model. In terms of particle content, $\mathbb{Z}_N$ demands the identification of the Nth bound state with the antiparticle. Since the fusion condition for the bound mass $m_b^2 = (p_1+p_2)^2 = m_1^2 + m_2^2 + 2m_1 m_2 ch(\theta_1-\theta_2)$ is only possible for a pure imaginary rapidity difference $\theta_{12} = \theta_1 - \theta_2 = i\alpha$ ("binding angle"). Hence, the binding of two "elementary" particles of mass m gives

$$m_2 = m\frac{\sin 2\alpha}{\sin\alpha}$$

and more generally of k particles with

$$m_k = m\frac{\sin k\alpha}{\sin\alpha}$$

so that the antiparticle mass condition $m_N = \bar{m} = m$ fixes the binding angle to $\alpha = 2\pi/N$. (The quotation mark is meant to indicate that in contrast to the Schrödinger QM there is "nuclear democracy" on the level of particles. The inexorable presence of interaction-caused vacuum polarization limits a fundamental/fused hierarchy to the fusion of charges.) The minimal (no additional physical poles) two-particle S-matrix in terms of which the n-particle S-matrix factorizes is therefore

$$S_{\min}^{(2)} = \frac{\sin(1/2)(\theta + (2\pi i)/N)}{\sin(1/2)(\theta - (2\pi i)/N)} \qquad [20]$$

(minimal = without so-called CDD poles) The SU(N) model as compared with the U(N) model requires a similar identification of bound states of $N-1$ particles with an antiparticle. This S-matrix enters as in the equation for the vacuum to n-particle meromorphic form factor of local operators; together with the crossing and the so-called "kinematical pole equation," one obtains a recursive infinite system linking a certain residue with a form factor involving a lower number of particles. The solutions of this infinite system form a linear space from which the form factors of specific tensor fields can be selected by a process which is analogous but more involved than the specification of a Wick basis of composite free fields. Although the statistics property of two-dimensional massive fields is not intrinsic but a matter of choice, it would be natural to realize, for example, the $\mathbb{Z}_N$ fields as $\mathbb{Z}_N$-anyons.

Another rich class of factorizing models are the Toda theories of which the sine-Gordon and sinh–Gordon are the simplest cases. For their descriptions, the quasiclassical use of Lagrangians (supported by integrability) turns out to be of some help in setting up their more involved bootstrap form-factor construction.

The unexpected appearance of objects with new fundamental (solitonic) charges (e.g., the Thirring field as the carrier of a solitonic sine-Gordon charge) and the unexpected confinement of charges (e.g., the CP(1) model as a confined SU(2) model) turn out to be opposite sides of the same coin and both cases have realizations in the setting of factorizing models (Schroer 2005).

Recent Developments

There are two ongoing developments which place the two-dimensional bootstrap form-factor program into a more general setting which permits to understand its position in the general context of local quantum physics.

One of these starts from the observation that the smallest spacetime localization region in which it is possible to find vacuum-polarization-free generators (PFG) in the presence of interactions is the wedge

region. If one demands in addition that these generators (necessarily unbounded operators) have the standard domain properties of QFT (which include stability of the domain under translations), then one finds that this leads precisely to the two-dimensional Z–F algebraic structure which in turn in this way a spacetime interpretation for the first time acquires. In these investigations (Schroer 2005), modular localization theory plays a prominent role and there are strong indications that with these methods one can show the nontriviality of intersections of wedge algebras which is the algebraic criterion for the existence of a model within local quantum physics.

There is a second constructive idea based on light-front holography which uses the radical reorganization of spacetime properties of the algebraic structure while maintaining the physical content including the Hilbert space. Since spacetime localization aspects (apart from the remark about wedge algebras and their PFG generators made before) are traditionally related to the concept of fields, holographic methods tend to de-emphasize the particle structure in favor of "field properties." Indeed, the transversely extended chiral theories which arise as the holographic image lead to simplification of many interesting properties with very similar aims to the old "light-cone quantization" except that light-front holography is another way of looking at the original local ambient theory without subjecting it to another quantization. (The price for this simplification is that as a result of the nonuniqueness of the holographic inversion certain problems cannot be formulated.)

Actually, as a result of the absence of a transverse direction in the two-dimensional setting, the family of factorizing models provides an excellent theoretical laboratory to study their rigorous "chiral encoding" which is conceptually very different from Zamolodchikov's perturbative relation (which is based on identifying a factorizing model in terms of a perturbation on a chiral theory).

It turns out that the issue of statistics of particles loses its physical relevance for two-dimensional massive models since they can be changed without affecting the physical content. Instead such notions as order/disorder fields and soliton take their place (Schroer 2005).

In accordance with its historical origin, the theory of two-dimensional factorizing models may also be viewed as an outgrowth of the quantization of classical integrable systems (Integrability and Quantum Field Theory). But in comparison with the rather involved structure of integrabilty (verifying the existence of sufficiently many commuting conservation laws), the conceptual setting of factorizing models within the scattering framework (factorization follows from existence of wedge-localized tempered PFGs) is rather simple and intrinsic (Schroer 2005).

Among the additional ongoing investigations in which the conceptual relation with higher-dimensional QFT is achieved via modular localization theory, we will select three which have caught our, active attention. One is motivated by the recent discovery of the adaptation of Einsteins classical principle of local covariance to QFT in curved spacetime. The central question raised by this work (*see* Algebraic Approach to Quantum Field Theory) is if all models of Minkowski spacetime QFTs permit a local covariant extension to curved spacetime and if not which models do? In the realm of chiral QFT, this would amount to ask if all Moebius-invariant models are also Diff(S^1)-covariant. It has been known for sometime that a QFT with all its rich physical content can be uniquely defined in terms of a carefully chosen relative position of a finite number of copies of one unique von Neumann operator algebra within one common Hilbert space. This is a perfect quantum field-theoretical illustration for Leibnitz's philosophical proposal that reality results from the relative position of "monades" (As opposed to the more common (Newtonian) view that the material reality originates from a material content being placed into a spacetime vessel) if one takes the step of identifying the hyperfinite typ III_1 Murray von Neumann factor algebra with an abstract monade from which the different copies result from different ways of positioning in a shared Hilbert space (Schroer 2005). In particular, Moebius-covariant chiral QFTs arise from two monades with a joint intersection defining a third monade in such a way that the relative positions are specified in terms of natural modular concepts (without reference to geometry). This begs the question whether one can extend these modular-based algebraic ideas to pass from the global vacuum preserving Moebius invariance to local Diff(S) covariance Moeb $\rightarrow$ Diff(S^1). This would be precisely the two-dimensional adaptation of the crucial problem raised by the recent successful generalization of the local covariance principle underlying Einstein's classical theory of gravity to QFT in curved spacetime: does every Poincaré covariant Minkowski spacetime QFT allow a unique correspondence with one curved spacetime (having the same abstract algebraic substrate but with a totally different spacetime encoding)? In the chiral context, one is led to the notion of "partially geometric modular groups" which only act geometrically if restricted to specific subalgebras (Schroer

2005). It is hard to imagine how one can combine quantum theory and gravity without understanding first the still mysterious links between spacetime geometry, thermal properties, and relative positioning of monades in a joint Hilbert space.

A second important umbilical cord with higher-dimensional theories is the issue of "Euclideanization" in particular the chiral counterpart of Osterwalder–Schrader localization and the closely related Nelson–Symanzik duality. In concrete chiral models (e.g., the models in the section "Chiral fields and two-dimensional conformal models"), it has been noted as a result of explicit calculations that the analytic continuation in the angular parametrization for thermal correlation functions leads to a duality relation in

$$\langle A(\varphi_1,\ldots,\varphi_n)\rangle_{\alpha,2\pi\beta_t} = \left(\frac{\mathrm{i}}{\beta_t}\right)^a \sum_\gamma S_{\alpha\gamma}\left\langle A\left(\frac{\mathrm{i}}{\beta_t}\varphi_1,\ldots,\frac{\mathrm{i}}{\beta_t}\varphi_n\right)\right\rangle_{\gamma,(2\pi/\beta_t)} \quad [21]$$

where the thermal correlation function is defined as

$$\langle A(\varphi_1,\ldots,\varphi_n)\rangle_{\rho_\alpha,2\pi\beta_t} := \mathrm{tr}_{H_{\rho_\alpha}} \mathrm{e}^{-2\pi\beta_t\left(L_0^{\rho_\alpha}-(c/24)\right)}\pi_{\rho_\alpha}(A(\varphi_1,\ldots,\varphi_n)) \quad [22]$$

$$A(\varphi_1,\ldots,\varphi_n) = \prod_{i=1}^n A_i(\varphi_i)$$

Compared with the thermally extended Nelson–Symanzik relation for two-dimensional QFT one notices that in addition to the expected behavior of real coordinates becoming imaginary and the 2π-periodicity changing role with the (suitably normalized) KMS inverse temperature, there is a rotation in the space of superselected charges in terms of a unitary matrix S whose origin lies in the braid group statistics (the statistics character matrix). The deeper structural explanation which shows that this relation is not just a property of special models, but rather a generic property of chiral QFT, comes from a very deep angular Euclideanization which is based on modular theory (Schroer). Specializing $A =$ identity, one obtains a relation for the partition function, the famous Verlinde identity which is part of the transformation law of the thermal angular correlation functions under the SL(2, R) modular group.

There are many additional important observations on factorizing models whose relation to the physical principles of QFT, unlike the bootstrap form-factor program, is not yet settled. The meaning of the c-parameter outside the chiral setting and ideas on its renormalization group flow as well as the various formulations of the thermodynamic Bethe ansatz belong to a series of interesting observations whose final relation to the principles of QFT still needs clarification.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Bosons and Fermions in External Fields; Euclidean Field Theory; Integrablility and Quantum Field Theory; Operator Product Expansion in Quantum Field Theory; Sine-Gordon Equation; Symmetries in Quantum Field Theory: Algebraic Aspects; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Tomita–Takesaki Modular Theory.

Further Reading

Abdalla E, Abdalla MCB, and Rothe K (1991) *Non-Perturbative Methods in 2-Dimensional Quantum Field Theory*. Singapore: World Scientific.

Belavin AA, Polyakov AM, and Zamolodchikov AB (1984) Infinite conformal symmetry in two-dimensional quantum field theory. *Nuclear Physics B* 241: 333.

Bisognano JJ and Wichmann EH (1975) *Journal of Mathematical Physics* 16: 985.

Dashen F, Hasslacher B, and Neveu A (1975) *Physics Reviews D* 11: 3424.

Di Francesco P, Mathieu P, and Sénéchal D (1996) *Conformal Field Theory*. Berlin: Springer.

Doplicher S, Haag R, and Roberts JE (1971/1974) *Communications in Mathematical Physics* 23: 199; 35: 49.

Fredenhagen K, Rehren KH, and Schroer B (1992) Superselection sector with braid group statistics and exchange algebras II: Geometric aspects and conformal invariance. *Reviews of Mathematical Physics* 1 (special issue): 113.

Furlan P, Sotkov G, and Todorov I (1989) Two-dimensional conformal quantum field theory. *Rivista del Nuovo Cimento* 12: 1–203.

Gabbiani F and Froehlich J (1993) Operator algebras and conformal field theory. *Communications in Mathematical Physics* 155: 569.

Glimm J and Jaffe A (1987) *Quantum Physics. A Functional Integral Point of View*. Berlin: Springer.

Ginsparg P (1990) Applied conformal field theory. In: Brezin E and Zinn-Justin J (eds.) *Fields, Strings and Critical Phenomena*, Les Houches 1988. Amsterdam: North-Holland.

Goddard P, Kent A, and Olive D (1985) Virasoro algebras and coset space models. *Physics Letters B* 152: 88.

Haag R (1992) *Local Quantum Physics*. Berlin: Springer.

Ising E (1925) *Zeitschrift für physik* 31: 253.

Jordan P (1937) Beiträge zur Neutrinotheorie des Lichts. *Zeitschrift für Physik* 114: 229 and earlier papers quoted therein.

Karowski M, Thun H-J, Truoung TT, and Weisz P (1977) *Physics Letters B* 67: 321.

Karowski M and Weisz P (1978) *Physics Reviews B* 139: 445.

Kawahigashi Y, Longo R, and Mueger M (2001) Multi-interval subfactors and modularity in representations of conformal field theory. *Communications in Mathematical Physics* 219: 631.

Lenz W (1920) *Physikalische Zeitschrift* 21: 613.

Lüescher M and Mack G (1975) Global conformal invariance in quantum field theory. *Communications in Mathematical Physics* 41: 203.

Rehren K-H and Schroer B (1987) Exchange algebra and Ising n-point functions. *Physics Letters B* 198: 84.

Rehren KH and Schroer B (1989) Einstein causality and Artin braids. *Nuclear Physics B* 312: 715.

Schroer B (2005) Two-dimensional models, a testing ground for principles and concepts of QFT, *Annals of Physics* (in print) (hep-th/0504206).

Schroer B and Swieca JA (1974) Conformational transformations for quantized fields. *Physics Reviews D* 10: 480.

Schroer B, Swieca JA, and Voelkel AH (1975) Global operator expansions in conformally invariant relativistic quantum field theory. *Physics Reviews D* 11: 11.

Schroer B, Truong TT, and Weisz P (1976) Towards an explicit construction of the Sine-Gordon field theory. *Annals of Physics (New York)* 102: 156.

Schwinger J (1962) *Physical Review* 128: 2425.

Schwinger J (1963) Gauge theory of vector particles. In: *Theoretical Physics Trieste Lectures 1962*. Wien: IAEA.

Smirnov FA (1992) *Advanced Series in Mathematical Physics 14*. Singapore: World Scientific.

Streater RF and Wightman AS (1964) *PCT, Spin and Statistics and All That*. New York: Benjamin.

Zamolodchikov AB and Zamolodchikov AB (1979) *Annals of Physics (New York)* 120: 253.

Quantum Field Theory

Quantum Field Theory: A Brief Introduction

L H Ryder, University of Kent, Canterbury, UK

© 2006 Elsevier Ltd. All rights reserved.

By any account quantum field theory occupies a prominent place in the history of mathematical physics. This article is, however, not intended to serve as an overview of this subject, but has the more modest aim of identifying a few areas which seem to me interesting and significant.

Historical Remarks; Second Quantization

At the time when quantum field theory was at the forefront of theoretical physics its *raison d'être* was to complete the quantum description of the subatomic world. Quantum mechanics had been amazingly successful in solving almost the whole of atomic physics by making explicit the quantum (wave) nature of the electron, according to the formulations of Heisenberg and Schrödinger. The introduction of the quantum idea into physics, however, by Planck in 1900 closely followed by Einstein in 1905 was the proposal of a quantum (particular) aspect of the electromagnetic field – the photon. In the mid-1920s the only force in nature to be considered was the electromagnetic interaction; this was before the theories of Yukawa and Fermi, concerning the strong and weak nuclear forces. Dirac, Heisenberg, Jordan, and others then addressed themselves to finding a formulation of quantum electrodynamics (QED) comparable in mathematical sophistication to the Heisenberg–Schrödinger formulation of quantum mechanics – which Planck's and Einstein's theories were not.

The idea that was pursued, at least in the early stages, was that the Schrödinger wave function ψ, taken as a wave field, should be "quantized"; Dirac

Notes:

seems to have taken this as a model for photons. Jordan further proposed that electrons should be treated as the quanta of an electron field, but recognized that their fermionic nature would modify the quantization procedure. This generic idea involved what was called "second quantization" – of a field into a particle.

One of the earliest quantization rules was Bohr's condition relating to the periodic orbits of electrons in atoms, $J = \int p\, \mathrm{d}q = nh$. At the hands of Heisenberg and Dirac this became upgraded to the commutation relation

$$[q, p] = \mathrm{i}\hbar$$

where the operators p and q are "observables." In their papers on quantum field theory, Dirac, Jordan and Wigner, and Heisenberg introduced creation and annihilation operators which had the function, as their name implied, of creating and destroying single particles – quanta of the field. These operators obeyed the commutation rules (with $[A, B] = AB - BA$)

$$[b_r, b_s^*] = \delta_{rs}, \quad [b_r, b_s] = [b_r^*, b_s^*] = 0$$

when the field quanta were bosons, and the anti-commutation rules

$$\{b_r, b_s^*\} = \delta_{rs}, \quad \{b_r, b_s\} = \{b_r^*, b_s^*\} = 0$$

(with $\{A, B\} = AB + BA$) when the field quanta were fermions (e.g., electrons). These steps constitute second quantization, but it may be noted that the creation and annihilation operators are not observables, as p and q are in the Heisenberg commutation relation. In addition, the second quantization conditions do not involve Planck's constant. "First" and "second" quantization are therefore not so similar as one might like to think.

The question of what exactly is being quantized was in fact the source of some confusion. In his paper of 1927, Dirac's attention is focussed on electromagnetic radiation, but he nevertheless discusses the difference between "a light-wave and the de Broglie or Schrödinger wave associated with the light-quanta." As Dirac points out, "their intensities are to be interpreted in different ways. The number of light quanta per unit volume associated with a monochromatic light-wave equals the energy per unit volume of the wave divided by the energy $(2\pi h)\nu$ of a single light quantum. On the other hand a monochromatic de Broglie wave of amplitude a (multiplied into the imaginary exponential factor) must be interpreted as representing a^2 light quanta per unit volume for all frequencies." There are at least two problematic issues here. First, is the Schrödinger wave function ψ to be considered as a "real" field, whose quanta result in "real" particles, or is it a probability field, whose significance lies in Born's probabilistic interpretation of quantum mechanics? Born wrote in 1926, "[Einstein said that] the waves are present only to show the corpuscular light quanta the way, and he spoke in the sense of a "ghost field". This determines the probability that a light quantum, the bearer of energy and momentum, takes a certain path; however, the field itself has no energy and no momentum." This is the first problem. The second one concerns the nature of the quantization itself. Is this a quantization of field energy, or a quantization of the field itself, as a substantial entity? If the field is real, the second of these does not imply the first.

Ambiguities surrounding the idea of second quantization survived into the 1960s. Wigner is recorded as saying, in an interview in 1963, "just as we get photons by quantising the electromagnetic fields, so we should be able to get material particles by quantising the Schrödinger field." And Rosenfeld, also in an interview in 1963, said, "in some sense or other, Jordan himself took the wave function, the probability amplitude, physically more seriously than most people [did]."

It would seem we are justified in concluding that the idea of second quantization contains flaws, but an even clearer indication of the need for rethinking is provided by the story of the Dirac equation. This is a wave equation for the electron, compatible with special relativity, and taking explicit account of its spin being $(1/2)\hbar$. The equation famously had both positive- and negative-energy solutions. This potential disaster was converted by Dirac into a triumph by reinterpreting the (absence of) negative-energy solutions as (positive-energy) antiparticles – positrons, particles with positive charge but the same mass and spin as the electron. Positrons were eventually discovered by Anderson. It was later shown that the existence of antiparticles is a general feature of quantum field theory, not just a peculiarity of spin-1/2 particles. The significance of this discovery, however, is that the twin requirements of relativity and quantum theory are not compatible with a single-particle state; rather, these requirements result in a two-particle state. Thus, in some sense the requirements of relativity and quantum mechanics already start to take us down the road to a quantum theory of fields.

Quantum field theory is then constructed on the following sort of framework: "classical" theories for fields with any spin may be written down and these are quantized by reinterpreting the field variables as operators and imposing Heisenberg-type commutation relations on the field and its corresponding

"momentum" variable. So, for example, for spinless fields we have the equal-time commutation relation

$$[\phi(\boldsymbol{x},t),\pi(\boldsymbol{y},t)] = i\hbar\delta^{(3)}(\boldsymbol{x}-\boldsymbol{y})$$

where $\pi = \partial \mathrm{L}/\partial(\partial_0\phi)$ and L is the Lagrange density. The mass and spin of particles are defined with reference to the Poincaré group (thereby incorporating special relativity) and the quantum requirement is the familiar one that physical states are represented by vectors in Hilbert space. The rest follows: as Weinberg says, "quantum field theory is the way it is because (with certain qualifications) this is the only way to reconcile quantum mechanics with special relativity."

Renormalization

A notorious problem in quantum field theory is the occurrence of infinities. In QED, for example, the electron acquires a self-energy – and therefore a contribution to its mass – by virtue of the emission and reabsorption of virtual photons. It turns out that this self-energy is infinite – it is given by a divergent integral – even in the lowest order of perturbation theory. In the early days, this was recognized as being a serious problem, and in fact it turns out to be a generic problem in quantum field theory. It was realized by Dyson, however, that in some field theories these divergences may be dealt with by redefining a small number of parameters (e.g., in QED, the electron mass, charge, and field amplitude) so that thereafter the theory is finite to all orders of perturbation theory. Such theories are called renormalizable, and QED is a renormalizable field theory.

Some important field theories, however, are not renormalizable; an example is Fermi's theory of weak interactions. To lowest order in perturbation theory, Fermi's theory works well (e.g., in accounting for the electron spectrum in neutron beta decay), but to higher orders divergent results are obtained, which cannot be waved away by redefining a finite number of parameters; that is to say, as the order of perturbation increases, so also does the number of parameters to be redefined. Nonrenormalizable theories of this type have traditionally been regarded as highly undesirable, not to say rather nasty.

The modern view of renormalization is, however, somewhat different. The problem with nonrenormalizable theories is that, in order to calculate a physical process to all orders in perturbation theory, an infinite number of parameters must be renormalized, so the theory has no predictive power. In practice, however, we do not need to calculate to all orders in perturbation theory, since any physical process (say a scattering process or a particle decay) will only be observed at a finite energy and comparison of theory and experiment therefore only requires calculation up to a finite order of perturbation theory. So even nonrenormalizable theories are perfectly acceptable as low-energy theories. This amounts to a philosophy of effective field theories; an effective field theory is a model which holds good up to a particular energy scale, or equivalently down to a particular length scale.

An important addition to the theoretical armoury is the renormalization group. Renormalization is implemented first of all by a scheme of regularization, which enables the divergences to be exhibited explicitly. The simplest type of regularization is the introduction of a cutoff in the momentum integrals, but in modern particle physics the favored scheme is dimensional regularization. The dimensionality of the integrals in momentum space is taken to be $d = 4 - \varepsilon$ and the divergent quantities have an explicit dependence on ε (which, of course, as the "real" world is approached, approaches zero). At the same time, a mass parameter μ is introduced in order to define dimensionless quantities, for example, a dimensionless coupling constant. The renormalized quantities then depend on the "bare" (unrenormalized) quantities and on μ and ε. The arbitrariness of μ enables a differential equation, for scattering amplitudes, for example, to be written down. While at first sight this renormalization group equation might seem to have no physical importance, in fact it gives a powerful way of studying scattering behavior at large momenta.

Most interestingly, the concept of the renormalization group also arises in condensed matter physics. Here, rather than, for example, a cutoff in momentum space, the relevant parameter is a distance scale. In the Ising model in statistical mechanics, for example, in which spins are located on a lattice, the parameter is the lattice spacing. To construct a theory that describes the physics on the macroscopic scale involves integrating out the details on the microscopic scale and one way to do this is via the "block spin" transformation originally introduced by Kadanoff. In this way the renormalization group has had a large impact in condensed matter physics, for example, in the study of critical phenomena.

Particle Physics and Cosmology

Probably the most spectacular success of quantum field theory in the twentieth century has been in particle physics. The "standard model" accounts for the strong, electromagnetic, and weak interactions

between elementary particles with outstanding success. The interactions are generalizations of Maxwell's electrodynamics, which is invariant under a symmetry group U(1) of gauge transformations. An enlargement of this group to $SU(2) \otimes U(1)$ accounts for the unified electroweak interaction (the unification resulting from the fact that the two U(1)'s above are not exactly the same; there is some on-diagonal mixing), and the strong interactions between quarks, which binds them into hadrons, are invariant under an SU(3) group of gauge transformations. The gauge fields are the photon γ, the W and Z bosons (both heavy; of the order of 100 times the proton mass), and the (massless) gluons mediating the force between quarks (quantum chromodynamics, QCD). An important feature of the standard model is spontaneous symmetry breaking, which is the mechanism by which the W and Z particles acquire a mass (but the photon does not, and neither do the gluons). This goes by the name of the Higgs mechanism.

The quantization of the standard model is most successfully carried out using the path-integral formalism, rather than canonical quantization, and the proof of the renormalizability of the model (of nonabelian gauge theories with spontaneous symmetry breaking) was given by 't Hooft. Details of these topics are now available in many textbooks.

Confidence that this is a realistic model of elementary particles – that is to say, of quarks and leptons – depends, of course, on particular experiments and their interpretation and an important milestone on this journey was Feynman's quark–parton model of deep inelastic electron–proton scattering. The interpretation of the data required a picture of an electron scattering from an individual quark in the proton, and this in turn required a negligible interaction between quarks; in other words, that at small distances (inside the proton) the quarks are (almost) free – despite the fact that at large distances they most certainly are not! The proof, by Gross, Politzer, and Wilczek, that nonabelian gauge are indeed asymptotically free (asymptotic in momentum space, that is) was therefore an important event in helping to establish the credibility of the standard model.

A characteristic contribution of quantum field theory to our view of the physical world is its picture of the vacuum, as being populated with virtual particle–antiparticle pairs. A consequence of this is the phenomenon of vacuum polarization – that the presence of an electric charge in free space polarizes these virtual pairs. This in turns leads to the phenomenon of screening in QED, and antiscreening in QCD, SU(3) having a more complicated structure than U(1). It also leads to a nonzero (in fact, quadratically divergent!) value for the energy of the vacuum. This is in effect the contribution of the zero-point energies of all the oscillators in the Fourier expansion of the scalar field operator. In any other interaction than gravity, this zero-point energy may be ignored, but in gravity it may be expected to have observable consequences, and indeed it turns out that it plays the same role as a cosmological constant Λ, and therefore acts as an agent of acceleration, rather than deceleration, of the universe.

A final topic worth noting is one whose existence would have been inconceivable in the early days of this subject. The nonlinearity of the (nonabelian) gauge field equations and the existence of a nontrivial group space allows new types of topologically nontrivial solutions to these equations: solitons, bounces, instantons, sphalerons, and so on. Effects such as fractional spin and nonconservation of fermion number also appear, and, on the cosmological scale, domain walls and cosmic strings. There is something here for theoretical physicists of many differing interests.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; BRST Quantization; Constrained Systems; Constructive Quantum Field Theory; Deformation Quantization; Electroweak Theory; Euclidean Field Theory; Exact Renormalization Group; Integrability and Quantum Field Theory; Nonperturbative and Topological Aspects of Gauge Theory; Perturbative Renormalization Theory and BRST; Quantum Chromodynamics; Quantum Electrodynamics and Its Precision Tests; Quantum Fields with Indefinite Metric: Non-Trivial Models; Quantum Fields with Topological Defects; Renormalization: General Theory; Standard Model of Particle Physics; Symmetries and Conservation Laws; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Topological Defects and Their Homotopy Classification; Topological Quantum Field Theory: Overview; Twistors.

Further Reading

Cao TY (1997) *Conceptual Developments of 20th Century Field Theories*. Cambridge: Cambridge University Press.

Davies P (ed.) (1989) *The New Physics*. Cambridge: Cambridge University Press.

Gross F (1993) *Relativistic Quantum Mechanics and Field Theory*. New York: Wiley.

Huang K (1998) *Quantum Field Theory*. New York: Wiley.

Itzykson C and Zuber J-B (1980) *Quantum Field Theory*. New York: McGraw-Hill.

Maggiore M (2005) *A Modern Introduction to Quantum Field Theory*. Oxford: Oxford University Press.

Rubakov V (2002) *Classical Theory of Gauge Fields*. Princeton: Princeton University Press.

Schweber SS (1994) *QED and the Men Who Made It*. Princeton: Princeton University Press.

Schwinger J (ed.) (1958) *Quantum Electrodynamics*. New York: Dover.

't Hooft G (1997) *In Search of the Ultimate Building Blocks*. Cambridge: Cambridge University Press.

Weinberg S (1995, 1996) *The Quantum Theory of Fields*, vol. 1 and 2. Cambridge: Cambridge University Press.

Wentzel G (1960) Quantum theory of fields (until 1947). In: Fierz M and Weisskopf VF (eds.) *Theoretical Physics in the Twentieth Century*, (Reprinted in Mehra J (1973) *The Physicist's Conception of Nature*. Dordrecht: Reidel.) New York: Interscience.

Zee A (2003) *Quantum Field Theory in a Nutshell*. Princeton: Princeton University Press.

Algebraic Approach to Quantum Field Theory

R Brunetti and K Fredenhagen, Universität Hamburg, Hamburg, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Quantum field theory may be understood as the incorporation of the principle of locality, which is at the basis of classical field theory, into quantum physics. There are, however, severe obstacles against a straightforward translation of concepts of classical field theory into quantum theory, among them the notorious divergences of quantum field theory and the intrinsic nonlocality of quantum physics. Therefore, the concept of locality is somewhat obscured in the formalism of quantum field theory as it is typically exposed in textbooks. Nonlocal concepts such as the vacuum, the notion of particles or the S-matrix play a fundamental role, and neither the

Notes:

relation to classical field theory nor the influence of background fields can be properly treated.

Algebraic quantum field theory (AQFT; synonymously, local quantum physics), on the contrary, aims at emphasizing the concept of locality at every instance. As the nonlocal features of quantum physics occur at the level of states ("entanglement"), not at the level of observables, it is better not to base the theory on the Hilbert space of states but on the algebra of observables. Subsystems of a given system then simply correspond to subalgebras of a given algebra. The locality concept is abstractly encoded in a notion of independence of subsystems; two subsystems are independent if the algebra of observables which they generate is isomorphic to the tensor product of the algebras of the subsystems.

Spacetime can then – in the spirit of Leibniz – be considered as an ordering device for systems. So, one associates with regions of spacetime the algebras of observables which can be measured in the pertinent region, with the condition that the algebras of subregions of a given region can be identified with subalgebras of the algebra of the region.

Problems arise if one aims at a generally covariant approach in the spirit of general relativity. Then, in order to avoid pitfalls like in the "hole problem," systems corresponding to isometric regions must be isomorphic. Since isomorphic regions may be embedded into different spacetimes, this amounts to a simultaneous treatment of all spacetimes of a suitable class. We will see that category theory furnishes such a description, where the objects are the systems and the morphisms the embeddings of a system as a subsystem of other systems.

States arise as secondary objects via Hilbert space representations, or directly as linear functionals on the algebras of observables which can be interpreted as expectation values and are, therefore, positive and normalized. It is crucial that inequivalent representations ("sectors") can occur, and the analysis of the structure of the sectors is one of the big successes of AQFT. One can also study the particle interpretation of certain states as well as (equilibrium and nonequilibrium) thermodynamical properties.

The mathematical methods in AQFT are mainly taken from the theory of operator algebras, a field of mathematics which developed in close contact to mathematical physics, in particular to AQFT. Unfortunately, the most important field theories, from the point of view of elementary particle physics, as quantum electrodynamics or the standard model could not yet be constructed beyond formal perturbation theory with the annoying consequence that it seemed that the concepts of AQFT could not be applied to them. However, it has recently been shown that formal perturbation theory can be reshaped in the spirit of AQFT such that the algebras of observables of these models can be constructed as algebras of formal power series of Hilbert space operators. The price to pay is that the deep mathematics of operator algebras cannot be applied, but the crucial features of the algebraic approach can be used.

AQFT was originally proposed by Haag as a concept by which scattering of particles can be understood as a consequence of the principle of locality. It was then put into a mathematically precise form by Araki, Haag, and Kastler. After the analysis of particle scattering by Haag and Ruelle and the clarification of the relation to the Lehmann–Symanzik–Zimmermann (LSZ) formalism by Hepp, the structure of superselection sectors was studied first by Borchers and then in a fundamental series of papers by Doplicher, Haag, and Roberts (DHR) (see, e.g., Doplicher *et al.* (1971, 1974)) (soon after Buchholz and Fredenhagen established the relation to particles), and finally Doplicher and Roberts uncovered the structure of superselection sectors as the dual of a compact group thereby generalizing the Tannaka–Krein theorem of characterization of group duals.

With the advent of two-dimensional conformal field theory, new models were constructed and it was shown that the DHR analysis can be generalized to these models. Directly related to conformal theories is the algebraic approach to holography in anti-de Sitter (AdS) spacetime by Rehren.

The general framework of AQFT may be described as a covariant functor between two categories. The first one contains the information on local relations and is crucial for the interpretation. Its objects are topological spaces with additional structures (typically globally hyperbolic Lorentzian spaces, possibly spin bundles with connections, etc.), its morphisms being the structure-preserving embeddings. In the case of globally hyperbolic Lorentzian spacetimes, one requires that the embeddings are isometric and preserve the causal structure. The second category describes the algebraic structure of observables. In quantum physics the standard assumption is that one deals with the category of C^*-algebras where the morphisms are unital embeddings. In classical physics, one looks instead at Poisson algebras, and in perturbative quantum field theory one admits algebras which possess nontrivial representations as formal power series of Hilbert space operators. It is the leading principle of AQFT that the functor $\mathscr{A}$ contains all physical information. In particular, two theories are equivalent if the corresponding functors are naturally equivalent.

In the analysis of the functor $\mathscr{A}$, a crucial role is played by natural transformations from other functors on the locality category. For instance, a field A may be defined as a natural transformation from the category of test function spaces to the category of observable algebras via their functors related to the locality category.

Quantum Field Theories as Covariant Functors

The rigorous implementation of the generally covariant locality principle uses the language of category theory.

The following two categories are used:

Loc: The class of objects obj(**Loc**) is formed by all (smooth) d-dimensional ($d \geq 2$ is held fixed), globally hyperbolic Lorentzian spacetimes M which are oriented and time oriented. Given any two such objects M_1 and M_2, the morphisms $\psi \in \hom_{\mathbf{Loc}}(M_1, M_2)$ are taken to be the isometric embeddings $\psi : M_1 \to M_2$ of M_1 into M_2 but with the following constraints:

(i) if $\gamma : [a, b] \to M_2$ is any causal curve and $\gamma(a), \gamma(b) \in \psi(M_1)$ then the whole curve must be in the image $\psi(M_1)$, that is, $\gamma(t) \in \psi(M_1)$ for all $t \in [a, b]$;

(ii) any morphism preserves orientation and time orientation of the embedded spacetime. The composition is defined as the composition of maps, the unit element in $\hom_{\mathbf{Loc}}(M, M)$ is given by the identical embedding $\mathrm{id}_M : M \mapsto M$ for any $M \in \mathrm{obj}(\mathbf{Loc})$.

Obs: The class of objects obj(**Obs**) is formed by all C^*-algebras possessing unit elements, and the morphisms are faithful (injective) unit-preserving $*$-homomorphisms. The composition is again defined as the composition of maps, the unit element in $\hom_{\mathbf{Obs}}(\mathcal{A}, \mathcal{A})$ is for any $\mathcal{A} \in \mathrm{obj}(\mathbf{Obs})$ given by the identical map $\mathrm{id}_{\mathcal{A}} : A \mapsto A, A \in \mathcal{A}$.

The categories are chosen for definitiveness. One may envisage changes according to particular needs, as, for instance, in perturbation theory where instead of C^*-algebras general topological $*$-algebras are better suited. Or one may use von Neumann algebras, in case particular states are selected. On the other hand, one might consider for **Loc** bundles over spacetimes, or (in conformally invariant theories) admit conformal embeddings as morphisms. In case one is interested in spacetimes which are not globally hyperbolic, one could look at the globally hyperbolic subregions (where one needs to be careful about the causal convexity condition (i) above).

The concept of locally covariant quantum field theory is defined as follows.

Definition 1

(i) A locally covariant quantum field theory is a covariant functor $\mathscr{A}$ from **Loc** to **Obs** and (writing α_ψ for $\mathscr{A}(\psi)$) with the covariance properties

$$\alpha_{\psi'} \circ \alpha_\psi = \alpha_{\psi' \circ \psi}, \qquad \alpha_{\mathrm{id}_M} = \mathrm{id}_{\mathscr{A}(M)}$$

for all morphisms $\psi \in \hom_{\mathbf{Loc}}(M_1, M_2)$, all morphisms $\psi' \in \hom_{\mathbf{Loc}}(M_2, M_3)$, and all $M \in \mathrm{obj}(\mathbf{Loc})$.

(ii) A locally covariant quantum field theory described by a covariant functor $\mathscr{A}$ is called "causal" if the following holds: whenever there are morphisms $\psi_j \in \hom_{\mathbf{Loc}}(M_j, M), j = 1, 2$, so that the sets $\psi_1(M_1)$ and $\psi_2(M_2)$ are causally separated in M, then one has

$$\left[\alpha_{\psi_1}(\mathscr{A}(M_1)), \alpha_{\psi_2}(\mathscr{A}(M_2))\right] = \{0\}$$

where the element-wise commutation makes sense in $\mathscr{A}(M)$.

(iii) One says that a locally covariant quantum field theory given by the functor $\mathscr{A}$ obeys the "time-slice axiom" if

$$\alpha_\psi(\mathscr{A}(M)) = \mathscr{A}(M')$$

holds for all $\psi \in \hom_{\mathbf{Loc}}(M, M')$ such that $\psi(M)$ contains a Cauchy surface for M'.

Thus, a quantum field theory is an assignment of C^*-algebras to (all) globally hyperbolic spacetimes so that the algebras are identifiable when the spacetimes are isometric, in the indicated way. This is a precise description of the generally covariant locality principle.

The Traditional Approach

The traditional framework of AQFT, in the Araki–Haag–Kastler sense, on a fixed globally hyperbolic spacetime can be recovered from a locally covariant quantum field theory, that is, from a covariant functor $\mathscr{A}$ with the properties listed above.

Indeed, let M be an object in obj(**Loc**). $\mathcal{K}(M)$ denotes the set of all open subsets in M which are relatively compact and also contain, with each pair of points x and y, all g-causal curves in M connecting x and y (cf. condition (i) in the definition of **Loc**). $O \in \mathcal{K}(M)$, endowed with the metric of M restricted to O and with the induced orientation and time orientation, is a member of obj(**Loc**), and the injection map $\iota_{M,O} : O \to M$, that is, the identical map restricted to O, is an element in $\hom_{\mathbf{Loc}}(O, M)$.

With this notation, it is easy to prove the following assertion:

Theorem 1 *Let $\mathscr{A}$ be a covariant functor with the above-stated properties, and define a map $\mathcal{K}(M) \ni O \mapsto \mathcal{A}(O) \subset \mathscr{A}(M)$ by setting*

$$\mathcal{A}(O) := \alpha_{\iota_{M,O}}(\mathscr{A}(O))$$

Then the following statements hold:

(i) *The map fulfills isotony, that is,*

$$O_1 \subset O_2 \Rightarrow \mathcal{A}(O_1) \subset \mathcal{A}(O_2)$$
$$\text{for all } O_1, O_2 \in \mathcal{K}(M)$$

(ii) *If there exists a group G of isometric diffeomorphisms $\kappa : M \to M$ (so that $\kappa * g = g$) preserving orientation and time orientation, then there is a representation $G \ni \kappa \mapsto \tilde{\alpha}_\kappa$ of G by C^*-algebra automorphisms $\tilde{\alpha}_\kappa : \mathscr{A}(M) \to \mathscr{A}(M)$ such that*

$$\tilde{\alpha}_\kappa(\mathcal{A}(O)) = \mathcal{A}(\kappa(O)), \quad O \in \mathcal{K}(M)$$

(iii) *If the theory given by $\mathscr{A}$ is additionally causal, then it holds that*

$$[\mathcal{A}(O_1), \mathcal{A}(O_2)] = \{0\}$$
for all $O_1, O_2 \in \mathcal{K}(M)$ with O_1 causally separated from O_2.

These properties are just the basic assumptions of the Araki–Haag–Kastler framework.

The Achievements of the Traditional Approach

In the Araki–Haag–Kastler approach in Minkowski spacetime $\mathbb{M}$, many results have been obtained in the last 40 years, some of them also becoming a source of inspiration to mathematics. A description of the achievements can be organized in terms of a length-scale basis, from the small to the large. We assume in this section that the algebra $\mathscr{A}(\mathbb{M})$ is faithfully and irreducibly represented on a Hilbert space $\mathcal{H}$, that the Poincaré transformations are unitarily implemented with positive energy, and that the subspace of Poincaré invariant vectors is one dimensional (uniqueness of the vacuum). Moreover, algebras correponding to regions which are spacelike to a nonempty open region are assumed to be weakly closed (i.e., von Neumann algebras on $\mathcal{H}$), and the condition of weak additivity is fulfilled, that is, for all $O \in \mathcal{K}(\mathbb{M})$ the algebra generated from the algebras $\mathcal{A}(O + x), x \in \mathbb{M}$ is weakly dense in $\mathscr{A}(\mathbb{M})$.

Ultraviolet Structure and Idealized Localizations

This section deals with the problem of inspecting the theory at very small scales. In the limiting case, one is interested in idealized localizations, eventually the points of spacetimes. But the observable algebras are trivial at any point $x \in \mathbb{M}$, namely

$$\bigcap_{O \ni x} \mathcal{A}(O) = \mathbb{C}1, \quad O \in \mathcal{K}(\mathbb{M})$$

Hence, pointlike localized observables are necessarily singular. Actually, the Wightman formulation of quantum field theory is based on the use of distributions on spacetime with values in the algebra of observables (as a topological $*$-algebra). In spite of technical complications whose physical significance is unclear, this formalism is well suited for a discussion of the connection with the Euclidean theory, which allows, in fortunate cases, a treatment by path integrals; it is more directly related to models and admits, via the operator-product expansion, a study of the short-distance behavior. It is, therefore, an important question how the algebraic approach is related to the Wightman formalism. The reader is referred to the literature for exploring the results on this relation.

Whereas these results point to an essential equivalence of both formalisms, one needs in addition a criterion for the existence of sufficiently many Wightman fields associated with a given local net. Such a criterion can be given in terms of a compactness condition to be discussed in the next subsection. As a benefit, one derives an operator-product expansion which has to be assumed in the Wightman approach.

In the purely algebraic approach, the ultraviolet structure has been investigated by Buchholz and Verch. Small-scale properties of theories are studied with the help of the so-called scaling algebras whose elements can be described as orbits of observables under all possible renormalization group motions. There results a classification of theories in the scaling limit which can be grouped into three broad classes: theories for which the scaling limit is purely classical (commutative algebras), those for which the limit is essentially unique (stable ultraviolet fixed point) and not classical, and those for which this is not the case (unstable ultraviolet fixed point). This classification does not rely on perturbation expansions. It allows an intrinsic definition of confinement in terms of the so-called ultraparticles, that is, particles which are visible only in the scaling limit.

Phase-Space Analysis

As far as finite distances are concerned, there are two apparently competing principles, those of

nuclearity and modularity. The first one suggests that locally, after a cutoff in energy, one has a situation similar to that of old quantum mechanics, namely a finite number of states in a finite volume of phase space. Aiming at a precise formulation, Haag and Swieca introduced their notion of compactness, which Buchholz and Wichmann sharpened into that of nuclearity. The latter authors proposed that the set generated from the vacuum vector Ω,

$$\{e^{-\beta H}A\Omega \mid A \in \mathcal{A}(O), \|A\| < 1\}$$

H denoting the generator of time translations (Hamiltonian), is nuclear for any $\beta > 0$, roughly stating that it is contained in the image of the unit ball under a trace class operator. The nuclear size $Z(\beta,O)$ of the set plays the role of the partition function of the model and has to satisfy certain bounds in the parameter β. The consequence of this constraint is the existence of product states, namely those normal states for which observables localized in two given spacelike separated regions are uncorrelated. A further consequence is the existence of thermal equilibrium states (KMS states) for all $\beta > 0$.

The second principle concerns the fact that, even locally, quantum field theory has infinitely many degrees of freedom. This becomes visible in the Reeh–Schlieder theorem, which states that every vector Φ which is in the range of $e^{-\beta H}$ for some $\beta > 0$ (in particular, the vacuum Ω) is cyclic and separating for the algebras $\mathcal{A}(O), O \in \mathcal{K}(\mathbb{M})$, that is, $\mathcal{A}(O)\Phi$ is dense in $\mathcal{H}$ (Φ is cyclic) and $A\Phi = 0, A \in \mathcal{A}(O)$ implies $A = 0$ (Φ is separating). The pair $(\mathcal{A}(O),\Omega)$ is then a von Neumann algebra in the so-called standard form. On such a pair, the Tomita–Takesaki theory can be applied, namely the densely defined operator

$$SA\Omega = A^*\Omega, \quad A \in \mathcal{A}(O)$$

is closable, and the polar decomposition of its closure $\bar{S} = J\Delta^{1/2}$ delivers an antiunitary involution J (the modular conjugation) and a positive self-adjoint operator Δ (the modular operator) associated with the standard pair $(\mathcal{A}(O),\Omega)$. These operators have the properties

$$J\mathcal{A}(O)J = \mathcal{A}(O)'$$

where the prime denotes the commutant, and

$$\Delta^{it}\mathcal{A}(O)\Delta^{-it} = \mathcal{A}(O), \quad t \in \mathbb{R}$$

The importance of this structure is based on the fact disclosed by Bisognano and Wichmann using Poincaré-covariant Wightman fields and local algebras generated by them, that for specific regions in Minkowski spacetime the modular operators have a geometrical meaning. Indeed, these authors showed for the pair $(\mathcal{A}(W),\Omega)$, where W denotes the wedge region $W = \{x \in \mathbb{M} \mid |x^0| < x^1\}$, that the associated modular unitary Δ^{it} is the Lorentz boost with velocity $\tanh(2\pi t)$ in the direction 1 and that the modular conjugation J is the CP_1T symmetry operator with parity P_1 the reflection with respect to the $x^1 = 0$ plane. Later, Borchers discovered that already on the purely algebraic level a corresponding structure exists. He proved that, given any standard pair $(\mathcal{A},\Phi)$ and a one-parameter group of unitaries $\tau \to U(\tau)$ acting on the Hilbert space $\mathcal{H}$ with a positive generator and such that Φ is invariant and $U(\tau)\mathcal{A}U(\tau)^* \subset \mathcal{A}, \tau > 0$, then the associated modular operators Δ and J fulfill the commutation relations

$$\Delta^{it}U(\tau)\Delta^{-it} = U(e^{-2\pi t}\tau)$$
$$JU(\tau)J = U(-\tau)$$

which are just the commutation relations between boosts and lightlike translations.

Surprisingly, there is a direct connection between the two concepts of nuclearity and modularity. Indeed, in the nuclearity condition, it is possible to replace the Hamiltonian operator by a specific function of the modular operator associated with a slightly larger region. Furthermore, under mild conditions, nuclearity and modularity together determine the structure of local algebras completely; they are isomorphic to the unique hyperfinite type III_1 von Neumann algebra.

Sectors, Symmetries, Statistics, and Particles

Large scales are appropriate for discussing global issues like superselection sectors, statistics and symmetries as far as large spacelike distances are concerned, and scattering theory, with the resulting notions of particles and infraparticles, as far as large timelike distances are concerned.

In purely massive theories, where the vacuum sector has a mass gap and the mass shell of the particles are isolated, a very satisfactory description of the multiparticle structure at large times can be given. Using the concept of almost local particle generators,

$$\Psi = A(t)\Omega$$

where Ψ is a single-particle state (i.e., an eigenstate of the mass operator), $A(t)$ is a family of almost local operators essentially localized in the kinematical region accessible from a given point by a motion with the velocities contained in the spectrum of Ψ, one obtains the multiparticle states as limits of products $A_1(t)\cdots A_n(t)\Omega$ for disjoint velocity supports. The corresponding closed subspaces are

invariant under Poincaré transformations and are unitarily equivalent to the Fock spaces of noninteracting particles.

For massless particles, no almost-local particle generators can be expected to exist. In even dimensions, however, one can exploit Huygens principle to construct asymptotic particle generators which are in the commutant of the algebra of the forward or backward lightcone, respectively. Again, their products can be determined and multiparticle states obtained.

Much less well understood is the case of massive particles in a theory which also possesses massless particles. Here, in general, the corresponding states are not eigenstates of the mass operator. Since quantum electrodynamics (QED) as well as the standard model of elementary particles have this problem, the correct treatment of scattering in these models is still under discussion. One attempt to a correct treatment is based on the concept of the so-called particle weights, that is, unbounded positive functionals on a suitable algebra. This algebra is generated by positive almost-local operators annihilating the vacuum and interpreted as counters.

The structure at large spacelike scales may be analyzed by the theory of superselection sectors. The best-understood case is that of locally generated sectors which are the objects of the DHR theory. Starting from a distinguished representation π_0 (vacuum representation) which is assumed to fulfill the Haag duality,

$$\pi_0(\mathcal{A}(O)) = \pi_0(\mathcal{A}(O'))'$$

for all double cones O, one may look at all representations which are equivalent to the vacuum representation if restricted to the observables localized in double cones in the spacelike complement of a given double cone. Such representations give rise to endomorphisms of the algebra of observables, and the product of endomorphisms can be interpreted as a product of sectors ("fusion"). In general, these representations violate the Haag duality, but there is a subclass of the so-called finite statistics sectors where the violation of Haag duality is small, in the sense that the nontrivial inclusion

$$\pi(\mathcal{A}(O)) \subset \pi(\mathcal{A}(O'))'$$

has a finite Jones index. These sectors form (in at least three spacetime dimensions) a symmetric tensor category with some further properties which can be identified, in a generalization of the Tannaka–Krein theorem, as the dual of a unique compact group. This group plays the role of a global gauge group. The symmetry of the category is expressed in terms of a representation of the symmetric group. One may then enlarge the algebra of observables and obtain an algebra of operators which transform covariantly under the global gauge group and satisfy Bose or Fermi commutation relations for spacelike separation.

In two spacetime dimensions, one obtains instead braided tensor categories. They have been classified under additional conditions (conformal symmetry, central charge $c < 1$) in a remarkable work by Kawahigashi and Longo. Moreover, in their paper, one finds that by using completely new methods (Q-systems) a new model is unveiled, apparently inaccessible by methods used by others. To some extent, these categories can be interpreted as duals of generalized quantum groups.

The question arises whether all representations describing elementary particles are, in the massive case, DHR representations. One can show that in the case of a representation with an isolated mass shell there is an associated vacuum representation which becomes equivalent to the particle representation after restriction to observables localized spacelike to a given infinitely extended spacelike cone. This property is weaker than the DHR condition but allows, in four spacetime dimensions, the same construction of a global gauge group and of covariant fields with Bose and Fermi commutation relations, respectively, as the DHR condition. In three space dimensions, however, one finds a braided tensor category, which has similar properties as those known from topological field theories in three dimensions.

The sector structure in massless theories is not well understood, due to the infrared problem. This is in particular true for QED.

Fields as Natural Transformations

In order to be able to interpret the theory in terms of measurements, one has to be able to compare observables associated with different regions of spacetime, or, even different spacetimes. In the absence of nontrivial isometries, such a comparison can be made in terms of locally covariant fields. By definition, these are natural transformations from the functor of quantum field theory to another functor on the category of spacetimes **Loc**.

The standard case is the functor which associates with every spacetime M its space $\mathcal{D}(M)$ of smooth compactly supported test functions. There, the morphisms are the pushforwards $\mathcal{D}\psi \equiv \psi_*$.

Definition 2 A locally covariant quantum field Φ is a natural transformation between the functors $\mathscr{D}$ and $\mathscr{A}$, that is, for any object M in obj(**Loc**) there exists a morphism $\Phi_M : \mathcal{D}(M) \to \mathscr{A}(M)$ such that for

any pair of objects M_1 and M_2 and any morphism ψ between them, the following diagram commutes:

$$\begin{array}{ccc} \mathcal{D}(M_1) & \xrightarrow{\Phi_{M_1}} & \mathcal{A}(M_1) \\ \psi_* \downarrow & & \downarrow \alpha_\psi \\ \mathcal{D}(M_2) & \xrightarrow[\Phi_{M_2}]{} & \mathcal{A}(M_2) \end{array}$$

The commutativity of the diagram means, explicitly, that

$$\alpha_\psi \circ \Phi_{M_1} = \Phi_{M_2} \circ \psi_*$$

which is the requirement sought for the covariance of fields. It contains, in particular, the standard covariance condition for spacetime isometries.

Fields in the above sense are not necessarily linear. Examples for fields which are also linear are the scalar massive free Klein–Gordon fields on all globally hyperbolic spacetimes and its locally covariant Wick polynomials. In particular, the energy–momentum tensors can be constructed as locally covariant fields, and they provide a crucial tool for discussing the back-reaction problem for matter fields.

An example for the more general notion of a field are the local S-matrices in the Stückelberg–Bogolubov–Epstein–Glaser sense. These are unitaries $S_M(\lambda)$ with $M \in \mathrm{obj}(\mathbf{Loc})$ and $\lambda \in \mathcal{D}(M)$ which satisfy the conditions

$$S_M(0) = 1$$

$$S_M(\lambda + \mu + \nu) = S_M(\lambda + \mu) S_M(\mu)^{-1} S_M(\mu + \nu)$$

for $\lambda, \mu, \nu \in \mathcal{D}(M)$ such that the supports of λ and ν can be separated by a Cauchy surface of M with $\mathrm{supp}\,\lambda$ in the future of the surface.

The importance of these S-matrices relies on the fact that they can be used to define a new quantum field theory. The new theory is locally covariant if the original theory is and if the local S-matrices satisfy the condition of the locally covariant field above. A perturbative construction of interacting quantum field theory on globally hyperbolic spacetimes was completed in this way by Hollands and Wald, based on previous work by Brunetti and Fredenhagen.

See also: Axiomatic Quantum Field Theory; Constructive Quantum Field Theory; Current Algebra; Deformation Quantization and Representation Theory; Dispersion Relations; Indefinite Metric; Integrability and Quantum Field Theory; Operads; Perturbative Renormalization Theory and BRST; Quantum Central Limit Theorems; Quantum Field Theory: A Brief Introduction; Quantum Field Theory in Curved Spacetime; Quantum Fields with Indefinite Metric: Non-Trivial Models; Quantum Fields with Topological Defects; Quantum Geometry and its Applications; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Scattering in Relativistic Quantum Field Theory: The Analytic Program; Spin Foams; Symmetries in Quantum Field Theory: Algebraic Aspects; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Tomita–Takesaki Modular Theory; Two-Dimensional Models; von Neumann Algebras: Introduction, Modular Theory and Classification Theory; von Neumann Algebras: Subfactor Theory.

Further Reading

Araki H (1999) *Mathematical Theory of Quantum Fields*. Oxford: Oxford University Press.

Baumgærtel H and Wollemberg M (1992) *Causal Nets of Operator Algebras*. Berlin: Akademie Verlag.

Borchers HJ (1996) *Translation Group and Particle Representation in Quantum Field Theory*, Lecture Notes in Physics. New Series m: Monographs, 40. Berlin: Springer.

Borchers HJ (2000) On revolutionizing quantum field theory with Tomita's modular theory. *Journal of Mathematical Physics* 41: 3604–3673.

Bratteli O and Robinson DW (1987) *Operator Algebras and Quantum Statistical Mechanics*, vol. 1 Berlin: Springer.

Brunetti R and Fredenhagen K (2000) Microlocal analysis and interacting quantum field theories: renormalization on physical backgrounds. *Communications in Mathematical Physics* 208: 623–661.

Brunetti R, Fredenhagen K, and Verch R (2003) The generally covariant locality principle – a new paradigm for local quantum field theory. *Communications in Mathematical Physics* 237: 31–68.

Buchholz D and Haag R (2000) The quest for understanding in relativistic quantum physics. *Journal of Mathematical Physics* 41: 3674–3697.

Dixmier J (1964) *Les C*-algèbres et leurs représentations*. Paris: Gauthier-Villars.

Doplicher S, Haag R, and Roberts JE (1971) Local observables and particle statistics I. *Communications in Mathematical Physics* 23: 199–230.

Doplicher S, Haag R, and Roberts JE (1974) Local observables and particle statistics II. *Communications in Mathematical Physics* 35: 49–85.

Evans DE and Kawahigashi Y (1998) *Quantum Symmetries on Operator Algebras*. New York: Clarendon Press.

Haag R (1996) *Local Quantum Physics*, 2nd edn. Berlin: Springer.

Haag R and Kastler D (1964) An algebraic approach to quantum field theory. *Journal of Mathematical Physics* 5: 848–861.

Hollands S and Wald RM (2001) Local Wick polynomials and time ordered products of quantum fields in curved spacetime. *Communications in Mathematical Physics* 223: 289–326.

Hollands S and Wald RM (2002) Existence of local covariant time ordered products of quantum field in curved spacetime. *Communications in Mathematical Physics* 231: 309–345.

Kastler D (ed.) (1990) *The Algebraic Theory of Superselection Sectors. Introductions and Recent Results*. Singapore: World Scientific.

Kawahigashi Y and Longo R (2004) Classification of local conformal nets. Case $c < 1$. *Annals of Mathematics* 160: 1–30.

Takesaki M (2003) *Theory of Operator Algebras I, II, III*, Encyclopedia of Mathematical Sciences, vols. 124, 125, 127. Berlin: Springer.

Wald RM (1994) *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics*. Chicago: University of Chicago Press.

Axiomatic Quantum Field Theory

B Kuckert, Universität Hamburg, Hamburg, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The term "axiomatic quantum field theory" subsumes a collection of research branches of quantum field theory analyzing the general principles of relativistic quantum physics. The content of the results typically is structural and retrospective rather than quantitative and predictive.

The first axiomatic activities in quantum field theory date back to the 1950s, when several groups started investigating the notion of scattering and S-matrix in detail (Lehmann, Symanzik, and Zimmermann 1955 (LSZ-approach), Bogoliubov and Parasiuk 1957, Hepp and Zimmermann (BPHZ-approach), Haag 1957–59 and Ruelle 1962 (Haag–Ruelle theory) (*see* Scattering, Asymptotic Completeness and Bound States and Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools).

Wightman (1956) analyzed the properties of the vacuum expectation values used in these approaches and formulated a system of axioms that the vacuum expectation values ought to satisfy in general. Together with Gårding (1965), he later formulated a system of axioms in order to characterize general quantum fields in terms of operator-valued functionals, and the two systems have been found to be equivalent.

A couple of spectacular theorems such as the PCT theorem and the spin–statistics theorem have been obtained in this setting, but no interacting quantum fields satisfying the axioms have been found so far (in $1+3$ spacetime dimensions). So, the development of alternatives and modifications of the setting got into the focus of the theory, and the axioms themselves became the objects of research. Their role as axioms – understood in the common sense – turned into the role of mere properties of quantum fields. Today, the term "axiomatic quantum field theory" is widely avoided for this reason.

In a long list of publications spread over the 1960s, Araki, Borchers, Haag, Kastler, and others worked out an algebraic approach to quantum field theory in the spirit of Segal's "postulates for general quantum Mechanics" (1947) (*see* Algebraic Approach to Quantum Field Theory).

The Wightman setting was the basis of a framework into which the causal construction of the S-matrix developed by Stückelberg (1951) and Bogoliubov and Shirkov (1959) has been fitted by Epstein and Glaser (1973). The causality principle fixes the time-ordered products up to a finite number of parameters at each order, which are to be put in as the renormalization constants.

Already in 1949, Dyson had seen that problems in the formulation of quantum electrodynamics (QED) could be avoided by "just" multiplying the time variable and, correspondingly, the energy variable by the imaginary unit constant ("Wick rotation"). Schwinger then investigated time-ordered Green functions of QED in this Euclidean setting. This approach was formulated in terms of axioms by Osterwalder and Schrader (1973, 1975) (*see* Euclidean Field Theory).

Other extensions of the aforementioned settings are objects of current research (*see* Indefinite Metric,

Notes:

Quantum Field Theory in Curved Spacetime, Symmetries in Quantum Field Theory of Lower Spacetime Dimensions, and Thermal Quantum Field Theory).

Quantum Fields

Gårding and Wightman characterized operator-valued quantum fields on the Minkowski spacetime $\mathbb{R}^{1+3}$ by a couple of axioms. Given additional assumptions concerning the high-energy behavior, the Gårding–Wightman fields are in one–one correspondence with algebraic field theories.

Without specifying or presupposing these additional assumptions, the axioms will now be formulated and discussed in detail and compared to the corresponding conditions in the algebraic setting. Adjoint operators are marked by an asterisk, and Einstein's summation convention is used.

Operator-valued functionals *The components of a field F are an n-tuple $F_1 \cdots F_n$ of linear maps that assign to each test function $\varphi \in C_0^\infty(\mathbb{R}^{1+3})$ linear operators $F_1(\varphi) \cdots F_n(\varphi)$ in a Hilbert space $\mathcal{H}$ with domains of definition $D(F_1(\varphi)) \cdots D(F_n(\varphi))$. There exists a dense subspace $\mathcal{D}$ of $\mathcal{H}$ with $\mathcal{D} \subset D(F_\nu(\varphi)) \cap D(F_\nu(\varphi)^*)$ and $F_\nu(\varphi)\mathcal{D} \cup F_\nu(\varphi)^*\mathcal{D} \subset \mathcal{D}$ for all indices ν. Consider m such fields $\mathbf{F}^1 \cdots \mathbf{F}^m$ with components F_ν^a, $1 \le a \le m$, $1 \le \nu \le n_a$. Assume there to be an involution $* : (1 \cdots m) \to (1 \cdots m)$ such that $F_\nu^{a^*}(\varphi) = F_\nu^a(\overline{\varphi})^*$, where $\overline{\varphi}(x) := \overline{\varphi(x)}$.*

Quantum fields cannot be operator-valued functions on $\mathbb{R}^{1+3}$ if one wants them to exhibit (part of) the properties to follow. But point fields can be quadratic forms; typically this is the case for fields in a Fock space.

For each component F_ν^a and each open region $\mathcal{O} \subset \mathbb{R}^{1+3}$, the field operators $F_\nu^a(\varphi)$ with $\operatorname{supp}\varphi \subset \mathcal{O}$ generate a *-algebra $\mathcal{F}_\nu^a(\mathcal{O})$ of operators defined on $\mathcal{D}$. These operators typically are unbounded, which is one of the differences with the traditional setting of the algebraic approach. There a C*-algebra $\mathfrak{A}(\mathcal{O})$ is assigned to each open region $\mathcal{O}$ in such a way that $\mathcal{O} \subset \mathcal{P}$ implies $\mathfrak{A}(\mathcal{O}) \subset \mathfrak{A}(\mathcal{P})$. Each C*-algebra is a *-algebra, but in contrast to a C*-algebra, a*-algebra does not need to be endowed with a norm. The fundamental observables in quantum theory are bounded positive operators (typically, but not always, projections), and these generate a C*-algebra.

There is no fundamental physical motivation for confining the setting to fields with a finite number of components, except that it includes most of the fields known from "daily life."

Continuity as a distribution *For all $\Phi, \Psi \in \mathcal{D}$, the linear functionals $T_{\Phi,\Psi,\nu}$ on $C_0^\infty(\mathbb{R}^{1+3})$ defined by*

$$T^a_{\nu,\Phi,\Psi}(\varphi) := \langle \Phi, F^a_\nu(\varphi)\Psi \rangle$$

are distributions. They can be extended to tempered distributions.

The Fourier transform of a tempered distribution is well defined as a tempered distribution. It is mainly due to the importance of Fourier transformations that the preceding assumption is convenient. Bogoliubov *et al.* (1975) remark that the assumption is not a mere technicality, since it rules out nonrenormalizable quantum fields.

Microcausality (Bose–Fermi alternative) *If φ and ψ are test functions with spacelike separated support, then*

$$F^a_\nu(\varphi)F^b_\mu(\psi)|_{\mathcal{D}} = \pm F^b_\mu(\psi)F^a_\nu(\varphi)|_{\mathcal{D}}.$$

The sign depends on the statistics of the fields, it is "$-$" if and only if both F^a and F^b are fermion fields.

Microcausality is closely related to Einstein causality. Einstein causality requires that any two observables located in spacelike separated regions commute in the strong sense, that is, their spectral measures commute. But fields with Fermi–Dirac statistics are not observables, and not even for Bose–Einstein fields with self-adjoint field operators does the above condition imply that the spectral projections commute, which is the criterion for commensurability. The sign on the right-hand side does, however, specify the statistics of the field.

This is a crucial difference with the algebraic approach. If $\mathcal{O}$ and P are spacelike separated open regions and if $A \in \mathfrak{A}(\mathcal{O})$ and $B \in \mathfrak{A}(P)$, then one assumes, like in the above case, that $AB = BA$ (locality). But being elements of C*-algebras, A and B are bounded operators (or can be represented accordingly), so if A and B are self-adjoint, they are, indeed, commensurable.

Doplicher, Haag, and Roberts (1974) and Buchholz and Fredenhagen (1984) have derived from this input of observables a field structure of *localized particle states*, and they showed that the statistics of these fields is Bose–Einstein, Fermi–Dirac, or some corresponding parastatistics (which is, *a priori*, forbidden if one assumes microcausality).

Recall that the unimodular group $\mathrm{SL}(2,\mathbb{C})$ is isomorphic to the universal covering group of the restricted Lorentz group $L_+^\uparrow$ (the connected component containing the unit element). Denote by $\Lambda : \mathrm{SL}(2,\mathbb{C}) \to L_+^\uparrow$ a covering map.

Covariance *There exist strongly continuous unitary representations U and T of $\mathrm{SL}(2,\mathbb{C})$ and $(\mathbb{R}^{1+3},+)$, respectively, and representations $\boldsymbol{D}^1\cdots\boldsymbol{D}^m$ of $\mathrm{SL}(2,\mathbb{C})$ in $\mathbb{C}^{n_1}\cdots\mathbb{C}^{n_m}$, respectively, such that*

$$U(g)F^a_\nu(\varphi)U(g)^* = D^a(g^{-1})^\mu_\nu F^a_\mu(\varphi(\Lambda(g)^{-1}\cdot))$$

and

$$T(y)F^a_\nu(\varphi)T(y)^* = F^a_\mu(\varphi(\cdot - y)),$$

where $D^a(g^{-1})^\mu_\nu$ are the elements of the matrix $\boldsymbol{D}^a(g^{-1})$. Dropping coordinate indices, this reads

$$U(g)\boldsymbol{F}^a(\varphi)U(g)^* = \boldsymbol{D}^a(g^{-1})\boldsymbol{F}^a(\varphi(\Lambda(g)^{-1}\cdot))$$

and

$$T(y)\boldsymbol{F}^a(\varphi)T(y)^* = \boldsymbol{F}^a(\varphi(\cdot - y)).$$

The representations U and T generate a representation of the universal covering of the restricted Poincaré group.

As it stands, this assumption is a very strong one, since it manifestly fixes the action of the representation on the field operators. In the algebraic approach, the covariance assumption is more modestly formulated. Namely, it is assumed that $U(g)\mathfrak{A}(\mathcal{O})U(g)^* = \mathfrak{A}(\Lambda(g)\mathcal{O})$ and $T(y)\mathfrak{A}(\mathcal{O})T(y)^* = \mathfrak{A}(\mathcal{O}+y)$, leaving open how the representation acts on the single local observables.

Vacuum vector *There exists a unique (up to a multiple) vector $\Omega\in\mathcal{D}$ that is invariant under the representations U and T and cyclic with respect to the algebra $\mathcal{F}(\mathbb{R}^{1+3})$ generated by all field operators $F^a_\nu(\varphi)$, that is, $\overline{\mathcal{F}(\mathbb{R}^{1+3})\Omega} = \mathcal{H}$.*

Spectrum condition *The joint spectrum of the components of the 4-momentum, i.e., of the generators of the spacetime translations, has support in the closed forward light cone $\overline{V}_+$, that is, the set $\{k^2\geq 0, k_0\geq 0\}$.*

The existence of an invariant ground state called the vacuum is standard in algebraic quantum field theory as well.

N-Point Functions

Consider the above fields $\boldsymbol{F}^1\cdots\boldsymbol{F}^m$. For each $N\in\mathbb{N}$ and each N-tuple $(a_1\cdots a_N)$ of natural numbers $\leq m$ (labeling fields), define families $(\boldsymbol{F}^{a_1\cdots a_N}) := (F^{a_1\cdots a_N}_{\nu_1\cdots\nu_N})_{\nu_i\leq n_{a_i}}$ and $\boldsymbol{w}^{a_1\cdots a_N} := (w^{a_1\cdots a_N}_{\nu_1\cdots\nu_N})_{\nu_i\leq n_{a_i}}$ of distributions on $(\mathbb{R}^{1+3})^N$ by

$$F^{a_1\cdots a_N}_{\nu_1\cdots\nu_N}(\varphi_1\otimes\cdots\otimes\varphi_N) := F^{a_1}_{\nu_1}(\varphi_1)\cdots F^{a_N}_{\nu_N}(\varphi_N)$$

(using the nuclear theorem) and

$$w^{a_1\cdots a_N}_{\nu_1\cdots\nu_N}(\psi) := \langle\Omega, F^{a_1\cdots a_N}_{\nu_1\cdots\nu_N}(\psi)\Omega\rangle. \qquad [1]$$

These distributions are called the "N-point functions" of the fields $\boldsymbol{F}^1\cdots\boldsymbol{F}^m$ and yield the vacuum expectation values of the theory. It is straightforward to deduce the following properties from the Gårding–Wightman axioms.

Microcausality (Bose–Fermi alternative) *If φ_i and φ_{i+1} have spacelike separated supports, then*

$$w^{a_1\cdots a_i a_{i+1}\cdots a_N}_{\nu_1\cdots\nu_i\nu_{i+1}\cdots\nu_N}(\varphi_1\otimes\cdots\otimes\varphi_i\otimes\varphi_{i+1}\otimes\cdots\otimes\varphi_N)$$
$$= \pm w^{a_1\cdots a_{i+1}a_i\cdots a_N}_{\nu_1\cdots\nu_{i+1}\nu_i\cdots\nu_N}(\varphi_1\otimes\cdots\otimes\varphi_{i+1}\otimes\varphi_i\otimes\cdots\otimes\varphi_N).$$

or dropping coordinate indices,

$$\boldsymbol{w}^{a_1\cdots a_i a_{i+1}\cdots a_N}(\varphi_1\otimes\cdots\otimes\varphi_i\otimes\varphi_{i+1}\otimes\cdots\otimes\varphi_N)$$
$$= \pm\boldsymbol{w}^{a_1\cdots a_{i+1}a_i\cdots a_N}(\varphi_1\otimes\cdots\otimes\varphi_{i+1}\otimes\varphi_i\otimes\cdots\otimes\varphi_N).$$

Invariance *For all $g\in\mathrm{SL}(2,\mathbb{C})$ and $y\in\mathbb{R}^{1+3}$, one has*

$$\begin{aligned} &w^{a_1\cdots a_N}_{\mu_1\cdots\mu_N}(\varphi_1\otimes\cdots\otimes\varphi_N)\\ &\quad= D^{a_1}(g^{-1})^{\nu_1}_{\mu_1}\cdots D^{a_N}(g^{-1})^{\nu_N}_{\mu_N}\\ &\qquad\times w^{a_1\cdots a_N}_{\nu_1\cdots\nu_N}(\Lambda(g)\varphi_1\otimes\cdots\otimes\Lambda(g)\varphi_N)\\ &\quad= w^{a_1\cdots a_N}_{\mu_1\cdots\mu_N}(\varphi_1(\cdot - y)\otimes\cdots\otimes\varphi_N(\cdot - y)) \end{aligned}$$

or dropping coordinate indices,

$$\begin{aligned} &\boldsymbol{w}^{a_1\cdots a_N}(\varphi_1\otimes\cdots\otimes\varphi_N)\\ &\quad= (\boldsymbol{D}^{a_1}(g^{-1})\otimes\cdots\otimes\boldsymbol{D}^{a_N}(g^{-1}))\\ &\qquad\times\boldsymbol{w}^{a_1\cdots a_N}(\Lambda(g)\varphi_1\otimes\cdots\otimes\Lambda(g)\varphi_N)\\ &\quad= \boldsymbol{w}^{a_1\cdots a_N}(\varphi_1(\cdot - y)\otimes\cdots\otimes\varphi_N(\cdot - y)). \end{aligned}$$

By translation invariance, the N-point functions $w^{a_1\cdots a_N}_{\nu_1\cdots\nu_N}(x_1\cdots x_N)$ only depend on the $N-1$ *relative*-position vectors $\xi_1 := x_1 - x_2$, $\xi_2 := x_2 - x_3,\ldots,$ $\xi_{N-1} := x_{N_1} - x_N$. This means that there are distributions $W^{a_1\cdots\nu_N}_{\nu_1\cdots\nu_N}$ on $(\mathbb{R}^{1+3})^{N-1}$ related to the N-point functions by the symbolic condition

$$w^{a_1\cdots a_N}_{\nu_1\cdots\nu_N}(x_1\cdots x_N) = W^{a_1\cdots a_N}_{\nu_1\cdots\nu_N}(\xi_1\cdots\xi_{N-1}).$$

In precise notation, this reads

$$w^{a_1\cdots a_N}_{\nu_1\cdots\nu_N}(\varphi) = \int_{1+3} W^{a_1\cdots a_N}_{\nu_1\cdots\nu_N}(\varphi_x)\,dx,$$

where

$$\varphi_x(\xi_1\cdots\xi_{N-1}) := \varphi(x, x-\xi_1, x-\xi_1-\xi_2,\ldots,x-\xi_1-\cdots-\xi_{N-1}).$$

The functions $W^{a_1\cdots a_N}_{\nu_1\cdots\nu_N}$ are called the *Wightman functions*, and they have the following property because of the spectrum condition of the field.

Spectrum condition *The support of the Fourier transform of each* $W^{a_1 \cdots a_N}_{\nu_1 \cdots \nu_N}$ *is contained in* $(\overline{V}_+)^{N-1}$.

The uniqueness of the vacuum vector (up to a phase) is equivalent to the following condition.

Cluster property *For* $N \geq 2$, *let* x *be a spacelike vector in* $\mathbb{R}^{1+3}$, *let* L *be a natural number* $< N$, *and let* φ *and* ψ *be tempered test functions on* $(\mathbb{R}^{1+3})^L$ *and* $(\mathbb{R}^{1+3})^{N-L}$, *respectively. then*

$$\lim_{0<\lambda\to\infty} w^{a_1 \cdots a_N}_{\nu_1 \cdots \nu_N}(\varphi \otimes \psi(\cdot - \lambda x)) = w^{a_1 \cdots a_L}_{\nu_1 \cdots \nu_L}(\varphi) w^{a_{L+1} \cdots a_N}_{\nu_{L+1} \cdots \nu_N}(\psi).$$

On the one hand, these properties have been deduced from the Gårding–Wightman axioms via eqn [1]. Conversely, a family of distributions labeled in the above fashion and satisfying the above properties may be used to construct a Gårding–Wightman field theory provided that two more conditions – which hold for all systems of N-point functions – are satisfied. This requires some elementary notation.

Define the index sets

$$\mathcal{I}_N := \left\{ \begin{pmatrix} a_1 \cdots a_N \\ \nu_1 \cdots \nu_N \end{pmatrix} : 1 \leq a_i \leq m, 1 \leq \nu_i \leq n_{a_i} \text{ for all } 1 \leq i \leq N \right\}, \quad N \in \mathbb{N}$$

$\mathcal{I}_0 := \{\emptyset\}$, and $\mathcal{I} := \bigcup_{N \in \mathbb{N}_0} \mathcal{I}_N$. On $\mathcal{I}$ a concatenation $\circ$ is defined by

$$\begin{pmatrix} a_1 \cdots a_N \\ \nu_1 \cdots \nu_N \end{pmatrix} \circ \begin{pmatrix} b_1 \cdots b_M \\ \mu_1 \cdots \mu_M \end{pmatrix} := \begin{pmatrix} a_1 \cdots a_N b_1 \cdots b_M \\ \nu_1 \cdots \nu_N \mu_1 \cdots \mu_M \end{pmatrix}$$

and

$$\emptyset \circ \kappa := \kappa \circ \emptyset := \kappa$$

and an involution $*$ by

$$\begin{pmatrix} a_1 \cdots a_N \\ \nu_1 \cdots \nu_N \end{pmatrix}^* := \begin{pmatrix} a_N^* \cdots a_1^* \\ \nu_N \cdots \nu_1 \end{pmatrix} \quad \text{and} \quad \emptyset^* := \emptyset.$$

Define an antilinear involution $*$ on $\mathcal{S}^N := \mathcal{S}((\mathbb{R}^{1+3})^N)$ by

$$\psi(x_1 \cdots x_N) := \overline{\psi(x_N \cdots x_1)}$$

for each $N \in \mathbb{N}$. Put $\mathcal{S}^0 := \mathbb{C}$ and $z^* := \bar{z}$ for all $z \in \mathbb{C}$.

Define $\mathcal{S}^{\mathcal{I}_N} := \mathcal{S}^N \times \mathcal{I}_N$, and $\mathcal{S}^{\mathcal{I}} := \bigcup_N \mathcal{S}^{\mathcal{I}_N}$. For each $\kappa \in \mathcal{I}_N$, the set $\mathcal{S}^\kappa := \mathcal{S}((\mathbb{R}^{1+3})^N) \times \{\kappa\}$ is a linear space. On the direct sum $\mathcal{B}^{\mathcal{I}} := \bigoplus_{\kappa \in \mathcal{I}} \mathcal{S}^\kappa$ define an associative product by

$$(\psi, \kappa)(\chi, \lambda) := (\psi \otimes \chi, \kappa \circ \lambda)$$

and an antilinear involution $*$ by $(\psi, \kappa)^* := (\psi^*, \kappa^*)$. This endows $\mathcal{B}^{\mathcal{I}}$ with the structure of a nonabelian $*$-algebra with unit element $1 = (1, \emptyset)$ (Borchers algebra).

If one defines $F_\emptyset(z) := z\mathbf{1}$, then $w_\emptyset(z) = z$, and the Wightman functions induce a $\mathbb{C}$-linear functional ω on $\mathcal{B}^{\mathcal{I}}$ by

$$\omega(\psi, \kappa) := w_\kappa(\psi) \qquad [2]$$

ω exhibits the following two properties, which are the announced additional conditions required for reconstructing the fields from the N-point functions.

Hermiticity $\omega(\xi^*) = \overline{\omega(\xi)}$.

Positivity $\omega(\xi^* \xi) \geq 0$.

To see Hermiticity, compute

$$\begin{aligned} \omega(\psi^*, \kappa^*) &= \langle \Omega, F_{\kappa^*}(\psi^*)\Omega \rangle \\ &= \langle F_\kappa(\psi)\Omega, \Omega \rangle = \overline{\omega(\psi, \kappa)} \end{aligned}$$

and use $\mathbb{C}$-linearity to prove the statement for arbitrary $\xi \in \mathcal{B}$. For positivity, write any ξ as a finite sum $\xi = (\psi_1, \kappa_1) + \cdots + (\psi_M, \kappa_M)$, and compute

$$\begin{aligned} \omega(\xi^* \xi) &= \omega\left(\sum_{i,j=1}^M (\psi_i, \kappa_i)^* (\psi_j, \kappa_j) \right) \\ &= \omega\left(\sum_{ij} (\psi_i^* \otimes \psi_j, \kappa_i^* \circ \kappa_j) \right) \\ &= \sum_{ij} w_{\kappa_i^* \circ \kappa_j}(\psi_i^* \otimes \psi_j) \\ &= \sum_{ij} \langle \Omega, F_{\kappa_i^* \circ \kappa_j}(\psi_i^* \otimes \psi_j)\Omega \rangle \\ &= \sum_{ij} \langle \Omega, F_{\kappa_i^*}(\psi_i^*) F_{\kappa_j}(\psi_j)\Omega \rangle \\ &= \sum_{ij} \langle F_{\kappa_i}(\psi_i)\Omega, F_{\kappa_j}(\psi_j)\Omega \rangle \\ &= \left\| \sum_i F_{\kappa_i}(\psi_i)\Omega \right\|^2 \geq 0. \end{aligned}$$

Theorem 1 (Wightman's reconstruction theorem). *Let* m *and* $n_1 \cdots n_m$ *be natural numbers, let* $\mathcal{I}_0, \mathcal{I}_1, \mathcal{I}_2, \ldots,$ *and* $\mathcal{I}$ *be the above index sets, and let* $\mathcal{B}^{\mathcal{I}}$ *be the above Borchers algebra. Let* $D_1 \cdots D_m$ *be matrix representations of* $\mathrm{SL}(2, \mathbb{C})$ *in* $\mathbb{C}^{n_1} \cdots \mathbb{C}^{n_m}$, *respectively.*

For each natural number N, *let* $(w_\kappa)_{\kappa \in \mathcal{I}_N}$ *be a family of distributions on* $(\mathbb{R}^{1+3})^N$. *Suppose the family* $(w_\kappa)_{\kappa \in \mathcal{I}}$ *defined this way satisfies microcausality, covariance, spectrum condition, and the cluster property. If the linear functional* ω *defined on* $\mathcal{B}^{\mathcal{I}}$ *by eqn* [2] *is Hermitian and positive, then*

there is (up to unitary equivalence) a unique family $F^1 \cdots F^m$ *of Gårding–Wightman fields with* $n_1 \cdots n_m$ *components such that eqn* [1] *holds.*

The proof uses the GNS construction known from the theory of operator algebras. The Borchers algebra plays several roles. On the one hand, it is a linear space with an inner product. The Hilbert space $\mathcal{H}$ and the invariant space $\mathcal{D}$ of the field theory are constructed from this structure. On the other hand, the Borchers algebra acts on itself as an algebra of linear operators by its own algebra multiplication. This is the structure the $*$-algebra of field operators is constructed from.

Results

The mathematical and structural analysis of quantum fields has improved the understanding of scattering theory in the different approaches mentioned above; see Bogoliubov *et al.* (1975) and the relevant articles in this encyclopedia. Apart from this, the following results deserve to be mentioned. Evidently, many others have to be omitted for practical reasons.

PCT Symmetry

An early famous result was Lüders's proof (1957) that all fields in the above setting exhibit PCT symmetry, that is, the symmetry under reflections in all space and time variables combined with a charge conjugation. This symmetry is exhibited by all particle reactions observed so far. The proof, like several of the main results, made extensive use of the fact that the N-point functions are boundary values of analytic functions due to the spectrum condition, and that a fundamental theorem by Bargmann, Hall, and Wightman (1957) yields invariant analytic extensions.

Reeh–Schlieder Theorem

For each field F^a_ν and each bounded open region $\mathcal{O} \subset \mathbb{R}^{1+3}$, the vacuum vector is cyclic with respect to $\mathcal{F}^a_\nu(\mathcal{O})$ (Reeh and Schlieder 1961). So excitations of the vacuum vector by field operators located in $\mathcal{O}$ are not to be considered as state vectors of a particle localized in $\mathcal{O}$, since they are not perpendicular to the excitations by field operators located outside $\mathcal{O}$.

Unruh Effect and Modular P_1CT Symmetry

In the 1970s, Bisognano and Wichmann (1975, 1976) discovered a surprising link of symmetries to the intrinsic algebraic structure of quantum fields, which is established by the Tomita–Takesaki modular theory (*see* Tomita–Takesaki Modular Theory). Namely, the unitary operators implementing the Lorentz boosts on the fields are elements of modular groups. This means that a uniformly accelerated observer perceives the vacuum as a thermal state with a temperature proportional to its acceleration, corresponding to the famous Unruh effect.

In addition, it was shown that P_1CT symmetries (i.e., PCT combined with rotations by the angle π) are implemented by modular conjugations (modular P_1CT symmetry). Modular P_1CT symmetry is a consequence of the Unruh effect (Guido and Longo 1995).

Spin and Statistics

Immediately following Lüders's PCT theorem, the spin–statistics theorem was proved for the N-point functions of the Wightman setting (Lüders and Zumino 1958, Burgoyne 1958, Dell'Antonio 1961). This was a remarkable and widely acknowledged progress. But as remarked earlier, the confinement to finite-component fields, which is used in the proof, cannot be motivated by physical first principles (i.e., in a truly axiomatic fashion). The representation D of $SL(2, \mathbb{C})$ acting on the components, however, is forced to be finite dimensional by this assumption, and since the representations D^a are objects of investigation, a considerable part of the result is assumed this way from the outset. Even more so, there are examples of fields with a "wrong" spin–statistics connection and infinitely many components.

This was one reason to continue working on the subject. At the beginning of the 1990s, it was found that the spin–statistics theorem can be derived from the symmetries discovered by Bisognano and Wichmann, and Unruh. Two approaches not referring to the number of internal degrees of freedom have been worked out: one assumes the Unruh effect (Guido and Longo 1995), the other modular P_1CT symmetry (Kuckert 1995, 2005, Kuckert and Lorenzen 2005). The first approach has been generalized to conformal fields, the second to the case that the symmetry group's homogeneous part is not $SL(2, \mathbb{C})$, but only $SU(2)$.

Both approaches can be applied to infinite-component fields. They yield existence theorems; a distinguished representation is constructed from the modular symmetries, and this representation exhibits Pauli's spin–statistics connection. As mentioned before, nothing more can be expected at this level of generality. The line of argument works in both the algebraic and the Wightman setting.

A Dynamical Property of the Vacuum

One can derive the spectrum condition, the Bisognano–Wichmann symmetries/the Unruh effect, and

covariance from the condition that no (inertial or) uniformly accelerated observer can extract mechanical energy from the field *in vacuo* by means of a cyclic process (Kuckert 2002).

Interacting Fields

The examples of interacting quantum fields that fit into the above settings live in one or two spatial dimensions only, and their relevance for physics mainly consists in being such examples. This has contributed to some frustration and to doubts on whether one is not, in fact, proving theorems on pretty empty sets, or in other words, working on "the most sophisticated theory of the free field."

The computations in quantum field theory are, like most of the computations in physics, perturbative. In order to be successful, they need to yield good agreement with experiment with reasonable computational efforts, that is, by evolution up to the second or third order. This asymptotic convergence is more important than convergence of the series as a whole. There are low-dimensional examples of interacting Wightman fields (e.g., $(\varphi^4)_2$; cf. the monograph by Glimm and Jaffe (1987)), and time will tell whether four-dimensional interacting Wightman fields exist. But there is no reason to expect convergence for general interacting fields; for example, QED does not fit into the Wightman framework.

The appropriate extension of the Wightman setting has been formulated by Epstein and Glaser (1973). It defines the *S-matrix* rather than the field itself as a (in general divergent) formal power series of operator-valued distributions.

The above results apply to this somewhat more modest setting as well, so the "axiomatic" approaches do help in understanding the known high-energy physics interactions. This even includes gauge theories (*see* Perturbative Renormalization Theory and BRST). The high-precision results of QED can be reproduced within this setting, and there occur no UV singularities: renormalization amounts to the need to extend distributions by fixing some parameters, that is, the renormalization constants. The infrared problem is circumvented by considering the *S*-matrix as a (position-dependent) distribution taking values in the unitary formal power series of distributions rather than as a single (global) unitary operator (or unitary power series).

Quantum Energy Inequalities

Energy densities of Wightman fields admit negative expectation values (Epstein, Glaser, and Jaffe 1965). This is in contrast to the positivity conditions that the energy–momentum tensors of classical general (and, hence, also special) relativity have to satisfy to ensure causality. But the conflict can be solved by smearing the densities out in space or time, as has first been realized by Ford (1991). The extent to which the energy density can become negative depends on the extent to which it is smeared out: "more smearing means less violation of positivity," so the classical positivity conditions are restored at medium and large scales. There are many ways to make this principle concrete. Quantum energy inequalities hold for thermodynamically well-behaved quantum fields on causally well-behaved classical spacetime backgrounds.

Bibliographic Notes

Important monographs on axiomatic quantum field theory are those by Streater and Wightman (1964), Jost (1965), Bogoliubov *et al.* (1975), and Bogoliubov *et al.* (1990). Note that the books of Bogoliubov *et al.* differ in setup fundamentally and that neither replaces the other. For a lecture notes volume, see also Völkel (1977), and for a review article, see Streater (1975). A valuable discussion of the Wightman axioms can also be found in the second volume of the series by Reed and Simon (1970).

The first monograph on the algebraic approach to quantum field theory is due to Haag (1992), a more recent one has been written by Araki (1999). Concerning the sufficient conditions for "switching" between the Gårding–Wightman and the algebraic approach, see Wollenberg (1988) and the Ph.D. thesis of Bostelmann (2000) and references given there. Dynamical and thermodynamical foundation of standard axioms, the Bisognano–Wichmann symmetries (Unruh effect), and the spin–statistics theorem, have been investigated by Kuckert (2002, 2005), see also the references given there for related work.

In different formulations and at differing degrees of mathematical sophistication, the causal approach to perturbation theory can be found in the monographs by Bogoliubov and Shirkov (1959), Scharf (1989, 2001), and Steinmann (2000). Two modern review articles have been written by Brunetti and Fredenhagen (2000) and by Dütsch and Fredenhagen (2004).

The reference original articles on the Euclidean axioms are those of Osterwalder and Schrader (1973, 1975). Note that the first one contains an error. (cf. also Zinoviev (1995)). A monograph on Euclidean field theory and its relations to the other axiomatic settings of quantum field theory and to statistical mechanics is that by Glimm and Jaffe (1987).

A recent review on quantum energy inequalities is due to Fewster (2003).

Acknowledgments

The author is a fellow of the Emmy-Noether Programme (DFG). Thanks for discussions are due to Professor D Arlt.

See also: Algebraic Approach to Quantum Field Theory; C^*-Algebras and Their Classification; Constructive Quantum Field Theory; Dispersion Relations; Euclidean Field Theory; Indefinite Metric; Perturbative Renormalization Theory and BRST; Quantum Field Theory: A Brief Introduction; Quantum Field Theory in Curved Spacetime; Scattering, Asymptotic Completeness and Bound States; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Scattering in Relativistic Quantum Field Theory: The Analytic Program; Symmetries in Quantum Field Theory: Algebraic Aspects; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Thermal Quantum Field Theory; Tomita–Takesaki Modular Theory; Two-Dimensional Models.

Further Reading

Araki H (1999) *Mathematical Theory of Quantum Fields.* Oxford: Oxford University Press.

Bogoliubov NN, Logunov AA, and Todorov IT (1975) *Introduction to Axiomatic Quantum Field Theory*, (Russian original edition: Nauka (Moskow) 1969). New York: Benjamin.

Bogoliubov NN, Logunov AA, Oksak AI, and Todorov IT (1990) *General Principles of Quantum Field Theory*, (Russian original edition Nauka (Moskow) 1987). Dordrecht–Boston–London: Kluwer.

Bostelmann H (2000) *Lokale Algebren und Operatorprodukte am Punkt* (in German). Ph.D. thesis, Göttingen.

Brunetti R and Fredenhagen K (2004) Microlocal Analysis and Interacting Quantum Field Theories: Renormalization on Physical Backgrounds. *Communications in Mathematical Physics* 208: 623.

Dütsch and Fredenhagen K (2004) Causal Perturbation Theory in terms of retarded products, and a proof of the Action Ward Identity, to appear in. *Rev. Math. Phys.*

Fewster CJ (2003) *Energy Inequalities in Quantum Field Theory.* Proceedings of the International Conference on Mathematical Physics (revised version under math-ph/0501073).

Glimm J and Jaffe A (1987) *Quantum Physics: A Functional Integral Point of View*, 2nd edn. Berlin–Heidelberg–New York: Springer.

Guido D and Longo R (1995) An algebraic spin and statistics Theorem. *Communications in Mathematical Physics* 172: 517.

Haag R (1992) *Local Quantum Physics.* Berlin–Heidelberg–New York: Springer.

Jost R (1965) *The General Theory of Quantized Fields.* American Mathematical Society.

Kuckert B (2002) Covariant thermodynamics of quantum systems: passivity, semipassivity, and the Unruh effect. *Annals of Physics* 295: 216.

Kuckert B (2005) Spin, statistics, and reflections, I. *Annales Henri Poincaré* 6: 849.

Kuckert B and Lorenzen R (2005) Spin, Statistics, and Reflections, II. Preprint (math-ph/0512068).

Osterwalder K and Schrader R (1973) Axioms for Euclidean Green's functions. *Communications in Mathematical Physics* 31: 83.

Osterwalder K and Schrader R (1975) Axioms for Euclidean Green's functions. 2. *Communications in Mathematical Physics* 42: 281.

Reed M and Simon B (1970) *Methods of Modern Mathematical Physics*, (4 volumes). London: Academic Press.

Scharf G (1989) *Finite Quantum Electrodynamics.* Berlin–Heidelberg–New York: Springer.

Scharf G (2001) *Quantum Gauge Theories A True Ghost Story.* Weinheim: Wiley.

Streater RF (1975) Outline of axiomatic quantum field theory. *Reports on Progress in Physics* 38: 771.

Streater RF and Wightman AS (1964) *PCT, Spin & Statistics, and All That.* New York: Benjamin.

Völkel AH (1977) Fields, Particles, and Currents. *Lecture Notes in Physics*, vol. 66. Berlin–Heidelberg–New York: Springer.

Wollenberg M (1988) The existence of quantum fields for local nets of observables. *Journal of Mathematical Physics* 29: 2106.

Zinoviev YM (1995) Equivalence of Euclidean and Wightman field theories. *Communications in Mathematical Physics* 174: 1.

Batalin–Vilkovisky Quantization

A C Hirshfeld, Universität Dortmund, Dortmund, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The Batalin–Vilkovisky formalism for quantizing gauge theories has a long history of development. It begins with the Faddeev–Popov procedure for quantizing Yang–Mills theory, involving the Faddeev–Popov ghost fields (Faddeev and Popov 1967). It continued with the discovery of BRST symmetry by Becchi *et al.* (1976). Then Zinn-Justin (1975) introduced sources for these transformations, and a symmetric structure in the space of fields and sources in his study of renormalizability of these theories. Finally, Batalin and Vilkovisky (1981) systematized and generalized these developments. A more detailed account of this history can be found in Gomis *et al.* (1994), where many worked examples of the Batalin–Vilkovisky formalism are given. At the present time, it is the most general treatment available. Alexandrov, Kontsevich, Schwarz, and Zabaronsky (AKSZ 1997) have presented a geometric interpretation for the case in which the action is topologically invariant.

Structure of the Set of Gauge Transformations

Consider a system whose dynamics is governed by a classical action $S[\phi^i]$ which depends on the fields $\phi^i(x), i=1,\ldots,n$. We employ a compact notation in which the multi-index i may denote the various fields involved, the discrete indices on which they depend, and the dependence on the spacetime variables as well. The generalized summation convention then means that a repeated index may denote not only a sum over discrete variables, but also integration over the spacetime variables. $\epsilon_i = \epsilon(\phi^i)$ denotes the

Notes:

Grassmann parity of the fields. Fields with $\epsilon_i = 0$ are called bosonic, with $\epsilon_i = 1$ fermionic. The graded commutation rule is

$$\phi^i(x)\phi^j(y) = (-1)^{\epsilon_i\epsilon_j}\phi^j(y)\phi^i(x) \qquad [1]$$

For a gauge theory the action is invariant under a set of gauge transformations with infinitesimal form

$$\delta\phi^i = R^i_\alpha \varepsilon^\alpha, \quad \alpha = 1 \text{ or } 2 \text{ or} \ldots m \qquad [2]$$

The ε^α are the infinitesimal gauge parameters and R^i_α the generators of the gauge transformations. When $\epsilon_\alpha = \epsilon(\varepsilon^\alpha) = 0$ we have an ordinary symmetry, when $\epsilon_\alpha = 1$ the equation is characteristic of a supersymmetry. The Grassmann parity of R^i_α is $\epsilon(R^i_\alpha) = \epsilon_i + \epsilon_\alpha \pmod 2$.

A subscript after a comma denotes the right derivative with respect to the corresponding field, that is, the field is to be commutated to the far right and then dropped. The field equations may then be written as

$$S_{0,i} = 0 \qquad [3]$$

where S_0 is the classical action. Let Σ denote the surface in the space of solutions where the field equations are satisfied:

$$S_{0,i}|_\Sigma = 0 \qquad [4]$$

If the gauge transformations are "independent" on-shell, that is,

$$\operatorname{rank} R^i_\alpha|_\Sigma = m \qquad [5]$$

the gauge theory is said to be "irreducible." We assume here that this is the case. When it is not, the theory is "reducible." For details of the treatment in that case, see Gomis, Paris, and Samuel. The classical solutions are $\phi_0 \in \Sigma$.

The Noether identities are

$$S_{0,i}R^i_\alpha = 0 \qquad [6]$$

The general solution to the Noether identity is

$$\lambda^i = R^i_\alpha T^\alpha + S_{0,j}E^{ji} \qquad [7]$$

The commutator of two gauge transformations is

$$[\delta_1, \delta_2]\phi^i = \left(R^i_{\alpha,j}R^j_\beta - (-1)^{\epsilon_\alpha\epsilon_\beta}R^i_{\beta,j}R^j_\alpha\right)\varepsilon^\beta_1\varepsilon^\alpha_2 \qquad [8]$$

Since this commutator is a symmetry of the action, it satisfies the Noether identity

$$S_{0,i}\left(R^i_{\alpha,j}R^j_\beta - (-1)^{\epsilon_\alpha\epsilon_\beta}R^i_{\beta,j}R^j_\alpha\right) = 0 \qquad [9]$$

which by eqn [7] implies that

$$R^i_{\alpha,j}R^j_\beta - (-1)^{\epsilon_\alpha\epsilon_\beta}R^i_{\beta,j}R^j_\alpha = R^i_\gamma T^\gamma_{\alpha\beta} + S_{0,j}E^{ji}_{\alpha\beta} \qquad [10]$$

Equations [8] and [10] lead to the following condition:

$$[\delta_1, \delta_2]\phi^i = \left(R^i_\gamma T^\gamma_{\alpha\beta} - S_{0,j}E^{ji}_{\alpha\beta}\right)\varepsilon^\beta_1\varepsilon^\alpha_2 \qquad [11]$$

The tensors $T^\gamma_{\alpha\beta}$ are called the structure constants of the gauge algebra, although they depend, in general, on the fields of the theory. When $E^{ij}_{\alpha\beta} = 0$, the gauge algebra is said to be "closed," otherwise it is "open." Equation [11] defines a Lie algebra if the algebra is closed and the $T^\gamma_{\alpha\beta}$ are independent of the fields.

The gauge tensors have the following graded symmetry properties:

$$\begin{aligned} T^\gamma_{\alpha\beta} &= -(-1)^{\epsilon_\alpha\epsilon_\beta}T^\gamma_{\beta\alpha} \\ E^{ij}_{\alpha\beta} &= -(-1)^{\epsilon_\alpha\epsilon_\beta}E^{ji}_{\alpha\beta} = -(-1)^{\epsilon_\alpha\epsilon_\beta}E^{ij}_{\beta\alpha} \end{aligned} \qquad [12]$$

The Grassmann parities are

$$\epsilon(T^\gamma_{\alpha\beta}) = \epsilon_\alpha + \epsilon_\beta + \epsilon_\gamma \pmod 2 \qquad [13]$$

and

$$\epsilon(E^{ij}_{\alpha\beta}) = \epsilon_i + \epsilon_j + \epsilon_\alpha + \epsilon_\beta \pmod 2 \qquad [14]$$

Various restrictions are imposed by the Jacobi identity

$$\sum_{\text{cyclic}(123)} [\delta_1, [\delta_2, \delta_3]] = 0 \qquad [15]$$

These restrictions are

$$\sum_{\text{cyclic}(123)} \left(R^i_\delta A^\delta_{\alpha\beta\gamma} - S_{0,j}B^{ji}_{\alpha\beta\gamma}\right)\varepsilon^\gamma\varepsilon^\beta\varepsilon^\alpha = 0 \qquad [16]$$

where

$$\begin{aligned} 3A^\delta_{\alpha\beta\gamma} \equiv{}& \left(T^\delta_{\alpha\beta k}R^k_\gamma - T^\delta_{\alpha\eta}T^\eta_{\beta\gamma}\right) + (-1)^{\epsilon_\alpha(\epsilon_\beta+\epsilon_\gamma)} \\ &\times \left(T^\delta_{\beta\gamma k}R^k_\gamma - T^\delta_{\beta\eta}T^\eta_{\gamma\alpha}\right) \\ &+ (-1)^{\epsilon_\gamma(\epsilon_\alpha+\epsilon_\beta)}\left(T^\delta_{\gamma\alpha k}R^k_\beta - T^\delta_{\gamma\eta}T^\eta_{\alpha\beta}\right) \end{aligned}$$

and

$$\begin{aligned} 3B^{ji}_{\alpha\beta\gamma} \equiv{}& \left(E^{ji}_{\alpha\beta k}R^k_\alpha - E^{ji}_{\alpha\delta}T^\delta_{\beta\gamma} - (-1)^{\epsilon_i\epsilon_\alpha}\right. \\ &\left.\times R^j_{\alpha,k}E^{ki}_{\beta\gamma} + (-1)^{\epsilon_j(\epsilon_i+\epsilon_\alpha)}R^i_{\alpha,k}E^{kj}_{\beta\gamma}\right) \\ &+ (-1)^{\epsilon_\alpha(\epsilon_\beta+\epsilon_\gamma)}(\alpha \to \beta \to \gamma) + (-1)^{\epsilon_\gamma(\epsilon_\alpha+\epsilon_\beta)} \\ &\times (\alpha \to \gamma \to \beta) \end{aligned}$$

As in the familiar Faddeev–Popov procedure, it is useful to introduce ghost fields C^α with opposite Grassmann parities to the gauge parameters ε^α:

$$\epsilon(C^\alpha) = \epsilon_\alpha + 1 \pmod 2 \qquad [17]$$

and to replace the gauge parameters by ghost fields. One must then modify the graded symmetry properties of the gauge structure tensors according to

$$T_{\alpha_1\alpha_2\alpha_3\alpha_4\ldots} \to (-1)^{\epsilon^{\alpha_2+\alpha_4+\cdots}} T_{\alpha_1\alpha_2\alpha_3\alpha_4\ldots} \qquad [18]$$

The Noether identities then take the form

$$S_{0,i} R^i_\alpha C^\alpha = 0 \qquad [19]$$

and the structure relations [10] become

$$(2R^i_{\alpha,j} R^j_\beta - R^i_\gamma T^\gamma_{\alpha\beta} + S_{0,j} E^{ji}_{\alpha\beta}) C^\beta C^\alpha = 0 \qquad [20]$$

Introducing the Antifields

We incorporate the ghost fields into the field set $\Phi^A = \{\phi^i, C^\alpha\}$, where $i = 1, \ldots, n$ and $\alpha = 1, \ldots, m$. Clearly $A = 1, \ldots, N$, where $N = n + m$. One then further increases the set by introducing an antifield Φ^*_A for each field Φ^A. The Grassmann parity of the antifields is

$$\epsilon(\Phi^*_A) = \epsilon(\Phi^A) + 1 \pmod 2 \qquad [21]$$

Each field is assigned a ghost number, with

$$\begin{aligned} \mathrm{gh}[\phi^i] &= 0 \\ \mathrm{gh}[C^\alpha] &= 1 \\ \mathrm{gh}[\Phi^*_A] &= -\mathrm{gh}[\Phi^A] - 1 \end{aligned} \qquad [22]$$

In the space of fields and antifields, the antibracket is defined by

$$(X, Y) = \frac{\partial_r X}{\partial \Phi^A} \frac{\partial_l Y}{\partial \Phi^*_A} - \frac{\partial_r X}{\partial \Phi^*_A} \frac{\partial_l Y}{\partial \Phi^A} \qquad [23]$$

where ∂_r denotes the right, ∂_l the left derivative. The antibracket is graded antisymmetric:

$$(X, Y) = -(-1)^{(\epsilon_X+1)(\epsilon_Y+1)} (Y, X) \qquad [24]$$

It satisfies a graded Jacobi identity

$$\begin{aligned} &((X, Y), Z) + (-1)^{(\epsilon_X+1)(\epsilon_Y+1)} \\ &\quad \times ((Y, Z), X) + (-1)^{(\epsilon_Z+1)(\epsilon_X+\epsilon_Y)} ((Z, X), Y) = 0 \end{aligned} \qquad [25]$$

It is a graded derivation

$$\begin{aligned} (X, YZ) &= (X, Y)Z + (-1)^{\epsilon_X \epsilon_Y} (X, Z)Y \\ (XY, Z) &= X(Y, Z) + (-1)^{\epsilon_X \epsilon_Y} Y(X, Z) \end{aligned} \qquad [26]$$

It has ghost number

$$\mathrm{gh}[(X, Y)] = \mathrm{gh}[X] + \mathrm{gh}[Y] + 1 \qquad [27]$$

and Grassmann parity

$$\epsilon((X, Y)) = \epsilon(X) + \epsilon(Y) + 1 \pmod 2 \qquad [28]$$

For bosonic fields

$$(B, B) = 2 \frac{\partial B}{\partial \Phi^A} \frac{\partial B}{\partial \Phi^*_A} \qquad [29]$$

for fermionic fields

$$(F, F) = 0 \qquad [30]$$

and for any X

$$((X, X), X) = 0 \qquad [31]$$

If one groups the fields and the antifields together into the set

$$z^a = \{\Phi^A, \Phi^*_A\}, \quad a = 1, \ldots, 2N \qquad [32]$$

then the antibracket is seen to define a symplectic structure on the space of fields and antifields

$$(X, Y) = \frac{\partial_r X}{\partial z^a} \omega^{ab} \frac{\partial_l Y}{\partial z^b} \qquad [33]$$

with

$$\omega^{ab} = \begin{pmatrix} 0 & \delta^A_B \\ -\delta^A_B & 0 \end{pmatrix} \qquad [34]$$

The antifields can be thought of as conjugate variables to the fields, since

$$(\Phi^A, \Phi^*_B) = \delta^A_B \qquad [35]$$

The Classical Master Equation

Let $S[\Phi^A, \Phi^*_A]$ be a functional of the fields and antifields with the dimension of an action, vanishing ghost number and even Grassmann parity. The equation

$$(S, S) = 2 \frac{\partial S}{\partial \Phi^A} \frac{\partial S}{\partial \Phi^*_A} = 0 \qquad [36]$$

is the classical master equation. Solutions of the classical master equation with suitable boundary conditions turn out to be generating functionals for the gauge structure of the theory. S is also the starting point for the quantization. One denotes by Σ the subspace of stationary points of the action in the space of fields and antifields:

$$\Sigma = \left\{ z^a \middle| \frac{\partial S}{\partial z^a} = 0 \right\} \qquad [37]$$

Given a classical solution ϕ_0 of S_0 one stationary point is

$$\phi^i = \phi^i_0, \quad C^a = 0, \quad \Phi^*_A = 0 \qquad [38]$$

An action which satisfies the classical master equation has its own set of invariances:

$$\frac{\partial S}{\partial z^a} R_b^a = 0 \qquad [39]$$

with

$$R_b^a = \omega^{ac} \frac{\partial_l \partial_r S}{\partial z^c \partial z^b} \qquad [40]$$

This equation implies

$$R_c^a R_b^a|_\Sigma = 0 \qquad [41]$$

One says that R_b^a is invariant on-shell. A nilpotent $2N \times 2N$ matrix has rank $\leq N$. Let r be the rank of the hessian of S at the stationary point:

$$r = \text{rank} \left. \frac{\partial_l \partial_r S}{\partial z^a \partial z^b} \right|_\Sigma \qquad [42]$$

We then have $r \leq N$. The relevant solutions of the classical master equation are those for which $r = N$. In this case the number of independent gauge invariances of the type in eqn [39] equals the number of antifields. When at a later stage the gauge is fixed, the nonphysical antifields are eliminated.

To ensure the correct classical limit, the proper solution must contain the classical action S_0 in the sense that

$$S[\Phi^A, \Phi_A^*]|_{\Phi_A^* = 0} = S_0[\phi^i] \qquad [43]$$

The action $S[\Phi^A, \Phi_A^*]$ can be expanded in a series in the antifields, while maintaining vanishing ghost number and even Grassmann parity:

$$S[\Phi, \Phi^*] = S_0 + \phi_i^* R_\alpha^i C^\alpha + C_a^* \tfrac{1}{2} T_{\beta\gamma}^\alpha (-1)^{\epsilon_\beta} C^\gamma C^\beta + \phi_i^* \phi_j^* (-1)^{\epsilon_i} \tfrac{1}{4} E_{\alpha\beta}^{ji} (-1)^{\epsilon_\alpha} C^\beta C^\alpha + \cdots \qquad [44]$$

When this is inserted into the classical master equation, one finds that this equation implies the gauge structure of the classical theory.

Gauge Fixing and Quantization

Equation [39] shows that the action S still possesses gauge invariances, and hence is not yet suitable for quantization via the path integral approach: a gauge-fixing procedure is necessary. In the Batalin–Vilkovisky approach the gauge is fixed, and the antifields eliminated, by use of a gauge-fixing fermion Ψ which has Grassmann parity $\epsilon(\Psi) = 1$ and $\text{gh}[\Psi] = -1$. It is a functional of the fields Φ^A only; its relation to the antifields is

$$\Phi_A^* = \frac{\partial \Psi}{\partial \Phi^A} \qquad [45]$$

We define a surface in functional space

$$\Sigma_\Psi = \left\{ (\Psi^A, \Psi_A^*) | \Psi_A^* = \frac{\partial \Psi}{\partial \Phi^A} \right\} \qquad [46]$$

so that for any functional $X[\Phi, \Phi^*]$

$$X|_{\Sigma_\Psi} = X\left[\Psi, \frac{\partial \Psi}{\partial \Phi}\right] \qquad [47]$$

To construct a gauge-fixing fermion Ψ of ghost number -1, one must again introduce additional fields. The simplest choice utilizes a trivial pair $\bar{C}_\alpha, \bar{\pi}_\alpha$ with

$$\epsilon(\bar{C}_\alpha) = \epsilon_\alpha + 1, \qquad \epsilon(\bar{\pi}_\alpha) = \epsilon_\alpha$$
$$\text{gh}[\bar{C}_\alpha] = -1, \qquad \text{gh}[\bar{\pi}_\alpha] = 0 \qquad [48]$$

The fields $\bar{C}_\alpha$ are the Faddeev–Popov antighosts. Along with these fields we include the corresponding antifields $\bar{C}^{*\alpha}, \bar{\pi}^{*\alpha}$. Adding the term $\bar{\pi}_\alpha \bar{C}^{*\alpha}$ to the action S does not spoil its properties as a proper solution to the classical master equation, and one gets the nonminimal action

$$S^{\text{non}} = S + \bar{\pi}_\alpha \bar{C}^{*\alpha} \qquad [49]$$

The simplest possibility for Ψ is

$$\Psi = \bar{C}_\alpha \chi^\alpha(\phi) \qquad [50]$$

where χ^α are the gauge-fixing conditions for the fields ψ. The gauge-fixed action is denoted by

$$S_\Psi = S^{\text{non}}|_{\Sigma_\Psi} \qquad [51]$$

Quantization is performed using the path integral to calculate a correlation function X, with the constraint [45] implemented by a δ-function:

$$I_\Psi(X) = \int D\Phi D\Phi^* \delta\left(\Phi_A^* - \frac{\partial \Psi}{\partial \Phi^A}\right) \times \exp\left(\frac{\mathrm{i}}{\hbar} W[\Phi, \Phi^*]\right) X[\Phi, \Phi^*] \qquad [52]$$

Here W is the quantum action, which reduces to S in the limit $\hbar \to 0$. An admissible Ψ leads to well-defined propagators when the path integral is expressed as a perturbation series expansion.

The results of a calculation should be independent of the gauge fixing. Consider the integrand in eqn [52],

$$I[\Phi, \Phi^*] = \exp\left(\frac{\mathrm{i}}{\hbar} W[\Phi, \Phi^*]\right) X[\Phi, \Phi^*] \qquad [53]$$

Under an infinitesimal change in Ψ

$$I_{\Psi + \delta\Psi}(X) - I_\Psi(X) \approx \int D\Phi \, \Delta I \delta\Psi \qquad [54]$$

where the Laplacian Δ is

$$\Delta = (-1)^{\epsilon_A+1}\frac{\partial}{\partial\Phi^A}\frac{\partial}{\partial\Phi^*_A} \qquad [55]$$

Obviously, the integral $I_\Psi(X)$ is independent of Ψ if $\Delta I = 0$. For $X = 1$ one gets the requirement

$$\Delta\exp\left(\frac{\mathrm{i}}{\hbar}W\right) = \exp\left(\frac{\mathrm{i}}{\hbar}W\right) \times\left(\frac{\mathrm{i}}{\hbar}\Delta W - \frac{1}{2\hbar^2}(W,W)\right) = 0 \qquad [56]$$

The formula

$$\tfrac{1}{2}(W,W) = \mathrm{i}\hbar\Delta W \qquad [57]$$

is the quantum master equation. A gauge-invariant correlation function satisfies

$$(X,W) = \mathrm{i}\hbar\Delta X \qquad [58]$$

The terms of higher order in $\hbar$ by which the quantum action W may differ from the solution of the classical master equation S correspond to the counter-terms of the renormalizable gauge theory if

$$\Delta S = 0 \qquad [59]$$

One must, of course, use a regularization scheme which respects the symmetries of the theory. For $W = S + O(\hbar)$ the quantum master equation [57] reduces in this case to the classical master equation

$$(S,S) = 0 \qquad [60]$$

Hence, up to possible counter-terms, one may simply choose $W = S$.

To implement the gauge fixing, one uses for the action $W = S^{\mathrm{non}}$. For the path integral $Z = I_\Psi(X = 1)$, the integration over the antifields in eqn [52] is performed by using the δ-function. The result is

$$Z = \int D\Phi\exp\left(\frac{\mathrm{i}}{\hbar}S_\Psi\right) \qquad [61]$$

Geometrical Interpretation of Topological Field Theories

The Batalin–Vilkovisky formalism for topological field theories has been given a geometrical interpretation by AKSZ (1997).

A supermanifold equipped with an odd vector field satisfying $Q^2 = 0$ is called a Q-manifold. A Q-manifold provided with an odd symplectic structure ω (P-structure) is called a QP-manifold if the odd symplectic structure is Q-invariant, that is, $L_Q\omega = 0$. Every solution to the classical master equation determines a QP-structure on M and vice versa. The geometric object corresponding to a classical mechanical system in the Batalin–Vilkovisky formalism is a QP-manifold.

The nondegenerate closed 2-form ω is written as

$$\omega = dz^a\zeta_{ab}dz^b \qquad [62]$$

where z^a are local coordinates in the supermanifold M. For functions on M, an (odd) Poisson bracket is defined as in eqn [33], where ω^{ab} stands for the inverse matrix of ω_{ab}. An even function S on M satisfies the classical master equation if $(S,S) = 0$. The correspondence between vector fields and functions on M is given by $K_F G = (G,F)$, where K_F is the vector field, F the given function, and G an arbitrary function. The function F is called the Hamiltonian of the vector field K_F.

Geometrically, equivalent QP-manifolds describe the same physics. In particular, one can consider an even Hamiltonian vector field K_F corresponding to an odd function F. This vector field determines an infinitesimal transformation preserving P-structure. It transforms a solution S to the classical master equation into the physically equivalent solution $S + \epsilon(S,F)$, where ϵ is an infinitesimally small parameter.

A submanifold L of a P-manifold M is called a Lagrangian submanifold if the restriction of the form ω to L vanishes. In the particular case when $M = \Pi T^*N$ (the cotangent bundle to N with reversed parity of fibres) with standard P-structure, one can construct many examples of Lagrangian submanifolds in the following way. Fix an odd function Ψ on N, the gauge fermion. The submanifold $L_\Psi \in M$ determined by the equation

$$\xi^a = \frac{\partial\Psi}{\partial x^a} \qquad [63]$$

where $\{x^a,\xi_a\}$ are coordinates corresponding to the identification of M, will be a Lagrangian submanifold of M.

The P-manifold M in the neighborhood of L can be identified with ΠT^*L. In other words, one can find such a neighborhood U of L in M and a neighborhood V of L in ΠT^*L that there exists an isomorphism of P-manifolds U and V leaving L intact. Using this isomorphism a function Ψ defined on a Lagrangian submanifold $L \subset M$ determines another Lagrangian submanifold $L_\Psi \subset M$.

Consider a solution S to the classical master equation on M. In the Batalin–Vilkovisky formalism we have to restrict S to a Lagrangian submanifold $L \in M$, then the quantization of S can be performed by integration of $\exp(\mathrm{i}S/\hbar)$ over L. One may construct an odd vector field Q on L in such a

way that the functional S restricted to L is Q-invariant. This invariance is BRST invariance.

AKSZ apply these geometric constructions to obtain in a natural way the action functionals of two-dimensional sigma-models (Witten 1998) and to show that the Chern–Simons theory (Axelrod and Singer 1991) in Batalin–Vilkovisky formalism arises as a sigma-model with target space $\Pi\mathcal{G}$, where $\mathcal{G}$ stands for a Lie algebra and Π denotes parity inversion.

The Poisson-Sigma Model

The quantization of the Poisson-sigma model was performed by Hirshfeld and Schwarzweller (2000) and by Cattaneo and Felder (2001). The Poisson-sigma model is the simplest topological field theory in two dimensions. It is a field theory on a two-dimensional world sheet without boundary (Schaller and Strobl 1994). It involves a set of bosonic scalar fields, which can be seen as a set of maps $X^i : M \to N$, where N is a Poisson manifold. In addition, one has a 1-form A on the world sheet M which takes values in $T^*(N)$, for x coordinates on M we have $A = A_{\mu i} \mathrm{d}x^i \wedge \mathrm{d}X^i$. Its action is

$$S_0[X, A] = \int_M \mu\left(\epsilon^{\mu\nu}(A_{\mu i}\partial_\nu X^i + P^{ij}(X)A_{\mu i}A_{\nu j}\right) \quad [64]$$

where $\epsilon^{\mu\nu}$ is the antisymmetric tensor and μ is the volume form on M. The gauge transformations of the model are

$$\delta X^i = P^{ij}(X)\varepsilon_j, \qquad \delta A_{\mu i} = D^j_{\mu i}\varepsilon_j \quad [65]$$

where $D^j_{\mu i} = \partial_\mu \delta^j_i + P^{kj}{}_{,i}A_{\mu k}$. The equations of motion are

$$\epsilon^{\mu\nu} D^j_{\mu i} A_{\nu j} = 0 \quad [66]$$

and

$$\epsilon^{\mu\nu}(\partial_\nu X^i + P^{ij}A_{\nu j}) = \epsilon^{\mu\nu} D_\nu X^i = 0 \quad [67]$$

The gauge algebra is given by

$$\begin{aligned}
[\delta(\varepsilon_1), \delta(\varepsilon_2)]X^i &= P^{ji}(P^{mn}{}_{,j}\varepsilon_{1n}\varepsilon_{2m}) \\
[\delta(\varepsilon_1), \delta(\varepsilon_2)]A_{\mu i} &= D^j_{\mu i}(P^{mn}{}_{,j}\varepsilon_{1n}\varepsilon_{2m}) \\
&\quad - (\epsilon^{\nu\rho} D_\rho X^j)\epsilon_{\nu\mu} P^{mn}{}_{,ji}\varepsilon_{1n}\varepsilon_{2m}
\end{aligned} \quad [68]$$

In our general notation the generators of the gauge transformations R are here P^{ij} and $D^j_{\mu i}$. The gauge tensors T and E are $P^{ij}{}_{,k}$ and $\epsilon_{\nu\mu}P^{mn}{}_{,ji}$. The higher-order gauge tensors A and B vanish.

The ghost fields are again denoted by C^i. The Noether identities are then

$$\int_M \mu\left(\epsilon^{\mu\nu} D^j_{\mu i} A_{\nu j} P^{ki} + (\epsilon^{\mu\nu} D_\nu X^i) D^k_{\mu i}\right) C_k = 0 \quad [69]$$

Considering the commutator of two gauge transformations leads to (see eqns [8]–[11])

$$\begin{aligned}
&\int_M \mu\left(2P^{mi}{}_{,j}P^{nj} - P^{ji}P^{mn}{}_{,j}\right) C_m C_n = 0 \\
&\int_M \mu\Big(2(P^{jk}{}_{,i}D^l_{\mu j} + P^{mk}{}_{,ij}A_{\mu m}P^{jl}) \\
&\quad - D^m_{\mu i}P^{kl}{}_{,m} + (\epsilon^{\rho\nu}D_\rho X^j)\epsilon_{\mu\nu}P^{kl}{}_{,ji}\Big) C_l C_k = 0
\end{aligned} \quad [70]$$

The Jacobi identity is

$$P^{ij}{}_{,m}P^{mk} C_i C_j C_k = 0 \quad [71]$$

The fields and antifields of the model are

$$\Phi^A = \{A^{\mu i}, X^i, C_i\} \quad \text{and} \quad \Phi^*_A = \{A^{\mu i *}, X^*_i, C^{i*}\} \quad [72]$$

The extended action is

$$\begin{aligned}
S = \int_M \mu\Big(&\epsilon^{\mu\nu}(A_{\mu i}\partial_\nu X^i + P^{ij}(X)A_{\mu i}A_{\nu j}) \\
&+ A^{\mu i *} D^j_{\mu i} C_j + X^*_i P^{ji}(X) C_j + \frac{1}{2} C^{i*} P^{jk}{}_{,i}(X) C_j C_k \\
&+ \frac{1}{4} A^{\mu i *} A^{\nu j *} \epsilon_{\mu\nu} P^{kl}{}_{,ij}(X) C_k C_l\Big)
\end{aligned} \quad [73]$$

The gauge-fixing conditions are taken to be of the form $\chi_i(A, X)$, so that the gauge fermion [50] becomes $\Psi = \bar{C}^i \chi_i(A, X)$. The antifields are then fixed to be

$$\begin{aligned}
A^*_{\mu i} &= \bar{C}_j \frac{\partial \chi_j(A, X)}{\partial A_{\mu i}} \\
X^*_i &= \bar{C}_j \frac{\partial \chi_j(A, X)}{\partial X^i} \\
C^*_i &= 0 \\
\bar{C}^*_i &= \chi_i(A, X)
\end{aligned} \quad [74]$$

The gauge-fixed action is

$$\begin{aligned}
S_\Psi = \int_M \mu\Big(&\epsilon^{\mu\nu}(A_{\mu i}\partial_\nu X^i + P^{ij}(X)A_{\mu i}A_{\nu j}) \\
&+ \bar{C}^k \frac{\partial \chi_k(A, X)}{\partial A_{\mu i}} D^j_{\mu i} C_j + \bar{C}^k \frac{\partial \chi_k(A, X)}{\partial X^i} P^{ij} C_j \\
&+ \frac{1}{4} \bar{C}^m \frac{\partial \chi_m(A, X)}{\partial A_{\mu i}} \bar{C}^n \frac{\partial \chi_n(A, X)}{\partial A_{\nu j}} \epsilon_{\mu\nu} P^{kl}{}_{,ij}(X) \\
&\times C_k C_l + \bar{\pi}^i \chi_i(A, X)\Big)
\end{aligned} \quad [75]$$

Now consider different gauge conditions:

1. First, the Landau gauge for the gauge potential $\chi_i = \partial^\mu A_{\mu i}$, so that the gauge fermion becomes $\Psi = \bar{C}^i \partial^\mu A_{\mu i}$. The antifields are fixed to be

$$\begin{aligned}
A^{*\mu i} &= \partial^\mu \bar{C}^i \\
X^*_i &= C^{*i} = 0 \\
\bar{C}^*_i &= \partial^\mu A_{\mu i}
\end{aligned} \quad [76]$$

for this gauge choice the gauge-fixed action is

$$S_\Psi = \int_M \mu\Big(\epsilon^{\mu\nu}(A_{\mu i}\partial_\nu X^i + P^{ij}(X)A_{\mu i}A_{\nu j}) + \bar{C}^i\partial^\mu D^j_{\mu i}C_j + \frac{1}{4}(\partial^\mu \bar{C}^i)(\partial^\nu \bar{C}^j)\epsilon_{\mu\nu}P^{kl}{}_{,ij}(X) \times C_k C_l - \bar{\pi}^i(\partial^\mu A_{\mu i})\Big) \quad [77]$$

Translating this action into the notation of Cattaneo and Felder, one sees that it is exactly the expression they use to derive the perturbation series.

2. Now consider the temporal gauge $\chi_i = A_{0i}$. The gauge fermion is given by $\Psi = \bar{C}^i A_{0i}$. The antifields are fixed to

$$\begin{aligned} A^{*0i} &= \bar{C}^i \\ A^{*1i} &= 0 \\ X^*_i &= C^{*i} = 0 \\ \bar{C}^*_i &= A_{0i} \end{aligned} \quad [78]$$

The gauge-fixed action is

$$S_\Psi = \int_M \mu\Big(\epsilon^{\mu\nu}(A_{\mu i}\partial_\nu X^i + P^{ij}(X)A_{\mu i}A_{\nu j}) + \bar{C}^i D^j_{0i}C_j - \bar{\pi}^i(A_{0i})\Big) \quad [79]$$

3. Finally consider the Schwinger–Fock gauge $\chi_i = x^\mu A_{\mu i}$. Then the antifields are fixed to be

$$\begin{aligned} A^{*\mu i} &= x^\mu \bar{C}^i \\ X^*_i &= C^{*i} = 0 \\ \bar{C}^*_i &= x^\mu A_{\mu i} \end{aligned} \quad [80]$$

for this gauge choice the gauge-fixed action is

$$S_\Psi = \int_M \mu\Big(\epsilon^{\mu\nu}(A_{\mu i}\partial_\nu X^i + P^{ij}(X)A_{\mu i}A_{\nu j}) + \bar{C}^i x^\mu D^j_{\mu i}C_j - \bar{\pi}^i(\partial^\mu A_{\mu i})\Big) \quad [81]$$

Notice that in the noncovariant gauges 2 and 3 the action simplifies, in that the term which arose because of the nonclosed nature of the gauge algebra vanishes.

See also: BF Theories; BRST Quantization; Constrained Systems; Graded Poisson Algebras; Operads; Perturbative Renormalization Theory and BRST; Supermanifolds; Topological Sigma Models.

Further Reading

Alexandrov M, Kontsevich M, Schwarz A, and Zaboronsky O (1997) Geometry of the Master Equation. *International Journal of Modern Physics* A12: 1405–1430.

Axelrod S and Singer IM (1991) *Chern–Simons Perturbation Theory*, Proceedings of the XXth Conference on Differential Geometric Methods in Physics, Baruch College/CUNY, NY. (hep-th/9110056).

Batalin IA and Vilkovisky GA (1977) Gauge algebra and quantization. *Physics Letters* 69B: 309–312.

Becchi C, Rouet A, and Stora R (1976) Renormalization of gauge theories. *Annals of Physics (NY)* 98: 287–321.

Cattaneo AS and Felder G (2001) On the AKSZ formulation of the Poisson–Sigma model. *Letters of Mathematical Physics* 56: 163–179.

Faddeev LD and Popov VN (1967) Feynman diagrams for the Yang–Mills field. *Physics Letters* 25B: 29–30.

Gomis J, Paris J, and Samuel S (1994) Antibracket Antifields and gauge-theory quantization. *Physics Reports* 269: 1–145.

Hirshfeld AC and Schwarzweller T (2000) Path integral quantization of the Poisson–Sigma model. *Annals of Physics (Leipzig)* 9: 83–101.

Schaller P and Strobl T (1994) Poisson structure induced (topological) field theories. *Modern Physics Letters* A9: 3129–3136.

Witten E (1988) Topological sigma models. *Communications in Mathematical Physics* 118: 411–449.

Zinn-Justin J (1975) Renormalization of gauge theories. In: Rollnik H and Dietz K (eds.) *Trends in Elementary Particle Physics*, Lecture Notes in Physics, vol. 37. Berlin: Springer.

Bosons and Fermions in External Fields

E Langmann, KTH Physics, Stockholm, Sweden

© 2006 Elsevier Ltd. All rights reserved.

Introduction

In this article we discuss quantum theories which describe systems of nondistinguishable particles interacting with external fields. Such models are of interest also in the nonrelativistic case (in quantum statistical mechanics, nuclear physics, etc.), but the relativistic case has additional, interesting complications: relativistic models are genuine quantum field theories, that is, quantum theories with an infinite number of degrees of freedom, with nontrivial features like divergences and anomalies. Since interparticle interactions are ignored, such models can be regarded as a first approximation to more complicated theories, and they can be studied by mathematically precise methods.

Models of relativistic particles in external electromagnetic fields have received considerable attention in the physics literature, and interesting phenomena like the Klein paradox or particle–antiparticle pair creation in overcritical fields have been studied; see Rafelski *et al.* (1978) for an extensive review. We will not discuss these physics questions but only describe some prototype examples and a general Hamiltonian framework which has been used in mathematically precise work on such models. The general framework for this latter work is the mathematical theory of Hilbert space operators (see, e.g., Reed and Simon (1975)), but in our discussion we try to avoid presupposing knowledge of that theory. As mentioned briefly in the end, this work has had close relations to various topics of recent interest in mathematical physics, including anomalies, infinite-dimensional geometry and group theory, conformal field theory, and noncommutative geometry.

We restrict our discussion to spin-0 bosons and spin-1/2 fermions, and we will not discuss models of particles in external gravitational fields but only refer the interested reader to DeWitt (2003). We also only mention in passing that external field problems have also been studied using functional integral approaches, and mathematically precise work on this can be found in the extensive literature on determinants of differential operators.

Examples

Consider the Schrödinger equation describing a nonrelativistic particle of mass m and charge e

Notes:

moving in three-dimensional space and interacting with an external vector and scalar potentials A and ϕ, respectively,

$$i\partial_t\psi = H\psi, \quad H = \frac{1}{2m}(-i\nabla + eA)^2 - e\phi \qquad [1]$$

(we set $\hbar = c = 1$, $\partial_t = \partial/\partial t$, and ψ, ϕ, and A can depend on the space and time variables $x \in \mathbb{R}^3$ and $t \in \mathbb{R}$). This is a standard quantum-mechanical model, with ψ the one-particle wave function allowing for the usual probabilistic interpretation.

One interesting generalization to the relativistic regime is the Klein–Gordon equation

$$\left[(i\partial_t + e\phi)^2 - (-i\nabla + eA)^2 - m^2\right]\psi = 0 \qquad [2]$$

with a C-valued function ψ. There is another important relativistic generalization, the Dirac equation

$$[(i\partial_t + e\phi) - (-i\nabla + eA)\cdot\alpha + m\beta]\psi = 0 \qquad [3]$$

with $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ and β Hermitian 4×4 matrices satisfying the relations

$$\alpha_i\alpha_j + \alpha_j\alpha_i = \delta_{ij}, \quad \alpha_i\beta = -\beta\alpha_i, \quad \beta^2 = 1 \qquad [4]$$

and a $\mathbb{C}^4$-valued function ψ (we also write 1 for the identity). These two relativistic equations differ by the transformation properties of ψ under Lorentz transformations: in [2] it transforms like a scalar and thus describes spin-0 particles, and it transforms like a spinor describing spin-1/2 particles in [3]. While these equations are natural relativistic generalizations of the Schrödinger equation, they no longer allow to consistently interpret ψ as one-particle wave functions. The physical reason is that, in a relativistic theory, high-energy processes can create particle–antiparticle pairs, and this makes the restriction to a fixed particle number inconsistent. This problem can be remedied by constructing a many-body model allowing for an arbitrary number of particles and antiparticles. The requirement that this many-body model should have a ground state is an important ingredient in this construction.

It is obviously of interest to formulate and study many-body models of nondistinguishable particles already in the nonrelativistic case. An important empirical fact is that such particles come in two kinds, bosons and fermions, distinguished by their exchange statistics (we ignore the interesting possibility of exotic statistics). For example, the fermion many-particle version of [1] for suitable ϕ and A is a useful model for electrons in a metal. An elegant method to go from the one- to the many-particle description is the formalism of second quantization: one promotes ψ to a quantum field operator with certain (anti-) commutator relations, and this is a convenient way to construct the appropriate many-particle Hilbert space, Hamiltonian, etc. In the nonrelativistic case, this formalism can be regarded as an elegant reformulation of a pedestrian construction of a many-body quantum-mechanical model, which is useful since it provides convenient computational tools. However, this formalism naturally generalizes to the relativistic case where the one-particle model no longer has an acceptable physical interpretation, and one finds that one can nevertheless give a consistent physical interpretation to [2] and [3] provided that ψ are interpreted as quantum field operators describing bosons and fermions. This particular exchange statistics of the relativistic particles is a special case of the spin-statistics theorem: *integer-spin particles are bosons and half-integer spin particles are fermions*. While many structural features of this formalism are present already in the simpler nonrelativistic models, the relativistic models add some nontrivial features typical for quantum field theories.

In the following, we discuss a precise mathematical formulation of the quantum field theory models described above. We emphasize the functorial nature of this construction, which makes manifest that it also applies to other situations, for example, where the bosons and fermions are also coupled to a gravitational background, are considered in other spacetime dimensions than $3 + 1$, etc.

Second Quantization: Nonrelativistic Case

Consider a quantum system of nondistinguishable particles where the quantum-mechanical description of one such particle is known. In general, this one-particle description is given by a Hilbert space h and one-particle observables and transformations which are self-adjoint and unitary operators on h, respectively. The most important observable is the Hamiltonian H. We will describe a general construction of the corresponding many-body system.

Example As a motivating example we take the Hilbert space $h = L^2(\mathbb{R}^3)$ of square-integrable functions $f(x)$, $x \in \mathbb{R}^3$, and the Hamiltonian H in [1]. A specific example for a unitary operator on h is the gauge transformation $(Uf)(x) = \exp(i\chi(x))f(x)$ with χ a smooth, real-valued function on $\mathbb{R}^3$.

In this example, the corresponding wave functions for N identical such particles are the L^2-functions $f_N(x_1, \ldots, x_N)$, $x_j \in \mathbb{R}^3$. It is obvious how to extend

one-particle observables and transformations to such N-particle states: for example, the N-particle Hamiltonian corresponding to H in [1] is

$$H_N = \sum_{j=1}^{N} \frac{1}{2m}(-i\nabla_{x_j} + eA(t,x_j))^2 - e\phi(t,x_j) \quad [5]$$

and the N-particle gauge transformation U_N is defined through multiplication with $\prod_{j=1}^{N} \exp(i\chi(x_j))$.

For systems of indistinguishable particles it is enough to restrict to wave functions which are even or odd under particle exchanges,

$$f_N(x_1,\ldots,x_j,\ldots,x_k,\ldots,x_N) = \pm f_N(x_1,\ldots,x_k,\ldots,x_j,\ldots,x_N) \quad [6]$$

for all $1 \le j < k \le N$, with the upper and lower signs corresponding to bosons and fermions, respectively (this empirical fact is usually taken as a postulate in nonrelativistic many-body quantum physics). It is convenient to define the zero-particle Hilbert space as $\mathbb{C}$ (complex numbers) and to introduce a Hilbert space containing states with all possible particle numbers: this so-called Fock space contains all states

$$\begin{pmatrix} f_0 \\ f_1(x_1) \\ f_2(x_1,x_2) \\ f_3(x_1,x_2,x_3) \\ \vdots \end{pmatrix} \quad [7]$$

with $f_0 \in \mathbb{C}$. The definition of H_N and U_N then naturally extends to this Fock space; see below.

General Construction

The construction of Fock spaces and many-particle observables and transformations just outlined in a specific example is conceptually simple. An alternative, more efficient construction method is to use "quantum fields," which we denote as $\psi(x)$ and $\psi^\dagger(x)$, $x \in \mathbb{R}^3$. They can be fully characterized by the following (anti-) commutator relations:

$$[\psi(x),\psi^\dagger(y)]_\mp = \delta^3(x-y), \quad [\psi(x),\psi(y)]_\mp = 0 \quad [8]$$

where $[a,b]_\mp \equiv ab \mp ba$, with the commutator and anticommutators (upper and lower signs, respectively) corresponding to the boson and fermion case, respectively. It is convenient to "smear" these fields with one-particle wave functions and define

$$\psi(f) = \int_{\mathbb{R}^3} d^3x \overline{f(x)}\psi(x)$$
$$\psi^\dagger(f) = \int_{\mathbb{R}^3} d^3x \psi^\dagger(x) f(x) \quad [9]$$

for all $f \in h$. Then the relations characterizing the field operators can be written as

$$[\psi(f),\psi^\dagger(g)]_\mp = (f,g)$$
$$[\psi(f),\psi(g)]_\mp = 0 \quad \forall f,g \in h \quad [10]$$

where

$$(f,g) = \int_{\mathbb{R}^3} d^3x \overline{f(x)} g(x)$$

is the inner product in h. The Fock space $\mathcal{F}_\mp(h)$ can then be defined by postulating that it contains a normalized vector Ω called "vacuum" such that

$$\psi(f)\Omega = 0 \quad \forall f \in h \quad [11]$$

and that all $\psi^{(\dagger)}(f)$ are operators on $\mathcal{F}_\mp(h)$ such that $\psi^\dagger(f) = \psi(f)^*$, where $*$ is the Hilbert space adjoint. Indeed, from this we conclude that $\mathcal{F}_\mp(h)$, as vector space, is generated by

$$f_1 \wedge f_2 \wedge \cdots \wedge f_N \equiv \psi^\dagger(f_1)\psi^\dagger(f_2)\cdots\psi^\dagger(f_N)\Omega \quad [12]$$

with $f_j \in h$ and $N = 0,1,2,\ldots$, and that the Hilbert space inner product of such vectors is

$$\langle f_1 \wedge f_2 \wedge \cdots \wedge f_N, g_1 \wedge g_2 \wedge \cdots \wedge g_M\rangle = \delta_{N,M} \sum_{P\in S_N} (\pm 1)^{|P|} \prod_{j=1}^{N} (f_j, g_{Pj}) \quad [13]$$

with S_N the permutation group, with $(+1)^{|P|} = 1$ always, and $(-1)^{|P|} = +1$ and -1 for even and odd permutations, respectively. The many-body Hamiltonian $q(H)$ corresponding to the one-particle Hamiltonian H can now be defined by the following relations:

$$q(H)\Omega = 0, \quad [q(H),\psi^\dagger(f)] = \psi^\dagger(Hf) \quad [14]$$

for all $f \in h$ such that Hf is defined. Indeed, this implies that

$$q(H) f_1 \wedge f_2 \wedge \cdots \wedge f_N = \sum_{j=1}^{N} f_1 \wedge f_2 \wedge \cdots \wedge (Hf_j) \wedge \cdots \wedge f_N \quad [15]$$

which defines a self-adjoint operator on $\mathcal{F}_\mp(h)$, and it is easy to check that this coincides with our down-to-earth definition of H_N above. Similarly, the many-body transformation $Q(U)$ corresponding to a one-particle transformation U can be defined as

$$Q(U)\Omega = \Omega, \quad Q(U)\psi^\dagger(f) = \psi^\dagger(Uf)Q(U) \quad [16]$$

for all $f \in h$, which implies that

$$Q(U) f_1 \wedge f_2 \wedge \cdots \wedge f_N = (Uf_1) \wedge (Uf_2) \wedge \cdots \wedge (Uf_N) \quad [17]$$

and thus coincides with our previous definition of U_N.

While we presented the construction above for a particular example, it is important to note that it actually does not make reference to what the one-particle formalism actually is. For example, if we had a model of particles on a space $\mathcal{M}$ given by some "nice" manifold of any dimension and with M internal degrees of freedom, we would take $h = L^2(\mathcal{M}) \otimes \mathbb{C}^M$ and replace [9] by

$$\psi(f) = \int_{\mathcal{M}} \mathrm{d}\mu(\boldsymbol{x}) \sum_{j=1}^{M} \overline{f_j(\boldsymbol{x})} \psi_j(\boldsymbol{x}) \qquad [18]$$

and its Hermitian conjugate, with the measure μ on $\mathcal{M}$ defining the inner product in h,

$$(f, g) = \int \mathrm{d}\mu(\boldsymbol{x}) \sum_j \overline{f_j(\boldsymbol{x})} g_j(\boldsymbol{x})$$

With that, all formulas after [9] hold true as they stand. *Given any one-particle Hilbert space h with inner product $(\cdot\,,\cdot)$, observable H, and transformation U, the formulas above define the corresponding Fock spaces $\mathcal{F}_{\mp}(h)$ and many-body observable $q(H)$ and transformation $Q(U)$.* It is also interesting to note that this construction has various beautiful general (functorial) properties: the set of one-particle observables has a natural Lie algebra structure with the Lie bracket given by the commutator (strictly speaking: i times the commutator, but we drop the common factor i for simplicity). The definitions above imply that

$$[q(A), q(B)] = q([A, B]) \qquad [19]$$

for one-particle observables A, B, that is, the above-mentioned Lie algebra structure is preserved under this map q. In a similar manner, the set of one-particle transformations has a natural group structure preserved by the map Q,

$$Q(U)Q(V) = Q(UV), \quad Q(U)^{-1} = Q(U^{-1}) \qquad [20]$$

Moreover, if A is self-adjoint, then $\exp(\mathrm{i}A)$ is unitary, and one can show that

$$Q(\exp(\mathrm{i}A)) = \exp(\mathrm{i}q(A)) \qquad [21]$$

For later use, we note that, if $\{f_n\}_{n\in\mathbb{Z}}$ is some complete, orthonormal basis in h, then operators A on h can be represented by infinite matrices $(A_{mn})_{m,n\in\mathbb{Z}}$ with $A_{mn} = (f_m, Af_n)$, and

$$q(A) = \sum_{m,n} A_{mn} \psi_m^\dagger \psi_n \qquad [22]$$

where $\psi_n^{(\dagger)} = \psi^{(\dagger)}(f_n)$ obey

$$[\psi_m, \psi_n^\dagger]_{\mp} = \delta_{m,n}, \qquad [\psi_m, \psi_n^\dagger]_{\mp} = 0 \qquad [23]$$

for all m, n. We also note that, in our definition of $q(A)$, we made a convenient choice of normalization, but there is no physical reason to not choose a different normalization and define

$$q'(A) = q(A) - b(A) \qquad [24]$$

where b is some linear function mapping self-adjoint operators A to real numbers. For example, one may wish to use another reference vector $\tilde{\Omega}$ instead of Ω in the Fock space, and then would choose $b(A) = \langle \tilde{\Omega}, q(A)\tilde{\Omega} \rangle$. Then the relations in [19] are changed to

$$[q'(A), q'(B)] = q'([A, B]) + S_0(A, B) \qquad [25]$$

where $S_0(A, B) = b([A, B])$. However, the C-number term $S_0(A, B)$ in the relations [25] is trivial, since it can be removed by going back to $q(A)$.

Physical Interpretation

The Fock space $\mathcal{F}_{\mp}(h)$ is the direct sum of subspaces of states with different particle numbers N,

$$\mathcal{F}_{\mp}(h) = \bigoplus_{N=0}^{\infty} h_{\mp}^{(N)} \qquad [26]$$

where the zero-particle subspace $h_{\mp}^{(0)} = \mathbb{C}$ is generated by the vacuum Ω, and $h_{\mp}^{(N)}$ is the N-particle subspace generated by the states $f_1 \wedge f_2 \wedge \cdots \wedge f_N$, $f_j \in h$. We note that

$$\mathcal{N} \equiv q(1) \qquad [27]$$

is the "particle-number operator," $\mathcal{N}F_N = NF_N$ for all $F_N \in h_{\mp}^{(N)}$. The field operators obviously change the particle number: $\psi^\dagger(f)$ increases the particle number by one (maps $h_{\mp}^{(N)}$ to $h_{\mp}^{(N+1)}$), and $\psi(f)$ decreases it by one. Since every $f \in h$ can be interpreted as one-particle state, it is natural to interpret $\psi^\dagger(f)$ and $\psi(f)$ as "creation" and "annihilation" operators, respectively: they create and annihilate one particle in the state $f \in h$. It is important to note that, in the fermion case, [10] implies that $\psi^\dagger(f)^2 = 0$, which is a mathematical formulation of the Pauli exclusion principle: *it is not possible to have two fermions in the same one-particle state.* In the boson case, there is no such restriction. Thus, even though the formalisms used to describe boson and fermion systems look very similar, they describe dramatically different physics.

Applications

In our example, the many-body Hamiltonian $\mathcal{H}_0 \equiv q(H)$ can also be written in the following suggestive form:

$$\mathcal{H}_0 = \int \mathrm{d}^3x\, \psi^\dagger(\boldsymbol{x})(H\psi)(\boldsymbol{x}) \qquad [28]$$

and similar formulas hold true for other observables and other Hilbert spaces $h = L^2(\mathcal{M}) \otimes \mathbb{C}^n$. It is rather easy to solve the model defined by such Hamiltonian: all necessary computations can be reduced to one-particle computations. For example, in the static case, where A and ϕ are time independent, a main quantity of interest in statistical physics is the free energy

$$\mathcal{E} \equiv -\beta^{-1} \log(\mathrm{tr}(\exp(-\beta[\mathcal{H}_0 - \mu\mathcal{N}]))) \qquad [29]$$

where $\beta > 0$ is the inverse temperature, μ the chemical potential, and the trace over the Fock space $\mathcal{F}_\mp(h)$. One can show that

$$\mathcal{E} = \pm \mathrm{tr}\left(\beta^{-1} \log(1 \mp \exp(-\beta[H - \mu]))\right) \qquad [30]$$

where the trace is over the one-particle Hilbert space h. Thus, to compute $\mathcal{E}$, one only needs to find the eigenvalues of H.

It is important to mention that the framework discussed here is not only for external field problems but can be equally well used to formulate and study more complicated models with interparticle interactions. For example, while the model with the Hamiltonian $\mathcal{H}_0$ above is often too simple to describe systems in nature, it is easy to write down more realistic models, for example, the Hamiltonian

$$\mathcal{H} = \mathcal{H}_0 + (e^2/2) \int \mathrm{d}^3x \int \mathrm{d}^3y\, \psi^\dagger(\boldsymbol{x})\psi^\dagger(\boldsymbol{y}) \times |\boldsymbol{x} - \boldsymbol{y}|^{-1} \psi(\boldsymbol{y})\psi(\boldsymbol{x}) \qquad [31]$$

describes electrons in an external electromagnetic field interacting through Coulomb interactions. This illustrates an important point which we would like to stress: the task in quantum theory is twofold, namely to formulate and to solve (exact of otherwise) models. Obviously, in the nonrelativistic case, it is equally simple to formulate many-body models with and without interparticle interactions, and only the latter are simpler because they are easier to solve: the two tasks of formulating and solving models can be clearly separated. As we will see, in the relativistic case, even the formulation of an external field problem is nontrivial, and one finds that one cannot formulate the model without at least partially solving it. This is a common feature of quantum field theories making them challenging and interesting.

Relativistic Fermion and Boson Systems

We now generalize the formalism developed in the previous section to the relativistic case.

Field Algebras and Quasifree Representations

In the previous section, we identified the field operators $\psi^{(\dagger)}(f)$ with particular Fock space operators. This is analogous to identifying the operators $p_j = -\mathrm{i}\partial_{x_j}$ and $q_j = x_j$ on $L^2(\mathbb{R}^M)$ with the generators of the Heisenberg algebra, as usually done. (We recall: the Heisenberg algebra is the star algebra generated by P_j and Q_j, $j = 1, 2, \ldots, M < \infty$, with the well-known relations

$$[P_j, P_k] = -\mathrm{i}\delta_{jk}, \qquad [P_j, P_k] = [P_j, Q_k] = 0$$
$$P_j^\dagger = P_j, \qquad Q_j^\dagger = Q_j \qquad [32]$$

for all j, k.) Identifying the Heisenberg algebra with a particular representation is legitimate since, as is well known, all its irreducible representations are (essentially) the same (this statement is made precise by a celebrated theorem due to von Neumann).

However, in case of the algebra generated by the field operators $\psi^{(\dagger)}(f)$, there exist representations which are truly different from the ones discussed in the last section, and such representations are needed to construct relativistic external field problems. It is therefore important to distinguish the fields as generators of an algebra from the operators representing them. We thus define the (boson or fermion) field algebra $\mathcal{A}_\mp(h)$ over a Hilbert space h as the star algebra generated by $\Psi^\dagger(f), f \in h$, such that the map $f \to \Psi(f)$ is linear and the relations

$$[\Psi(f), \Psi^\dagger(g)]_\mp = (f, g)$$
$$[\Psi(f), \Psi(g)]_\mp = 0 \qquad [33]$$
$$\Psi^\dagger(f)^\dagger = \Psi(f)$$

are fulfilled for all $f, g \in h$, with $\dagger$ the star operation in $\mathcal{A}_\mp(h)$. The particular representation of this algebra discussed in the last section will be denoted by π_0, $\pi_0(\Psi^{(\dagger)}(f)) = \psi^{(\dagger)}(f)$. Other representations π_P can be constructed from any projection operators P_- on h, that is, any operator P_- on h satisfying $P_-^* = P_-^2 = P_-$. Writing $\hat{\psi}^{(\dagger)}(f)$ short for $\pi_{P_-}(\Psi^{(\dagger)}(f))$, this so-called quasifree representation is defined by

$$\hat{\psi}^\dagger(f) = \psi^\dagger(P_+ f) + \psi(\overline{P_- f})$$
$$\hat{\psi}(f) = \psi(P_+ f) \mp \psi^\dagger(\overline{P_- f}) \qquad [34]$$

where the bar means complex conjugation. It is important to note that, while the star operation is identical with the Hilbert space adjoint $*$ in the fermion case, we have

$$\hat{\psi}(f)^\dagger = \psi(Ff)^* \quad \text{with}$$
$$F = P_+ - P_- \quad \text{for bosons} \qquad [35]$$

where F is a grading operator, that is, $F^* = F$ and $F^2 = 1$. We stress that the "physical" star operation always is $*$, that is, physical observables A obey $A = A^*$.

The present framework suggests to regard quantization as the procedure which amounts to going from a one-particle Hilbert space h to the corresponding field algebra $\mathcal{A}_+(h)$. Indeed, the Heisenberg algebra is identical with the boson field algebra $\mathcal{A}_-(\mathbb{C}^M)$ (since the latter is obviously identical with the algebra of M harmonic oscillators), and thus conventional quantum mechanics can be regarded as boson quantization in the special case where the one-particle Hilbert space is finite dimensional. It is interesting to note that "fermion quantum mechanics" $\mathcal{A}_-(\mathbb{C}^M)$ is the natural framework for formulating and studying lattice fermion and spin systems which play an important role in condensed matter physics.

In the following, we elaborate the naive interpretations of the relativistic equations in [2] and [3] as a quantum theory of one particle, and we discuss why they are unphysical. For simplicity, we assume that the electromagnetic fields ϕ, $\boldsymbol{A}$ are time independent. We then show that quasifree representations as discussed above can provide physically acceptable many-particle theories. We first consider the Dirac case, which is somewhat simpler.

Fermions

One-particle formalism Recalling that $i\partial_t$ is the energy operator, we define the Dirac Hamiltonian D by rewriting [3] in the following form:

$$i\partial_t\psi = D\psi, \quad D = (-i\nabla + e\boldsymbol{A})\cdot\boldsymbol{\alpha} + m\beta - e\phi \qquad [36]$$

This Dirac Hamiltonian is obviously a self-adjoint operator on the one-particle Hilbert space $h = L^2(\mathbb{R}^4)\otimes\mathbb{C}^4$, but, different from the Schrödinger Hamiltonian in [1], it is not bounded from below: for any $E_0 > -\infty$, one can find a state f such that the energy expectation value (f, Df) is less than E_0. This can be easily seen for the simplest case where the external potential vanishes, $\boldsymbol{A} = \phi = 0$. Then the eigenvalues of D can be computed by Fourier transformation, and one finds

$$E = \pm\sqrt{\boldsymbol{p}^2 + m^2}, \quad \boldsymbol{p}\in\mathbb{R}^3 \qquad [37]$$

Due to the negative energy eigenvalues we conclude that there is no ground state, and the Dirac Hamiltonian thus describes an unstable system, which is physically meaningless.

To summarize: a (unphysical) one-particle description of relativistic fermions is given by a Hilbert space h together with a self-adjoint Hamiltonian D unbounded from below. Other observables and transformations are given by self-adjoint and unitary operators on h, respectively.

Many-body formalism We now explain how to construct a physical many-body description from these data. To simplify notation, we first assume that D has a purely discrete spectrum (which can be achieved by using a compact space). We can then label the eigenfunctions f_n by integers n such that the corresponding eigenvalues $E_n \geq 0$ for $n \geq 0$ and $E_n < 0$ for $n < 0$. Using the naive representation of the fermion field algebra discussed in the last section, we get (we use the notation introduced in [22])

$$q(D) = \sum_{n\geq 0} |E_n|\psi_n^\dagger\psi_n - \sum_{n<0} |E_n|\psi_n^\dagger\psi_n \qquad [38]$$

which is obviously not bounded from below and thus not physically meaningful. However, $\psi_n^\dagger\psi_n = 1 - \psi_n\psi_n^\dagger$, which suggests that we can remedy this problem by interchanging the creation and annihilation operators for $n < 0$. This is possible: it is easy to see that

$$\hat{\psi}_n \equiv \psi_n \quad \forall n \geq 0 \quad \text{and} \quad \hat{\psi}_n \equiv \psi_n^\dagger \quad \forall n < 0 \qquad [39]$$

provides a representation of the algebra in [23]. We thus define

$$\hat{q}(D) \equiv \sum_{n\in\mathbb{Z}} E_n : \hat{\psi}_n^\dagger\hat{\psi}_n : \qquad [40]$$

with the so-called normal ordering prescription

$$:\psi_m^\dagger\psi_n: \equiv \psi_m^\dagger\psi_n - \langle\Omega, \psi_m^\dagger\psi_n\Omega\rangle \qquad [41]$$

where we made use of the freedom of normalization explained after [23] to eliminate unwanted additive constants. We get $q(D) = \sum_{n\in\mathbb{Z}} |E_n|\psi_n^\dagger\psi_n$, which is manifestly a non-negative self-adjoint operator with Ω as ground state. We thus found a physical many-body description for our model. We can now define for other one-particle observables,

$$\hat{q}(A) \equiv \sum_{n\in\mathbb{Z}} A_{mn} : \hat{\psi}_m^\dagger\hat{\psi}_n : \qquad [42]$$

and, by straightforward computations, we obtain

$$[\hat{q}(A), \hat{q}(B)] = \hat{q}([A, B]) + S(A, B) \qquad [43]$$

where $S(A, B) = \sum_{m<0}\sum_{n\geq 0}(A_{mn}B_{nm} - B_{mn}A_{nm})$, that is,

$$S(A, B) = \mathrm{tr}(P_-AP_+BP_- - P_-BP_+AP_-) \qquad [44]$$

with $P_- = \sum_{n<0} f_n(f_n, \cdot)$ the projection onto the subspace spanned by the negative energy eigenvectors of D and $P_+ = 1 - P_-$. One can show that $\hat{q}(A)$ is no longer defined for all operators but only if

$$P_-AP_+ \text{ and } P_+AP_- \text{ are Hilbert–Schmidt operators} \qquad [45]$$

(we recall that a is a Hilbert–Schmidt operator if $\mathrm{tr}(a^*a) < \infty$). The C-number term $S(A,B)$ in [43] is

often called Schwinger term and, different from the similar term in [25], it is now nontrivial, that is, it is no longer possible to remove it by a redefinition $\hat{q}'(A) = \hat{q}(A) - b(A)$. This Schwinger term is an example of an anomaly, and it has various interesting implications.

In a similar manner, one can construct the many-body transformations $\hat{Q}(U)$ of unitary operators U on h satisfying the very Hilbert–Schmidt condition in [45], and one obtains

$$\hat{Q}(U)\hat{Q}(V) = \chi(U,V)\hat{Q}(UV) \qquad [46]$$

with interesting phase-valued functions χ.

More generally, for any one-particle Hilbert space h and Dirac Hamiltonian D, the physical representation is given by the quasifree representation π_{P_-} in [34] with P_- the projection onto the negative energy subspace of D. The results about $\hat{q}$ and $\hat{Q}$ mentioned hold true in any such representation.

Thus the one-particle Hamiltonian D determines which representation one has to use, and one therefore cannot construct the "physical" representation without specific information about D. However, not all these representations are truly different: *if there is a unitary operator $\mathcal{U}$ on the Fock space $\mathcal{F}_+(h)$ such that*

$$\mathcal{U}^* \pi_{P_-^{(1)}}(\psi^{(\dagger)}(f))\mathcal{U} = \pi_{P_-^{(2)}}(\psi^{(\dagger)}(f)) \qquad [47]$$

for all $f \in h$, then the quasifree representations associated with the different projections $P_-^{(1)}$ and $P_-^{(2)}$ are physically equivalent: one could equally well formulate the second model using the representation of the first. Two such quasifree representations are called unitarily equivalent, and a fundamental theorem due to Shale and Stinespring states that *two quasifree representations $\pi_{P_-^{(1,2)}}$ are unitarily equivalent if and only if $P_-^{(1)} - P_-^{(2)}$ is a Hilbert–Schmidt operator* (a similar result holds true in the boson case).

Bosons

One-particle formalism Similarly as for the Dirac case, the solutions of the Klein–Gordon equation in [2] also do not define a physically acceptable one-particle quantum theory with a ground state: the energy eigenvalues in [37] for $A = \phi = 0$ are a consequence the relativistic invariance and thus equally true for the Klein–Gordon case. However, in this case there is a further problem. To find the one-particle Hamiltonian, one can rewrite the second-order equation in [2] as a system of first-order equations,

$$\mathrm{i}\partial_t \Phi = K\Phi$$
$$\Phi = \begin{pmatrix} \psi \\ \pi^\dagger \end{pmatrix}, \quad K = \begin{pmatrix} C & \mathrm{i} \\ -\mathrm{i}B^2 & C \end{pmatrix} \qquad [48]$$

with

$$B^2 \equiv (-\mathrm{i}\nabla + eA)^2 + m^2, \qquad C \equiv -e\phi \qquad [49]$$

Thus, one sees that the natural one-particle Hilbert space for the Klein–Gordon equation is $h = L^2(\mathbb{R}^3) \otimes \mathbb{C}^2$; here, and in the following, we identify h with $h_0 \oplus h_0, h_0 = L^2(\mathbb{R}^3)$, and use a convenient 2×2 matrix notation naturally associated with that splitting. However, the one-particle Hamiltonian is not self-adjoint but rather obeys

$$K^* = JKJ, \quad J \equiv \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} \qquad [50]$$

with $*$ *the Hilbert space adjoint. It is important to* note that J is a grading operator. Thus, we can define a sesquilinear form

$$(f,g)_J \equiv (f,Jg) \quad \forall f,g \in h \qquad [51]$$

with $(\cdot,\cdot)$ the standard inner product, and [50] is equivalent to K being self-adjoint with respect to this sesquilinear form; in this case, we say that K is J-self-adjoint. Thus, in the Klein–Gordon case, this sesquilinear form takes the role of the Hilbert space inner product and, in particular, not (Φ,Φ) but $(\Phi,\Phi)_J$ is preserved under time evolution. However, different from $\Phi^\dagger\Phi$, $\Phi^\dagger J\Phi$ is not positive definite, and it is therefore not possible to interpret it as probability density as in conventional quantum mechanics. For consistency, one has to require that one-particle transformations U are unitary with respect to $(\Phi,\Phi)_J$, that is, $U^{-1} = JUJ$. We call such operators J-unitary.

To summarize: a (unphysical) one-particle description of relativistic bosons is given by a Hilbert space of the form $h = h_0 \oplus h_0$, the grading operator J in [50], and a J-self-adjoint Hamiltonian K of the form as in eqn [48], where $B \geq 0$ and C are self-adjoint operators on h_0. Other observables and transformations are given by J-self-adjoint and J-unitary operators on h, respectively.

Many-body formalism We first consider the quasifree representation $\pi_{P_-^{(0)}}$ of the boson field algebra $\mathcal{A}_-(h)$ so that the grading operator in [35] is equal to J, that is, $P_-^{(0)} = (1-J)/2$. Writing $\pi_{P_-^{(0)}}(\Psi^{(\dagger)}(f)) = \psi^{(\dagger)}(f)$, one finds that

$$q(A)^* = q(JAJ), \qquad Q(U)^* = Q(JU^*J) \qquad [52]$$

and thus J-self-adjoint operators and J-unitary operators are mapped to proper observables and transformations. In particular, $q(K)$ is a self-adjoint

operator, which resolves one problem of the one-particle theory. However, $q(K)$ is not bounded from below, and thus $\pi_{P(0)}$ is not yet the physical representation.

The physical representation can be constructed using the operators

$$T = \frac{1}{\sqrt{2}} \begin{pmatrix} B^{1/2} & iB^{-1/2} \\ B^{1/2} & iB^{-1/2} \end{pmatrix}, \quad F = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad [53]$$

(for simplicity, we restrict ourselves to the case $C = 0$ and $B > 0$; we use the calculus of self-adjoint operators here) with the following remarkable properties:

$$T^{-1} = JT^*F$$
$$TKT^{-1} = \begin{pmatrix} B & 0 \\ 0 & -B \end{pmatrix} \equiv \hat{K} \qquad [54]$$

One can check that

$$\hat{\psi}^\dagger(f) \equiv \psi^\dagger(Tf), \qquad \hat{\psi}(f) \equiv \psi(T^{-1}f) \qquad [55]$$

is a quasifree representation π_{P_-} of $\mathcal{A}_-(h)$ with $P_- = (1 - F)/2$. With that the construction of $\hat{q}$ and $\hat{Q}$ is very similar to the fermion case described above (the crucial simplification is that $\hat{K}$ and F now are diagonal). In particular, $\hat{q}(K)$ is a non-negative operator with the ground state Ω, and $\hat{q}(A)$ and $\hat{Q}(U)$ are self-adjoint and unitary for every one-particle observable A and transformation U, respectively. One also gets relations as in [43] and [46].

Related Topics of Recent Interest

The impossibility to construct relativistic quantum-mechanical models played an important role in the early history of quantum field theory, as beautifully discussed in chapter 1 of Weinberg (1995).

The abstract formalism of quasifree representations of fermion and boson field algebras was developed in many papers (see, e.g., Ruijsenaars (1977), Grosse and Langmann (1992), and Langmann (1994) for explicit results on $\hat{Q}$ and χ). A nice textbook presentation with many references can be found in chapter 13 of Gracia-Bondía *et al.* (2001) (this chapter is rather self-contained but mainly restricted to the fermion case).

Based on the Shale–Stinespring theorem, there has been considerable amount of work to investigate whether the quasifree representations associated with different external electromagnetic fields ψ_1, A_1 and ψ_2, A_2 are unitarily equivalent, if and which time-dependent many-body Hamiltonians exist, etc. (see chapter 13 of Gracia-Bondía *et al.* (2001), and references therein).

The infinite-dimensional Lie algebra g_2 of Hilbert space operators satisfying the condition in [45] is an interesting infinite-dimensional Lie algebra with a beautiful representation theory. This subject is closely related to conformal field theory (see, e.g., Kac and Raina (1987) for a textbook presentation and Carey and Ruijsenaars (1987) for a detailed mathematical account within the framework described by us).

It turns out that the mathematical framework discussed in the previous section is sufficient for constructing fully interacting quantum field theories, in particular Yang–Mills gauge theories, in $1 + 1$ but not in higher dimensions. The reason is that, in $3 + 1$ dimensions, the one-particle observables A of interest do not obey the Hilbert–Schmidt condition in [45] but only the weaker condition

$$\mathrm{tr}(a^*a)^n < \infty, \quad a = P_\mp A P_\pm \qquad [56]$$

with $n = 2$, and the natural analog of g_2 in $3 + 1$ dimensions thus seems to be the Lie algebra g_{2n} of operators satisfying this condition with $n = 2$. Various results on the representation theory of such Lie algebras $g_{2n>2}$ have been developed (see Mickelsson (1989), where various interesting relations to infinite-dimensional geometry are also discussed).

As mentioned, the Schwinger term $S(A,B)$ in [44] is an example of an anomaly. Mathematically, it is a nontrivial 2-cocycle of the Lie algebra g_2, and analogs for the groups $g_{2n>2}$ have been found. These cocycles provide a natural generalization of anomalies (in the meaning of particle physics) to operator algebras. They not only shed some interesting light on the latter, but also provide a link to notions and results from noncommutative geometry (see, e.g., Gracia-Bondía *et al.* (2001)). We believe that this link can provide a fruitful driving force and inspiration to find ways to deepen our understanding of quantum Yang–Mills theories in $3 + 1$ dimensions (Langmann 1996).

See also: Anomalies; C^*-Algebras and Their Classification; Dirac Fields in Gravitation and Nonabelian Gauge Theory; Dirac Operator and Dirac Field; Gerbes in Quantum Field Theory; Quantum Field Theory in Curved Spacetime; Quantum n-Body Problem; Superfluids; Two-Dimensional Models.

Further Reading

Carey AL and Ruijsenaars SNM (1987) On fermion gauge groups, current algebras and Kac–Moody algebras. *Acta Applicandae Mathematicae* 10: 1–86.

DeWitt B (2003) *The Global Approach to Quantum Field Theory*, International Series of Monographs on Physics, vols. 1 and 2, p. 114. New York: Oxford University Press.

Gracia-Bondía JM, Várilly JC, and Figueroa H (2001) *Elements of Noncommutative Geometry*, Birkhäuser Advanced Texts: Basel Textbooks. Boston: Birkhäuser.

Grosse H and Langmann E (1992) A superversion of quasifree second quantization. *Journal of Mathematical Physics* 33: 1032–1046.

Kac VG and Raina AK (1987) *Bombay Lectures on Highest Weight Representations of Infinite-Dimensional Lie Algebras*,

Advanced Series in Mathematical Physics, vol. 2. Teaneck: World Scientific Publishing.

Langmann E (1994) Cocycles for boson and fermion Bogoliubov transformations. *Journal of Mathematical Physics* 96–112.

Langmann E (1996) Quantum gauge theories and noncommutative geometry. *Acta Physica Polonica B* 27: 2477–2496.

Mickelsson J (1989) *Current Algebras and Groups*, Plenum Monographs in Nonlinear Physics. New York: Plenum Press.

Rafelski J, Fulcher LP, and Klein A (1978) Fermions and bosons interacting with arbitrary strong external fields. *Physics Reports* 38: 227–361.

Reed M and Simon B (1975) *Methods of Modern Mathematical Physics. II. Fourier Analysis, Self-Adjointness*. New York: Academic Press.

Ruijsenaars SNM (1977) On Bogoliubov transformations for systems of relativistic charged particles. *Journal of Mathematical Physics* 18: 517–526.

Weinberg S (1995) *The Quantum Theory of Fields*, vol. I (English summary) Foundations. Cambridge: Cambridge University Press.

Constrained Systems

M Henneaux, Université Libre de Bruxelles, Brussels, Belgium

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Consider a dynamical system with coordinates q^i $(i=1,\ldots,n)$ and Lagrangian $L(q^i,\dot{q}^i)$ (field theory is formally covered by regarding the spatial coordinates as a continuous index). When going to the Hamiltonian formulation, it is usually assumed that the Legendre transformation between the velocities $\dot{q}^i$ and the momenta

$$p_i = \frac{\partial L}{\partial \dot{q}^i} \qquad [1]$$

can be inverted to yield the velocities as functions of the q's and the p's. This "regular" situation occurs for most systems appearing in standard classical mechanics and enables one to proceed to the Hamiltonian formulation of the theory without difficulty.

In field theory, however, the regular case is the exception rather than the rule. This is due to gauge invariance and first-order Lagrangians.

- *Gauge invariance* A system possesses gauge symmetries if it is invariant under transformations that involve arbitrary functions of time (gauge transformations). In that case, the solution of the equations of motion with given initial data is not unique, since it is always possible to perform a gauge transformation in the course of the evolution without changing the initial data. It is then clear that the Legendre transformation cannot be invertible, for if it were, one could rewrite the equations of motion in the standard canonical form $\dot{q}^i = \partial H/\partial p_i, \dot{p}_i = -\partial H/\partial q^i$. These canonical equations are in normal form and have a unique solution for given initial data, which would contradict the presence of a gauge symmetry.

  A simple example that illustrates this phenomenon is given by the following model for three variables q^1, q^2, and λ, the Lagrangian of which reads

$$L = \tfrac{1}{2}\left((\dot{q}^1 - \lambda)^2 + (\dot{q}^2 - \lambda)^2\right) \qquad [2]$$

  This model is inspired by electromagnetism: the variables q^1 and q^2 play a role somewhat similar to that of the spatial components of the vector potential, while λ corresponds to the temporal component. The Lagrangian is invariant under the gauge transformations

$$q^1 \to q^1 + \varepsilon, \qquad q^2 \to q^2 + \varepsilon, \qquad \lambda \to \lambda + \dot{\varepsilon} \qquad [3]$$

  where ε is an arbitrary function of time. The conjugate momenta are

$$p_1 = \dot{q}^1 - \lambda, \qquad p_2 = \dot{q}^2 - \lambda, \qquad \pi_\lambda = 0$$

  One cannot invert the Legendre transformation since one cannot express the velocity $\dot{\lambda}$ in terms of the momenta.
- *First-order Lagrangians* Fermionic fields obey first-order equations. Their Lagrangian is linear in the derivatives, so that the conjugate momenta p_i depend on the coordinates q^i only. It is then clearly impossible to express the velocities in terms of the momenta through the Legendre transformation. More generally, any first-order Lagrangian with or without gauge symmetry leads to a noninvertible Legendre transformation.

Notes:

A simple system that exhibits this feature is described by the Lagrangian

$$L = z^2\dot{z}^1 - \tfrac{1}{2}(z^2)^2 \qquad [4]$$

for two bosonic degrees of freedom (z^1, z^2). This is in fact the canonical form of the Lagrangian for a free particle in one dimension (z^2 is the momentum conjugate to the position z^1): the system is already in Hamiltonian form. There is no gauge invariance, but because the Lagrangian is first order, the Legendre transformation with [4] as starting point,

$$p_1 = z^2, \qquad p_2 = 0 \qquad [5]$$

is non invertible for the velocities (which do not even appear in the formulas for the momenta).

Dirac showed how to develop the Hamiltonian formalism in the case when the Legendre transformation is not invertible. One can still reformulate the equations in phase space and write them in terms of brackets with the Hamiltonian, but a new major feature emerges, namely the canonical variables are no longer free. Rather, the permissible phase-space points are constrained to be on the so-called "constrained surface." For this reason, systems for which the Legendre transformation is not invertible are also called "constrained Hamiltonian systems." We shall adopt this terminology here.

The purpose of this article is to explain the main ideas underlying the Dirac method. To simplify the discussions and to focus on the features peculiar to the Dirac construction, we shall assume as a rule that all necessary smoothness conditions are fulfilled by the functions, surfaces, etc., appearing in the formalism. How to develop the analysis when some of the smoothness conditions are not fulfilled is of definite interest but goes beyond the scope of this review. We shall also assume, for definiteness, that all the variables are bosonic in order to avoid straightforward but somewhat cumbersome sign factors in the formulas.

General Theory

Dirac Algorithm

Primary constraints When the Legendre transformation [1] cannot be inverted, the momenta p_i's do not span an n-dimensional space but are constrained by relations

$$\phi_m(q,p) = 0, \quad m = 1, \ldots, M \qquad [6]$$

which follow from their definition. These equations reduce to identities when the momenta are replaced by their expression [1] in terms of the coordinates and the velocities. They are called primary constraints. We shall assume that the matrix

$$\frac{\partial(\phi_m)}{\partial(p_i, q^i)}$$

is everywhere of constant (maximum) rank M on the phase-space surface defined by eqns [6] which is assumed to be smooth. This surface is of dimension $2n - M$.

Canonical Hamiltonian The next step in the Dirac procedure is to define the canonical Hamiltonian H through

$$H = \dot{q}^i p_i - L \qquad [7]$$

As shown by Dirac, H can be re-expressed as a function $H(q, p)$ of the momenta and the coordinates, even when the Legendre transformation is not invertible: the canonical Hamiltonian H depends on the velocities only through the p_i's. Furthermore, the original equations of motion in Lagrangian form are equivalent to the Hamiltonian equations

$$\dot{q}^i = \frac{\partial H}{\partial p_i} + u^m \frac{\partial \phi_m}{\partial p_i} \qquad [8]$$

$$\dot{p}_i = -\frac{\partial H}{\partial q^i} - u^m \frac{\partial \phi_m}{\partial q^i} \qquad [9]$$

$$\phi_m(q, p) = 0 \qquad [10]$$

where the u^m's are parameters, some of which will be determined through the consistency algorithm to be discussed shortly. (In [7]–[9] and everywhere below, there is a summation over the repeated indices.)

Secondary constraints The equations of motion [8] and [9] can be rewritten as

$$\dot{F} = [F, H] + u^m [F, \phi_m] \qquad [11]$$

where $F = F(q,p)$ is any function of the canonical variables. Here, the Poisson bracket is defined as usual by

$$[G, F] = \frac{\partial G}{\partial q^i}\frac{\partial F}{\partial p_i} - \frac{\partial G}{\partial p_i}\frac{\partial F}{\partial q^i} \qquad [12]$$

If one takes for F one of the primary constraints ϕ_m, one should get zero, $\dot{\phi}_m = 0$. This yields the consistency conditions

$$[\phi_m, H] + u^{m'}[\phi_m, \phi_{m'}] = 0 \qquad [13]$$

These conditions can imply further restrictions on the canonical variables and/or impose conditions on the

variables u^m. Any new relation $X(q,p)=0$ on the canonical variables leads, in turn, to a further consistency condition $\dot{X}=[X,H]+u^{m'}[X,\phi_{m'}]=0$, which can bring in either further restriction on the constraint surface or fix more variables u^m. Constraints that follow from the consistency algorithm are called "secondary constraints." Finally, one is left with a certain number of secondary constraints, which are denoted by $\phi_k=0$, $k=M+1,\ldots,M+K$. We assume again that all the constraints (primary and secondary) define a smooth surface, called the "constraint surface," and fulfill the condition that $\partial(\phi_k)/\partial(q^i,p_i)$ is of maximum rank $J\equiv M+K$ on the constraint surface. (We also assume for simplicity that there is no branching in the consistency algorithm.)

Restrictions on the u's Having a complete set of constraints

$$\phi_j=0, \quad j=1,\ldots,M+K\equiv J \qquad [14]$$

we can now investigate more precisely the restrictions on the variables u^m. These read

$$[\phi_j,H]+u^m[\phi_j,\phi_m]\approx 0, \quad j=1,\ldots,J \qquad [15]$$

where the notation $\approx$ means "equal modulo the constraints." In [15], m is summed from 1 to M. Equations [15] are a set of J linear, inhomogeneous equations for the u's, with coefficients that are functions of the canonical variables q^i, p_i. The general solution of this system is of the form

$$u^m=U^m+u^aV_a^m \qquad [16]$$

where U^m is a particular solution and where the V_a^m $(a=1,\ldots,A)$ provide a complete set of independent solutions of the homogeneous system

$$V_a^m[\phi_j,\phi_m]\approx 0 \qquad [17]$$

The coefficients $u^a(a=1,\ldots,A)$ are completely arbitrary.

We thus see the emergence of another new feature in the theory, in addition to the appearance of constraints. It is that the general solution of the equations of motion may contain arbitrary functions of time (when $A\neq 0$), in agreement with the possible presence of a gauge symmetry.

First- and Second-Class Constraints

First- and second-class functions A function $F(q,p)$ is called a first-class function if it generates a canonical transformation that maps the constraint surface on itself. Thus, $F(q,p)$ is first class if its Poisson brackets with all the constraints vanish weakly (i.e., are zero on the constraint surface),

$$[F,\phi_j]\approx 0, \quad j=1,\ldots,J \qquad [18]$$

A function is second class otherwise, that is, if there is at least one constraint ϕ_j such that $[F,\phi_j]\neq 0$ (not even weakly). Second-class functions generate canonical transformations that do not leave the constraint surface invariant. Since canonical transformations that map the constraint surface on itself form a group, the Poisson bracket of two first-class functions is itself a first-class function.

Because the system is constrained to lie on the constraint surface, the only allowed canonical transformations are those that are generated by first-class functions. The importance of the distinction between first-class and second-class functions stems from this elementary fact. Note, in particular, that the time evolution is generated – as it should – by a first-class generator since the equations of motion [11] can be rewritten as

$$\dot{F}\approx[F,H']+u^a[F,V_a^m\phi_m] \qquad [19]$$

with

$$H'=H+U^m\phi_m \qquad [20]$$

One has both $[H',\phi_m]\approx 0$ and $[V_a^m\phi_m,\phi_j]\approx 0$.

Splitting of the constraints One can separate the constraints between first-class and second-class constraints. This can be achieved by considering the matrix $C_{jj'}$ of the Poisson bracket of the constraints,

$$C_{jj'}=[\phi_j,\phi_{j'}], \quad j,j'=1,\ldots,J \qquad [21]$$

One has the following theorem due to Dirac.

Theorem 1 *If* $\det C_{jj'}\approx 0$, *there exists at least one first-class constraint among the* ϕ_j*'s.*

Proof Straightforward: if $\det C_{jj'}\approx 0$, one can find a nontrivial solution λ^j of $\lambda^jC_{jj'}\approx 0$. The corresponding constraint $\lambda^j\phi_j$ is easily verified to be first class.

By redefining the constraints as $\phi_j\rightarrow\bar{\phi}_j=a_j^{j'}\phi_{j'}$ with $a_j^{j'}(q,p)$ invertible, one can bring the Poisson brackets of the constraints to the form

$$[\gamma_a,\gamma_b]=0, \qquad [\gamma_a,\chi_\alpha]=0, \qquad [\chi_\alpha,\chi_\beta]=C_{\alpha\beta} \qquad [22]$$

with $(\bar{\phi}_j)\equiv(\gamma_a,\chi_\alpha)$ and where the matrix $C_{\alpha\beta}$ is invertible. (We assume, for simplicity, throughout that the rank of the matrix $C_{jj'}$ is constant on the constraint surface ("regular case").) In this representation, the constraints are completely split into first-class constraints (γ_a) and second-class

constraints (χ_α): there is no first-class constraint left among the χ_α's, and the set $\{\gamma_a\}$ exhausts all the first-class constraints. Note that now the index $a = 1, \ldots, A, A+1, \ldots, \bar{A}$ runs over all (primary and secondary) first-class constraints.

This separation of the constraints into first-class and second-class constraints is quite important because, as already seen above, the first-class constraints generate admissible canonical transformations, while the second-class constraints do not.

For a bosonic system, the matrix $C_{\alpha\beta}$ is antisymmetric. As $C_{\alpha\beta}$ is invertible, this implies that the number of second-class constraints is even. In the fermionic case, $C_{\alpha\beta}$ is symmetric (in the fermionic sector) and, therefore, the number of second-class constraints can be even or odd.

First-class constraints and gauge symmetries The first-class constraints not only map the constraint surface on itself, but generate, in fact, transformations that do not change the physical state of the system, that is, gauge transformations. Indeed, the presence of arbitrary functions in the solutions of the equations of motion indicates that the q's and the p's involve some redundancy and are not all physically distinct. Only those phase-space functions whose time evolution does not depend on the arbitrary functions u^a are observables.

That the first-class constraints generate gauge transformations is rather clear in the case of the first-class primary constraints, since these appear explicitly in the generator of the time evolution multiplied by arbitrary functions. That it also holds for the first-class secondary constraints is known as the "Dirac conjecture." This conjecture can be proved under reasonable assumptions (see, e.g., Henneaux *et al.* 1990). The reason that the secondary first-class constraints also correspond to gauge transformations is that they appear in the brackets of the Hamiltonian with the primary first-class constraints. Thus, different choices of arbitrary functions u^a in the dynamical equations of motion will lead to phase-space points that differ by a canonical transformation whose generator involves the secondary first-class constraints as well.

In any case, as noted below, one must identify the phase-space points in the same orbit generated by all the first-class constraints (primary and secondary) in order to get a reduced space with a symplectic structure ("reduced phase space"). For this reason, one postulates that the first-class constraints always generate gauge transformations, even for systems which are counterexamples to the Dirac conjecture (i.e., in that case, one defines the gauge transformations as being the transformations generated by the first-class constraints).

The extended Hamiltonian H_E is defined to be the sum of the first-class Hamiltonian [20] and of all the first-class constraints γ_a multiplied by an arbitrary Lagrange multiplier,

$$H_E = H' + v^a \gamma_a \qquad [23]$$

(with a summed from 1 to $\bar{A}$). It is the generator of the time evolution in which the complete gauge symmetry is fully displayed.

Elimination of second-class constraints – Dirac brackets Second-class constraints do not generate permissible canonical transformations, since they do not map the constraint surface on itself. For this reason, it is convenient to eliminate them. This can consistently be done by using the Dirac brackets instead of the Poisson brackets. By definition, the Dirac bracket $[F, G]_D$ of two phase-space functions F and G is given by

$$[F, D]_D = [F, G] - [F, \chi_\alpha] C^{\alpha\beta} [\chi_\beta, G] \qquad [24]$$

where $C^{\alpha\beta}$ is the inverse to $C_{\alpha\beta}$,

$$C^{\alpha\beta} C_{\beta\gamma} = \delta^\alpha_\gamma$$

(which exists since the χ_α's are second class). As shown by Dirac, the bracket [24] is indeed a bracket (antisymmetry, derivation property, and Jacobi identity). Furthermore, it fulfills the crucial property that the Dirac bracket of anything with any second-class constraint is zero,

$$[F, \chi_\alpha]_D = 0 \quad (F \text{ arbitrary}) \qquad [25]$$

Thus, one can consistently eliminate the second-class constraints and replace the Poisson bracket by the Dirac bracket. Once this is done, one has fewer canonical variables and only first-class constraints remain (if any). It also follows from the definition that the Dirac bracket of two first-class functions is equal to their Poisson bracket.

Gauge conditions One can push the reduction procedure further and eliminate the first-class constraints by means of gauge conditions. Gauge conditions $C_a = 0$ are conditions on the phase-space variables which do not follow from the Lagrangian and which have the property that they cut each gauge orbit once and only once. Since the gauge transformations are generated by the first-class constraints, this requirement is (locally) equivalent to

$$[C_a, \gamma_b] \varepsilon^b \approx 0 \Rightarrow \varepsilon^b \approx 0 \qquad [26]$$

That is, the constraints (γ_a, C_b) form together a second-class system: there is no first-class constraint left once the conditions $C_a = 0$ are included. One can then eliminate all the constraints and gauge conditions and introduce the corresponding Dirac bracket. For gauge-invariant functions, this Dirac bracket coincides with the original Poisson bracket.

The reduced phase space is the unconstrained space obtained after this reduction, equipped with the Dirac bracket. It has dimension $2n - s - 2\bar{A}$, where $2n$ is the dimension of the original phase space, s is the number of second-class constraints, and $\bar{A}$ is the number of first-class constraints. In the bosonic case, this number is even (as it should) because s is even. One sees that "first-class constraints strike twice" since they need gauge conditions.

The observables of the theory are the reduced phase-space functions. They form a Poisson algebra, the relevant reduced phase-space bracket being the Dirac bracket associated with all the constraints and gauge conditions. The symplectic structure defined in the reduced phase space is nondegenerate because one has removed all the first-class constraints.

The definition of reduced phase space given above is useful in practice but has the conceptual drawback of relying on gauge conditions. This approach does not display clearly its intrinsic significance and, furthermore, in the case of the so-called Gribov problems (global obstructions to cutting each gauge orbit once and only once), may yield the incorrect expectation that the reduced phase space does not exist. We shall provide a more intrinsic definition below, which does not involve gauge conditions.

Examples

First example (see eqn [2]). There is here one primary constraint, namely $\pi_\lambda = 0$. The canonical Hamiltonian is $(1/2)((p_1)^2 + (p_2)^2) + \lambda(p_1 + p_2)$. The consistency algorithm yields the secondary constraint $p_1 + p_2 = 0$ and no condition on the u's. The constraints are first class. They generate the gauge transformations $q^1 \to q^1 + \varepsilon$, $q^2 \to q^2 + \varepsilon$, and $\lambda \to \lambda + \eta$, which coincide with the Lagrangian gauge transformations if one identifies η with $\dot{\varepsilon}$ (ε and $\dot{\varepsilon}$ are, of course, independent at any given time). One can fix the gauge by means of the gauge conditions $\lambda = 0$, $q^1 + q^2 = 0$. The reduced phase space is two-dimensional and the observables can be identified with the functions of the gauge-invariant variables $(1/2)(q^1 - q^2)$ and $p_1 - p_2$, which are conjugate. Any other gauge condition leads to the same reduced phase space.

Second example (see eqn [4]). The primary constraints are $p_1 - z^2 = 0$ and $p_2 = 0$ and define a two-dimensional plane in the four-dimensional phase space (z^1, z^2, p_1, p_2). The consistency algorithm forces $u^1 = z^2$ and $u^2 = 0$ and does not bring any further constraint. The constraints are second class since $[p_2, p_1 - z^2] = 1$. One can eliminate p_1 and p_2 through the constraints. The Dirac brackets of the remaining variables vanish, except $[z^1, z^2] = 1$. The reduced phase is the space of the z's, with z^2 conjugate to z^1. The Hamiltonian is the free-particle Hamiltonian , $H = (1/2)(z^2)^2$. Thus, one recovers the original description which was already in Hamiltonian form. (The recognition that a system is already in first-order form often enables one to shortcut some aspects of the Dirac procedure by not introducing the unnecessary momenta which would in any case be eliminated in the end.)

Quantization

The phase space of physical interest is the reduced phase space and the physical algebra is the algebra of the observables. The quantization of the theory then amounts to quantizing the algebra of the observables. This can be achieved along two different lines:

1. *Reduce then quantize*: In this direct approach, one represents as quantum operators only the reduced phase-space functions. There is no operator associated with non-gauge-invariant functions.
2. *Quantize then reduce*: In this approach, one represents as quantum operators the bigger algebra of functions of all the phase-space variables. One must then take into account the constraints. The second-class constraints are enforced as operator equations, which is consistent with the correspondence rule that the commutator in the quantum theory is $i\hbar$ times the Dirac bracket,

$$AB - BA = i\hbar[A, B]_D \qquad [27]$$

(plus higher-order terms in $\hbar$). The first-class constraints are implemented in a more subtle way. It would be inconsistent to impose them as operator equations since in general $[\gamma_a, F]_D \neq 0$ (even in the Dirac bracket). What one does is to impose them as conditions on the physical states: these are defined as the states annihilated by the first-class constraints,

$$\gamma_a|\psi\rangle = 0 \qquad [28]$$

For simple systems, it is easy to verify that the two procedures are equivalent. There is yet another

approach, in which one extends the system rather than reduce it. This is the Becchi–Rouet–Stora–Tyutin (BRST) approach, in which the new variables are called ghosts.

Geometric Description

We defined above first-class and second-class constraints through algebraic means. It turns out that these definitions also have a geometrical interpretation, which sheds considerable insight into their nature.

The phase-space symplectic 2-form σ induces, by pullback, a 2-form σ_Σ on the constraint surface Σ. While σ is of maximal rank, this may not be the case for the induced σ_Σ, which may be degenerate. In fact, the rank of σ_Σ fails to be equal to the maximum rank $2n - J$ (where J is the total number of constraints) by precisely the number $\bar{A}$ of first-class constraints.

Indeed, the Hamiltonian vector fields $X_{\gamma a}$ associated with the first-class constraints are tangent to the constraint surface Σ and are null eigenvectors of σ_Σ,

$$\sigma_\Sigma(X_{\gamma a}, Y) = 0 \quad \forall Y \text{ tangent to } \Sigma \qquad [29]$$

as an immediate consequence of the first-class property. Here, all first-class constraints (primary and secondary) yield a null eigenvector. The integral surfaces of the vector fields $X_{\gamma a}$ are the gauge orbits. The reduced phase space is nothing else but the quotient space of the constraint surface by the gauge orbits. The 2-form induced in the quotient space is invertible because one has removed all degeneracy directions (including the ones associated with secondary first-class constraints). Reaching the reduced phase space falls under the scope of Hamiltonian reduction. The observables are the functions on the reduced phase space.

Thus, the reduced phase space is obtained through a two-step procedure. First, one restricts the functions to functions on the constraint surface Σ. One may view the algebra $C^\infty(\Sigma)$ of smooth functions on Σ as the quotient algebra $C^\infty(P)/\mathcal{N}$ of the algebra $C^\infty(P)$ of smooth phase-space functions by the ideal $\mathcal{N}$ of phase-space functions that vanish on the constraint surface σ. The second step in the reduction procedure is to impose the gauge-invariant condition on the functions in $C^\infty(\Sigma)$, that is, to impose that they are constant along the gauge orbits $\mathcal{O}$. Assuming all necessary smoothness and regularity conditions to be fulfilled (i.e., that the orbits fiber which is, for instance, the case if the gauge orbits are the orbits of a free and proper group action), one may denote the algebra of observables as $C^\infty(\Sigma/\mathcal{O})$. This algebra is a Poisson algebra because the induced 2-form on the quotient space $\Sigma/\mathcal{O}$ is nondegenerate. The algebraic description of the observables underlies the BRST construction.

It is interesting to note that in the covariant approach to phase space, a similar two-step reduction procedure occurs. What plays the role of the constraint surface is the stationary surface in the space of all histories $q^i(t)$ of the dynamical variables. The gauge symmetry acts on this space and the reduced phase space is just the quotient space. One can establish the equivalence of the two descriptions (Barnich *et al.* 1991).

See also: Batalin–Vilkovisky Quantization; BRST Quantization; Canonical General Relativity; Operads; Perturbative Renormalization Theory and BRST; Quantum Dynamics in Loop Quantum Gravity; Quantum Field Theory: A Brief Introduction.

Further Reading

Anderson JL and Bergmann PG (1951) Constraints in covariant field theories. *Physical Review* 83: 1018.

Barnich G, Henneaux M, and Schomblond C (1991) On the covariant description of the canonical formalism. *Physical Review D* 44: 939.

Dirac PAM (1950) Generalized Hamiltonian dynamics. *Canadian Journal of Mathematics* 2: 129.

Dirac PAM (1967) *Lectures on Quantum Mechanics*. New York: Academic Press.

Flato M, Lichnerowicz A, and Sternheimer D (1976) Deformations of Poisson brackets, Dirac brackets and applications. *Journal of Mathematical Physics* 17: 1754.

Hanson A, Regge T, and Teitelboim C (1976) *Constrained Hamiltonian Systems*. Rome: Accad. Naz. dei Lincei.

Henneaux M and Teitelboim C (1992) *Quantization of Gauge Systems*. Princeton: Princeton University Press.

Henneaux M, Teitelboim C, and Zanelli J (1990) Gauge invariance and degree of freedom count. *Nuclear Physics B* 332: 169.

Marsden JE and Weinstein A (1974) Reduction of symplectic manifolds with symmetry. *Reports on Mathematical Physics* 5: 121.

Constructive Quantum Field Theory

G Gallavotti, Università di Roma "La Sapienza," Rome, Italy

© 2006 G Gallavotti. Published by Elsevier Ltd. All rights reserved.

Euclidean Quantum Fields

The construction of a relativistic quantum field is still an open problem for fields in spacetime dimension $d \geq 4$. The conceptual difficulty that sometimes led to fear an incompatibility between nontrivial quantum systems and special relativity has however been solved in the case of dimension $d = 2, 3$ although, so far, has not influenced the corresponding debate on the foundations of quantum mechanics, still much alive.

It began in the early 1960s with Wightman's work on the axioms and the attempts at understanding the mathematical aspects of renormalization theory and with Hepps' renormalization theory for scalar fields. The breakthrough idea was, perhaps, Nelson's realization that the problem could really be studied in Euclidean form. A solution in dimensions $d = 2, 3$ has been obtained in the 1960s and 1970s through a remarkable series of papers by Nelson, Glimm, Jaffe, and Guerra. While the works of Nelson and Guerra relied on the "Euclidean approach" (see below) and on $d = 2$, the early works of Glimm and Jaffe dealt with $d = 3$ making use of the "Minkowskian approach" (based on second quantization) but making already use of a "multiscale analysis" technique. The latter received great impulsion and systematization by the adoption of Wilson's views and methods on renormalization: in physics terminology, renormalization group methods; a point of view taken here following the Euclidean approach. The solution dealt initially with scalar fields but it has been subsequently considerably extended.

The Euclidean approach studies quantum fields through the following problems:

1. existence of the functional integrals defining the generating functions (see below) of the probability distribution of the interacting fields in finite volume: the "ultraviolet stability problem,"
2. existence of the infinite-volume limit of the generating functions: the "infrared problem," and
3. check that the infinite volume generating functions satisfy the axioms needed to pass from the Euclidean, probabilistic, formulation to a Minkowskian formulation guaranteeing the existence of the Hamiltonian operator, relativistic covariance, Ruelle–Haag scattering theory: the "reconstruction problem."

The characteristic problem for the construction of quantum fields is (1) and here attention will be confined to it with the further restriction to the paradigmatic massive scalar fields cases. The dimension d of the spacetime will be $d = 2, 3$ unless specified otherwise.

Given a cube Λ of side L, $\Lambda \subset \mathbb{R}^d$, consider the following functional integral on the space of the fields on Λ, that is, on functions $\varphi_\xi^{(\leq N)}$ defined for $\xi \in \Lambda$,

$$Z_N(\Lambda, f) = \int \exp\left(-\int_\Lambda \left(\lambda_N \varphi_\xi^{(\leq N)^4} + \mu_N \varphi_\xi^{(\leq N)^2} + \nu_N + f_\xi \varphi_\xi^{(\leq N)}\right) d\xi\right) P_N(d\varphi^{(\leq N)}) \qquad [1]$$

The fields $\varphi_\xi^{(\leq N)}$ are called "Euclidean" fields with ultraviolet cutoff $N > 0$, f_ξ is a smooth function with compact support bounded by $|f_\xi| \leq 1$ (for definiteness), the constants $\lambda_N > 0, \mu_N, \nu_N$ are called "bare couplings," and P_N is a Gaussian probability distribution defining the free-field distribution with mass m and ultraviolet cutoff N; the probability distribution P_N is determined by its "covariance" $C_{\xi,\eta}^{(\leq N)} \stackrel{\text{def}}{=} \int \varphi_\xi^{(\leq N)} \varphi_\eta^{(\leq N)} dP_N$, which in the physics literature is called a "propagator," given by

$$C_{\xi,\eta}^{(\leq N)} = \frac{1}{(2\pi)^d} \sum_{n \in \mathbb{Z}^d} \int \frac{e^{ip\cdot(\xi-\eta+nL)}}{p^2 + m^2} \chi_N(|p|) d^d p \qquad [2]$$

The sum over the integers $n \in \mathbb{Z}^d$ is introduced so that the field $\varphi_\xi^{(\leq N)}$ is periodic over the box Λ: this is not really necessary as in the limit $L \to \infty$ either translation invariance would be recovered or lack of it properly understood, but it makes the problem more symmetric and generates a few technical simplifications; here $\chi_N(z)$ is a "regularizer" and a standard choice is

$$\chi_N(|p|) = \frac{m^2(\gamma^{2N} - 1)}{p^2 + \gamma^{2N} m^2}$$

with $\gamma > 1$, which is such that

$$\frac{\chi_N(|p|)}{p^2 + m^2} \equiv \frac{1}{p^2 + m^2} - \frac{1}{p^2 + \gamma^{2N} m^2} \equiv \sum_{h=1}^{N} \left(\frac{1}{p^2 + \gamma^{2(h-1)} m^2} - \frac{1}{p^2 + \gamma^{2h} m^2}\right) \qquad [3]$$

here $\gamma > 1$ can be chosen arbitrarily: so $\gamma = 2$. If $d > 3$, the above regularization will not be sufficient and a χ_N decaying faster than p^{-2} would be needed.

A simple estimate yields, if $\varepsilon \in (0,1)$ is fixed and c is suitably chosen,

$$\left|C_{\xi,\eta}^{(\leq N)}\right| \leq c\gamma^{(d-2)N} e^{-m|\xi-\eta|}$$
$$\left|C_{\xi,\eta}^{(\leq N)} - C_{\xi,\eta'}^{(\leq N)}\right| \leq c\gamma^{(d-2)N}(\gamma^N m|\eta-\eta'|)^{\varepsilon} \quad [4]$$

with $\gamma^{(d-2)N}$ interpreted as N if $d=2$.

The

$$\zeta(f) = \log \frac{Z_N(\Lambda, f)}{Z_N(\Lambda, 0)}$$

defines a "generating function" of a probability distribution P_{int} over the fields on Λ which will be called the "distribution with φ^4-interaction" regularized on Λ and at length scale $m^{-1}\gamma^{-N}$: the integral, in [1],

$$V_N\left(\varphi^{(\leq N)}\right) \stackrel{\text{def}}{=} -\int_\Lambda \left(\lambda_N \varphi_\xi^{(\leq N)^4} + \mu_N \varphi_\xi^{(\leq N)^2} + \nu_N + f_\xi \varphi_\xi^{(\leq N)}\right) d^d\xi \quad [5]$$

will be called the "interaction potential" with external field f. The regularization is introduced to guarantee that the integral [1], $\int e^{V_N}\, dP_N$, is well defined if $\lambda_N > 0$. The momenta of P_{int} are the functional derivatives of $\zeta(f)$: they are called "Schwinger functions."

The problem (1) can now be made precise: it is to show the existence of λ_N, μ_N, ν_N so that the limit

$$\lim_{N\to\infty} \frac{Z_N(\Lambda, f)}{Z_N(\Lambda, 0)}$$

exists for all f and is not Gaussian, that is, it is not the exponential of a quadratic form in f: which would be the case if $\lambda_N, \mu_N \to 0$ fast enough: the last requirement is of course essential because the Gaussian case describes, in the physical interpretation, free fields and noninteracting particles, that is, it is trivial. Note that ν_N does not play a role: its introduction is useful to be able to study separately the numerator and the denominator of the fraction

$$\frac{Z_N(\Lambda, f)}{Z_N(\Lambda, 0)}$$

For more details, the reader is referred to Wightman and Gärding (1965), Streater and Wightman (1964), Nelson (1966), Guerra (1972), Osterwalder and Schrader (1973), and Simon (1974).

The Regularized Free Field

Since the propagator, see [4], decays exponentially over a scale m^{-1} and is smooth over a scale $m^{-1}\gamma^{-N}$, the fields $\varphi_\xi^{(\leq N)}$ sampled with distribution P_N are rather singular objects. Their properties cannot be described by a single length scale: they are extremely large for large N, take independent values only beyond distances of order m^{-1} but, at the same time, they look smooth only on the much smaller scale $m^{-1}\gamma^{-N}$. Their essential feature is that fixed $\varepsilon < 1$, for example, $\varepsilon = 1/2$, with P_N-probability 1 there is $B > 0$ such that (interpreting $\gamma^{(d-2)/2N}$ as N if $d=2$)

$$\left|\varphi_\xi^{(\leq N)}\right| \leq B\gamma^{N(d-2)/2}$$
$$\left|\varphi_\xi^{(\leq N)} - \varphi_\eta^{(\leq N)}\right| < B\gamma^{N(d-2)/2}\left(\gamma^N m|\xi-\eta|\right)^{\varepsilon/2} \quad [6]$$

and furthermore the probability of the relations in [6] will be N-independent, that is, $\varphi_\xi^{(\leq N)}$ are bounded and roughly of size $\gamma^{N(d-2)/2}$ as $N \to \infty$ and, on a very small length scale $m^{-1}\gamma^{-N}$, almost constant.

Substantial control on the field $\varphi_\xi^{(\leq N)}$ statistically sampled with distribution P_N can be obtained by decomposing it, through [3], into "components of various scales": that is, as a sum of statistically mutually independent fields whose properties are entirely characterized by a single scale of length. This means that they have size of order 1 and are independent and smooth on the same length scale.

Assuming the side of Λ to be an integer multiple of m^{-1}, let $\mathcal{Q}_h$ be a pavement of Λ into boxes of side $m^{-1}\gamma^{-h}$, imagined hierarchically arranged so that the boxes of $\mathcal{Q}_h$ are exactly paved by those of $\mathcal{Q}_{h+1}$.

Define $z_\xi^{(h)}$ to be the random field with propagator $C_{\xi,\eta}^{(h)}$ with Fourier transform

$$\sum_{n\in\mathbb{Z}^d}\left(\frac{1}{p^2+\gamma^{-2}m^2} - \frac{1}{p^2+m^2}\right) e^{in\cdot p L\gamma^h}$$

so that $\varphi_\xi^{(\leq N)}$ and its propagator $C_{\xi,\eta}^{(\leq N)}$ can be represented, see [2], [3], as

$$\varphi_\xi^{(\leq N)} \equiv \sum_{h=1}^{N} \gamma^{h(d-2)/2} z_{\gamma^h\xi}^{(h)}$$
$$C_{\xi,\eta}^{(\leq N)} = \sum_{h=1}^{N} \gamma^{h(d-2)} C_{\gamma^h\xi,\gamma^h\eta}^{(h)} \quad [7]$$

where the fields $z^{(h)}$ are independently distributed Gaussian fields. Note that the fields $z^{(h)}$ are also almost identically distributed because their propagator is obtained by periodizing over the period $\gamma^h L$ the same function

$$\overline{C}_{\xi,\eta}^{(0)} \stackrel{\text{def}}{=} \int \frac{e^{ip\cdot(\xi-\eta)}dp}{(2\pi)^d}\left(\frac{1}{p^2+\gamma^{-2}m^2} - \frac{1}{p^2+m^2}\right)$$

that is, their propagator is

$$C^{(h)}_{\xi,\eta} = \sum_{n\in\mathbb{Z}^d} \overline{C}^{(0)}_{\xi,\eta+\gamma^h nL}$$

The reason why they are not exactly equally distributed is that the field $z^{(h)}_\xi$ is periodic with period $\gamma^h L$ rather than L. But proceeding with care the sum over n in the above expressions can be essentially ignored: this is a little price to pay if one wants translation invariance built in the analysis since the beginning.

The representation [7] defines a "multiscale representation" of the field $\varphi^{(\le N)}_\xi$. Smoothness properties for the field $\varphi^{(\le N)}_\xi$ can be read from those of its "components" $z^{(h)}$. Define, for $\Delta \in \mathcal{Q}_0$,

$$\left\| z^{(h)} \right\|_\Delta = \max_{\substack{\xi\in\Delta,\eta\in\Lambda \\ |\xi-\eta|\le m^{-1}}} \left(\left| z^{(h)}_\xi \right| + \tau \frac{\left| z^{(h)}_\xi - z^{(h)}_\eta \right|}{|\xi-\eta|^{1/4}} \right) \qquad [8]$$

and τ will be chosen $\tau=0$ or $\tau=1$ as needed (in practice $\tau=0$ if $d=2$ and $\tau=1$ if $d=3$): $\tau=1$ will allow us to discuss some smoothness properties of the fields which will be necessary (e.g., if $d=3$). Then the size $\|z\|_\Delta$ of any field $z^{(h)}$, for all $h\ge 1$, is estimated by

$$\begin{aligned} &P\Big(\max_{\Delta\subset\mathcal{Q}_0} \|z\|_\Delta \le B \Big) \ge e^{-ce^{c'B^2}|\Lambda|} \\ &P(\|z\|_\Delta \ge B_\Delta, \forall \Delta \in \mathcal{D}) \le \prod_{\Delta\in\mathcal{D}} c e^{-c' B_\Delta^2} \end{aligned} \qquad [9]$$

where P is the Gaussian probability distribution of z, $\mathcal{D}$ is any collection of boxes $\Delta \in \mathcal{Q}_0$ and $c, c' > 0$ are suitable constants. The [9] imply in particular [6]. The estimates [9] follow from the Markovian nature of the Gaussian field $z^{(h)}$, that is, from the fact that the propagator is the Green's function of an elliptic operator (of fourth order, see the first of [3]), with constant coefficients which implies also the inequalities (fixing $\varepsilon \in (0,1)$)

$$\begin{aligned} &\left| C^{(h)}_{\xi,\eta} \right| \equiv \left| \int z_\xi z_\eta P(dz) \right| \le c e^{-mc'|\xi-\eta|} \\ &\left| C^{(h)}_{\xi,\eta} - C^{(h)}_{\xi,\eta'} \right| \le c(m|\eta-\eta'|)^\varepsilon \end{aligned} \qquad [10]$$

where $|\xi-\eta|$ is reinterpreted as the distance between ξ, η measured over the periodic box $\gamma^h\Lambda$ (hence $|\xi-\eta|$ differs from the ordinary distance only if the latter is of the order of $\gamma^h L$). The interpretation of [10] is that $z^{(h)}_\xi$ are essentially bounded variables which, on scale $\sim m^{-1}$, are essentially constant and furthermore beyond length $\sim m^{-1}$ are essentially independently distributed.

For more details, the reader is referred to Wilson (1970, 1972) and Gallavotti (1981, 1985).

Perturbation Theory

The naive approach to the problem is to fix $\lambda_N \equiv \lambda > 0$ and to develop $Z_N(\Lambda, f)$ or, more conveniently and equivalently, $(1/|\Lambda|)\log Z_N(\Lambda, f)$ in powers of λ. If one fixes *a priori* μ_N, ν_N independent of N, however, even a formal power series is not possible: this is trivially due to the divergence of the coefficients of the power series, already to second order, for generic f in the limit $N\to\infty$. Nevertheless it is possible to determine $\mu_N(\lambda), \nu_N(\lambda)$ as functions of N and λ so that a formal power series exists (to all orders in λ): this is the key result of renormalization theory.

To find the perturbative expansion, the simplest is to use a graphical representation of the coefficients of the power expansion in λ, μ_N, ν_N, f and the Gaussian integration rules which yield (after a classical computation) that the coefficient of $\lambda^n \mu_N^p f_{\xi_1} \dots f_{\xi_r}$ is obtained by considering the graph elements shown in **Figure 1**, where the segments will be called half-lines and the graph elements will be called, respectively, "coupling" or "φ^4-vertex," "mass vertex," "vacuum vertex," and "external vertex."

The half-lines of the graph elements are considered distinct (i.e., imagine a label attached to distinguish them). Then consider all possible *connected* graphs G obtained by first drawing, respectively, n, p, r graph elements in **Figure 1**, which are not vacuum vertices, with their nodes marked by points in Λ named $\xi_1, \dots, \xi_n, \xi_{n+1}, \dots, \xi_{n+p+r}$; and form all possible graphs obtained by attaching pairs of half-lines emerging from the vertices of the graph elements. These are the "nontrivial graphs." Furthermore, consider also the single "trivial" graph formed just by the third graph element and consisting of a single point. All graphs obtained in this way are particular Feynman graphs.

Given a nontrivial graph G (there are many of them) we define its value to be the product

$$\begin{aligned} &W_G(\xi_1, \dots, \xi_n, \xi_{n+1}, \dots, \xi_{n+p+r}) \\ &\quad = (-1)^{n+p+r} \frac{\lambda^n \mu_N^p \prod f_{\xi_{n+p+i}}}{n!p!r!} \prod_\ell C^{(\le N)}_{\xi_\ell,\eta_\ell} \end{aligned} \qquad [11]$$

where the last product runs over all pairs $\ell = (\xi_\ell, \eta_\ell)$ of half-lines of G that are joined and connect two vertices labeled by points ξ_ℓ, η_ℓ: "call line of G" any such pair. If the graph consists of the single vacuum

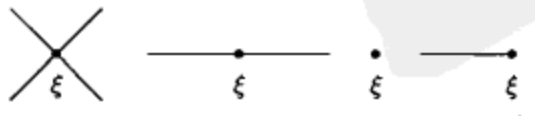


Figure 1 The graph elements to representing $\varphi^{(\le N)4}_\xi, \varphi^{(\le N)2}_\xi$, a constant $\varphi^{(\le N)}$.

vertex its value will be ν_N. The series for $(1/|\Lambda|)\log Z_N(\Lambda, f)$ is then

$$-\nu_N + \frac{1}{|\Lambda|}\sum_G \int W_G(\xi_1, \ldots, \xi_{n+p+r}) \prod_{j=1}^{n+p+r} d\xi_j \qquad [12]$$

and the integral will be called the integrated graph value.

Suppose first that $\mu_N = \nu_N = 0$. Then if a graph G contains subgraphs like in **Figure 2**, the corresponding respective contribution to the integral in [12] (considering only the integrals over η and suitably taking care of the combinatorial factors) is a factor obtained by integrating over ξ the quantities

$$-6\lambda C_{\alpha\xi}^{(\le N)} C_{\xi\xi}^{(\le N)} C_{\xi\beta}^{(\le N)}$$

$$\text{or} \qquad \frac{4^2 \cdot 3!}{2!}\lambda^2 C_{\alpha\xi}^{(\le N)} \int C_{\xi\eta}^{(\le N)3} C_{\eta\beta}^{(\le N)} d\eta \qquad [13]$$

which if $d = 3$ diverge as $N \to \infty$ as γ^N or, respectively, as N; the second factor does not diverge in dimension $d = 2$ while the first still diverges as N. The divergences arise from the fact that as $\xi - \eta \to 0$ the propagator behaves as $|\xi - \eta|^{-N}$ if $d = 3$ or as $-\log|\xi - \eta|$ if $d = 2$, all the way until saturation occurs at distance $|\xi - \eta| \simeq m^{-1}\gamma^{-N}$: for this reason the latter divergences are called "ultraviolet divergences."

However, if we set $\mu_N \neq 0$, then for every graph containing a subgraph like those in **Figure 2** there is another one identical except that the points α, β are connected via a mass vertex, see **Figure 1**, with the vertex in ξ, by a line $\alpha\xi$ and a line $\xi\beta$; the new graph value receives a contribution from the mass vertex inserted in ξ between α and β simply given by a factor $-\mu_N$. Therefore if we fix, for $d = 3$,

$$\mu_N = -6\lambda C_{\xi\xi}^{(\le N)} + \frac{4^2 \cdot 3!}{2}\lambda^2 \times \int_\Lambda C_{\xi\eta}^{(\le N)3} d\eta \stackrel{\text{def}}{=} -6\lambda C_{\xi\xi}^{(\le N)} + \delta\mu_N \qquad [14]$$

we can simply consider graphs which do not contain any mass graph element and in which there are no subgraphs like the first in **Figure 2** while the subgraphs like the second in **Figure 2** do not contribute a factor $\int C_{\alpha\xi}^{(\le N)} C_{\xi\eta}^{(\le N)3} C_{\eta\beta}^{(\le N)} d\eta$ but a renormalized factor $\int C_{\alpha\xi}^{(\le N)} C_{\xi\eta}^{(\le N)3} (C_{\eta\beta}^{(\le N)} - C_{\xi\beta}^{(\le N)}) d\eta$. If $d = 2$, we only need to define μ_N as the first term on the right-hand side (RHS) of [14] and we can leave the subgraphs like the second in **Figure 2** as they are (without any renormalization).

Graphs without external lines are called vacuum graphs and there are a few such graphs which are divergent. Namely, if $d = 3$, they are the first three drawn in **Figure 3**; furthermore, if μ_N is set to the above nonzero value a new vacuum graph, the fourth in **Figure 3**, can be formed. Such graphs contribute to the graph value, respectively, the terms in the sum

$$-3\lambda C_{\xi_1\xi_1}^{(\le N)2} + \frac{4!}{2}\lambda^2 \int C_{\xi_1\xi_2}^{(\le N)4} d\xi_2 - \frac{2^3 \cdot 3!^3}{3!}\lambda^3 \times \int C_{\xi_1\xi_2}^{(\le N)2} C_{\xi_2\xi_3}^{(\le N)2} C_{\xi_3\xi_1}^{(\le N)2} d\xi_2 d\xi_3 - \mu_N C_{\xi_1\xi_1}^{(\le N)} \qquad [15]$$

and diverge, respectively, as $\gamma^{2N}, \gamma^N, N, \gamma^{2N}$ if $d = 3$ while, if $d = 2$, only the first and the last (see [14]) diverge, like N^2.

Therefore, if we fix ν_N as minus the quantity in [15] we can disregard graphs like those in **Figure 3**; if $d = 2$ ν_N can be defined to be the sum of the first and last terms in [15].

The formal series in λ and f thus obtained is called the "renormalized series" for the field φ^4 in dimension $d = 2$ or, respectively, $d = 3$. Note that with the given definitions and choices of μ_N, ν_N the only graphs G that need to be considered to construct the expansion in λ and f are formed by the first and last graph elements in **Figure 1**, paying attention that the graphs in **Figure 3** do not contribute and, if $d = 3$, the graphs with subgraphs like the second in **Figure 2** have to be computed with the modification described.

In the next section, it will be shown that the above are the only sources of divergences as $N \to \infty$ and therefore the problem of studying [1] is solved at the level of formal power series by the subtraction in [14]. This also shows that giving a meaning to the series thus obtained is likely to be much easier if $d = 2$ than if $d = 3$.

The coefficients of order k of the expansion in λ of $(1/|\Lambda|)\log Z_N(\Lambda, f)$ can be ordered by the number $2n$ of vertices representing external fields: and have the form $\int S_{2n}^{(k)}(\xi_1, \ldots, \xi_{2n}) \prod_{i=1}^{2n} (f_{\xi_i} d\xi_i)$: the kernels $S_{2n}^{(k)}$ are the Schwinger functions of order $2n$, see the section "Euclidean quantum fields."

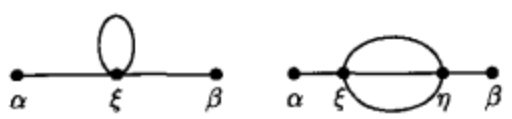


Figure 2 Divergent subgraphs, if $d = 3$. If $d = 2$ only the first diverges.



Figure 3 Divergent vacuum graphs.

Remark If $d=4$, the regularization at cutoff N in [2] is not sufficient as in the subtraction procedure smoothness of the first derivatives of the field $\varphi^{(\leq N)}$ is necessary, while the regularization [2] does not even imply [6], that is, not even Hölder continuity. A higher regularization (i.e., using a χ_N like the square of the χ_N in [3]). Furthermore, the subtractions discussed in the case $d=3$ are not sufficient to generate a formal power series and many more subtractions are needed: for instance, graphs with a subgraph like the one in **Figure 4** would give a contribution to the graph value which is a factor

$$\lambda^2 \ell_N \stackrel{\text{def}}{=} \frac{2 \cdot 6^2}{2!} \lambda^2 \int_\Lambda C_{\xi\eta}^{(\leq N)2} d\eta$$

also divergent as $N \to \infty$ proportionally to N. Although this divergence could be canceled by changing λ into $\lambda_N = \lambda + \lambda^2 \ell_N$ the previously discussed cancelations would be affected and a change in the value of μ_N would become necessary; furthermore, the subtraction in [14] will not be sufficient to make finite the graphs, not even to second order in λ, unless a new term $-\alpha_N \int (\partial_\xi \varphi_\xi^{(\leq N)})^2 \, d\xi$ with $\alpha_N = (1/2)\lambda^2 \int \partial_\eta C_{\xi\eta}^{(\leq N)3} (\xi - \eta)^2$ is added in the exponential in [1].

But all this will not be enough and still new divergences, proportional to λ^3, will appear.

And so on indefinitely, the consequence being that it will be necessary to define $\lambda_N, \mu_N, \alpha_N, \nu_N$ as formal power series in λ (with coefficients diverging as $N \to \infty$) in order to obtain a formal power series in λ for [1] in which all coefficients have a finite limit as $N \to \infty$. Thus, the interpretation of the formal renormalized series in the case $d=4$ is substantially different and naturally harder than the cases $d=2,3$. Beyond formal perturbation expansions, the case $d=4$ is still an open problem: the most widespread conjecture is that the series cannot be given a meaning other than setting to 0 all coefficients of $\lambda^j, j>0$. In other words, the conjecture claims, there should be no nontrivial solution to the ultraviolet problem for scalar φ^4 fields in $d=4$. But this is far from being proved, even at a heuristic level. The situation is simpler if $d \geq 5$: in such cases, it is impossible to find formal power series in λ for $(1/|\Lambda|) \log Z_N(\Lambda, f)$, even allowing $\lambda_N, \mu_N, \alpha_N, \nu_N$ to be formal power series in λ with divergent coefficients.



Figure 4 The simplest new divergent subgraph on $d=4$.

The distinctions between the cases $d=2,3,4,>4$ explain the terminology given to the φ^4-scalar field theories calling them super-renormalizable if $d=2,3$, renormalizable if $d=4$ and nonrenormalizable if $d>4$. Since the (divergent) coefficients in the formal power series defining $\lambda_N, \mu_N, \alpha_N, \nu_N$ are called counter-terms, the φ^4-scalar fields require finitely many counter-terms (see [14]) in the super-renormalizable cases and infinitely many in the renormalizable case. The nonrenormalizable cases ($d>4$) cannot be treated in a way analogous to the renormalizable ones.

For more details, the reader is referred to Gallavotti (1985), Aizenman (1982), and Fröhlich (1982).

Finiteness of the Renormalized Series, $d=2,3$: "Power Counting"

Checking that the renormalized series is well defined to all orders is a simple dimensional estimate characteristic of many multiscale arguments that in physics have become familiar with the name of "renormalization group arguments."

Consider a graph G with $n+r$ vertices built over n graph elements with vertices $\xi_1, \ldots, \xi_n$ each with four half-lines and r graph elements with vertices $\xi_{n+1}, \ldots, \xi_{n+r}$ representing the external fields: as remarked in the previous section, these are the only graphs to be considered to form the renormalized series.

Develop each propagator into a sum of propagators as in [7]. The graph G value will, as a consequence, be represented as a sum of values of new graphs obtained from G by adding scale labels on its lines and the value of the graph will be computed as a product of factors in which a line joining $\xi\eta$ and bearing a scale label h will contribute with $C_{\xi\eta}^{(h)}$ replacing $C_{\xi\eta}^{(\leq N)}$. To avoid proliferation of symbols, we shall call the graphs obtained in this way, i.e., with the scale labels attached to each line, still G: no confusion should arise as we shall, henceforth, only consider graphs G with each line carrying also a scale label.

The scale labels added on the lines of the graph G allow us to organize the vertices of G into "clusters": a cluster of scale h consists in a maximal set of vertices (of the graph elements in the graph) connected by lines of scale $h' \geq h$ among which one at least has scale h.

It is convenient to consider the vertices of the graph elements as "trivial" clusters of highest scale: conventionally call them clusters of scale $N+1$.

The clusters can be of "first generation" if they contain only trivial clusters, of "second generation"

if they contain only clusters which are trivial or of the first generation, and so on.

Imagine to enclose in a box the vertices of graph elements inside a cluster of the first generation and then into a larger box the vertices of the clusters of the second generation and so on: the set of boxes ordered by inclusion can then be represented by a rooted tree graph whose nodes correspond to the clusters and whose "top points" are nodes representing the trivial clusters (i.e., the vertices of the graph).

If the maximum number of nodes that have to be crossed to reach a top point of the tree starting from a node v is n_v (v included and the top nodes included), then the node v represents a cluster of the n_vth generation. The first node before the root is a cluster containing all vertices of G and the root of the tree will not be considered a node and it can conventionally bear the scale label 0: it represents symbolically the value of the graph.

For instance, in **Figure 5** a tree θ is drawn: its nodes correspond to clusters whose scale is indicated next to them; in the second part of the drawing, the trivial clusters as well as the clusters of the first generation are enclosed into boxes.

Then consider the next generation clusters, that is, the clusters which only contain clusters of the first generation or trivial ones, and draw boxes enclosing all the graph vertices that can be reached from each of them by descending the tree, etc. **Figure 6** represents all boxes (of any generation) corresponding to the nodes of the tree in **Figure 5**. The representations of the clusters of a graph G by a tree or by hierarchically ordered boxes (see **Figures 5** and **6**) are completely equivalent provided inside each box not representing a top point of the tree the scale h_v of the corresponding cluster v is marked. For

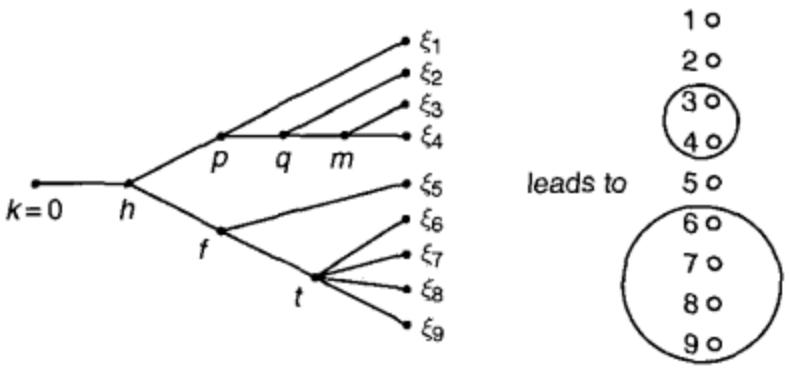


Figure 5 A tree and its clusters of generation 1 and 2.

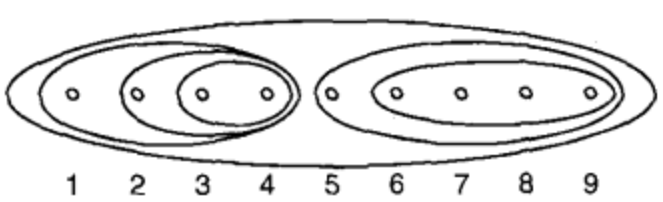


Figure 6 All clusters of any generation for the tree in **Figure 5**.

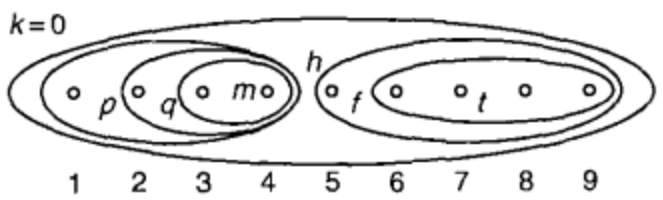


Figure 7 The clusters in **Figure 6** after affixing the scale labels.

instance, in the case of **Figure 6** one gets **Figure 7**. By construction, if two top points ξ and η are inside the same box b_v of scale h_v but not in inner boxes, then there is a path of graph lines joining ξ and η all of which have scales $\geq h_v$ and one at least has scale h_v.

Given a graph G, fix one of its points ξ_1 (say) and integrate the absolute value of the graph over the positions of the remaining points. The exponential decay of the propagators implies that if a point η is linked to a point η' by a line of scale h the integration over the position of η' is essentially constrained to extend only over a distance $\gamma^{-h}m^{-1}$. Furthermore, the maximum size of the propagator associated with a line of scale h is bounded proportionally to $\gamma^{(d-2)h}$. Therefore, recalling that $|f_\xi|$ is suppose bounded by 1, the mentioned integral can be immediately bounded by

$$\frac{\lambda^n}{n!r!} C^{n+r} I \stackrel{\text{def}}{=} \frac{\lambda^n C^{n+r}}{n!r!} \prod_\ell \gamma^{(d-2)/2h_\ell} \prod_v \gamma^{-dh_v(s_v-1)} \quad [16]$$

where, C being a suitable constant, the first product is over the half-lines ℓ composing the graph lines and the second is over the tree nodes (i.e., over the clusters of the graph G), s_v is the number of subclusters contained in the cluster v but not in inner clusters; and in [16] the scale of a half-line ℓ is h_ℓ if ℓ is paired with another half-line to form a line ℓ (in the graph G) of scale label h_ℓ.

Denoting by v' the cluster immediately containing v in G, by n_v^{inner} the number of half-lines in the cluster v, by n_v, r_v the numbers of graph elements of the first type or of the fourth type in **Figure 1** with vertices in the cluster v, and denoting by n_v^e the number of lines which are not in the cluster v but have one extreme on a vertex in v ("lines external to v"), the identities ($k=0$)

$$\sum_{v>\text{root}} (h_v - k)(s_v - 1) \equiv \sum_{v>\text{root}} (h_v - h_{v'})(n_v + r_v - 1)$$

$$\sum_{v>\text{root}} (h_v - k) n_v^{\text{inner}} \equiv \sum_{v>\text{root}} (h_v - h_{v'}) \tilde{n}_v^{\text{inner}} \quad [17]$$

with

$$\tilde{n}_v^{\text{inner}} \stackrel{\text{def}}{=} 4n_v + r_v - n_v^e$$

hold, so that the estimate [16] can be elaborated into

$$I \leq \prod_{v>r} \gamma^{-\rho_v(h_v - h_{v'})}$$
$$\rho_v \stackrel{\text{def}}{=} -d + (4-d)n_v + r_v \frac{d+2}{2} + \frac{d-2}{2} n_v^e \qquad [18]$$

where $h_{v'} = k = 0$ if v is the first nontrivial node (i.e., $v' = \text{root}$), and an estimate of the integral of the absolute value of the graphs G with given tree structure but different scale labels is proportional to $\Sigma_{\{h_v\}} I < \infty$ if (and only if) $\rho_v > 0, \forall v$.

But there may be clusters v with only two external lines $n_v^e = 2$ and two graph vertices inside: for which $\rho_v = 0$. However, this can happen only if $d = 3$ and in only one case: namely if the graph G contains a subgraph of the second type in **Figure 2** and the three intermediate lines form a cluster v of scale h_v while the other two lines are external to it: hence on scale $h' > h_v$. In this case, one has to remember that the subtraction in the previous section has led to a modification of the contribution of such a subgraph to the value of the graph (integrated over the position labels of the vertices). As discussed in the previous section, the change amounts to replacing the propagator $C_{\eta,\beta}^{(h')}$ by $C_{\eta,\beta}^{(h')} - C_{\xi,\beta}^{(h')}$.

This improves, in [18], the estimate of the contribution of the line joining η to β from being proportional to $\int C_{\xi\eta}^{(\leq h_v)3} C_{\eta\beta}^{(\leq h')} \, d\eta$ to being proportional to $\int C_{\xi\eta}^{(\leq h_v)3} (C_{\eta\beta}^{(\leq h')} - C_{\xi\beta}^{(\leq h')}) \, d\eta$; and this changes the contribution of the line $\eta\beta$ from $\gamma^{(d-2)h'}$ to $\int e^{-m\gamma^{h_v}|\xi-\eta|} (\gamma^{h'}|\xi-\eta|)^{1/2} \, d\eta$ because $C^{(h')}$ is regular on scale $\gamma^{-h'} m^{-1}$, see [10] with $\varepsilon = 1/2$.

Since ξ, η are in a cluster of higher scale h_v this means that the estimate is improved by $\gamma^{-(1/2)(h_v - h')}$. In terms of the final estimate, this means that ρ_v in [18] can be improved to $\bar{\rho}_v = \rho_v + 1/2$ for the clusters for which $\rho_v = 0$. Hence, the integrated value of the graph G (after taking also into account the integration over the initially selected vertex ξ_1, trivially giving a further factor $|\Lambda|$ by translation invariance), and summed over the possible scale labels is bounded proportionally to $|\Lambda| \Sigma_{\{h_v\}} I < \infty$ once the estimate of I is improved as described.

Note that the graphs contributing to the perturbation series for $(1/|\Lambda|) \log Z_N(\Lambda, f)$ to order λ^n are finitely many because the number r of external vertices is $r \leq 2n + 2$ (since graphs must be connected). Hence, the perturbation series is finite to all orders in λ.

The above is the renormalizability proof of the scalar φ^4-fields in dimension $d = 2, 3$. The theory is renormalizable even if $d = 4$ as mentioned in the remark at the end of the previous section. The analysis would be very similar to the above: it is just a little more involved power-counting argument.

For more details, the reader is referred to Hepp (1966), Gallavotti (1985), sections 8 and 16.

Asymptotic Freedom ($d = 2, 3$). Heuristic Analysis

Finiteness to all orders of the perturbation expansions is by no means sufficient to prove the existence of the ultraviolet limit for $Z_N(\Lambda, f)$ or for $(1/|\Lambda|) \log Z_N(\Lambda, f)$: and *a priori* it might not even be necessary. For this purpose, the first step is to check uniform (upper and lower) boundedness of $Z_N(\Lambda, f)$ as $N \to \infty$.

The reason behind the validity of a bound $e^{|\Lambda| E_-(\lambda, f)} \leq Z_N(\Lambda, f) \leq e^{|\Lambda| E_+(\lambda, f)}$ with $E_\pm(\lambda, f)$ cutoff independent has been made very clear after the introduction of the renormalization group methods in field theory. The approach studies the integral $Z_N(\Lambda, f)$, recursively, decomposing the field $\varphi_\xi^{(\leq N)}$ into its regular components $z_\xi^{(h)}$, see [7], and integrating first over $z^{(N)}$, then over $z^{(N-1)}$ and so on.

The idea emerges naturally if the potential V_N in [1] and [4] is written in terms of the "normalized" variables $X_\xi^{(N)} \stackrel{\text{def}}{=} \gamma^{-N(d-2)/2} \varphi_\xi^{(\leq N)}$, see [6]; here if $d = 2$ the factor $\gamma^{(d-2)/2N}$ is interpreted as $N^{1/2}$.

The key remark is that as far as the integration over the small-scale component $z^{(N)}$ is concerned the field $X_\xi^{(N)}$ is a sum of two fields of size of order 1 (statistically),

$$X_\xi^{(N)} \equiv z_{\gamma^N \xi}^{(N)} + \gamma^{-(d-2)/2} X_\xi^{(N-1)}$$

if $d = 2$ this becomes

$$X_\xi^{(N)} \equiv \frac{1}{N^{1/2}} z_{\gamma^N \xi}^{(N)} + \frac{(N-1)^{1/2}}{N^{1/2}} X_\xi^{(N-1)}$$

and it can be considered to be smooth on scale $m^{-1} \gamma^{-N}$ (also statistically). Hence, approximately constant and of size of order $O(1)$ on the small cubes Δ of volume $\gamma^{-dN} m^{-d}$ of the pavement $\mathcal{Q}_N$ introduced before [7]; at the same time it can be considered to take (statistically) independent values on different cubes of $\mathcal{Q}_N$. This is suggested by the inequalities [8]–[10].

Therefore, it is natural to decompose the potential V_N, see [5], as a sum over the small cubes Δ of volume $\gamma^{-dN} m^{-d}$ of the pavement $\mathcal{Q}_N$ as (see [14] for the definition of μ_N, ν_N), taking henceforth $m = 1$,

$$V_N(z^{(N)}) \stackrel{\text{def}}{=} -\sum_{\Delta \in \mathcal{Q}_N} \gamma^{-Nd} \int_\Delta \Big(\lambda \gamma^{2(d-2)N} X_\xi^{(N)\,4} + \mu_N \gamma^{(d-2)N} X_\xi^{(N)\,2} + \nu_N + f_\xi \gamma^{(d-2)/2N} X_\xi^{(N)} \Big) \frac{d\xi}{|\Delta|} \qquad [19]$$

where $\gamma^{(d-2)N}$ is interpreted as N if $d=2$. Hence, if $d=3$ it is

$$V_N(z^{(N)}) \stackrel{\text{def}}{=} -\sum_{\Delta\in\mathcal{Q}_N}\gamma^{-N}\int_\Delta\left(\lambda X_\xi^{(N)4}+\bar\mu_N X_\xi^{(N)2}+\bar\nu_N+f_\xi\gamma^{-\frac{3}{2}N}X_\xi^{(N)}\right)\frac{d\xi}{|\Delta|} \qquad [20]$$

where

$$\bar\mu_N \stackrel{\text{def}}{=} (-6\lambda c_N+\lambda^2 N\gamma^{-N}c'_N),$$
$$\bar\nu_N \stackrel{\text{def}}{=} 3\lambda c_N^2+\lambda^2\gamma^{-N}b_N+\lambda^3 N\gamma^{-2N}b'_N$$

and c_N, c'_N, b_N, b'_N, computable from [15] and [14], admit a limit as $N\to\infty$. While if $d=2$ it is

$$V_N(z^{(N)}) \stackrel{\text{def}}{=} -\sum_{\Delta\in\mathcal{Q}_N}N^2\gamma^{-2N}\int_\Delta\left(\lambda X_\xi^{(N)4}+\bar\mu_N X_\xi^{(N)2}+\bar\nu_N+f_\xi N^{-\frac{3}{2}}X_\xi^{(N)}\right)\frac{d\xi}{|\Delta|} \qquad [21]$$

where $\bar\mu_N \stackrel{\text{def}}{=} -6\lambda c_N$ and $\bar\nu_N = 3\lambda c_N^2$ and c_N, computable from [13], admits a limit as $N\to\infty$.

The fields $z^{(N)}$ and $X^{(N-1)}$ can be considered constant over boxes $\Delta\in\mathcal{Q}_N$: $z_\xi^{(N)}=s_\Delta, X_\xi^{(N-1)}=x_\Delta$ for $\xi\in\Delta$ and the s_Δ can be considered statistically independent on the scale of the lattice $\mathcal{Q}_N$.

Therefore, [20] and [21] show that integration over $z^{(N)}$ in the integral defining $Z_N(\Lambda,f)$ is not too different from the computation of a partition function of a lattice continuous spin model in which the "spins" are s_Δ and, most important, interact extremely weakly if N is large. In fact, the coupling constants are of order of a power of $|X^{(N-1)}|$ times $O(\gamma^{-N})$ if $d=3$ ($O(N^2\gamma^{-2N})$ if $d=2$), or of order $O(\gamma^{-N(d+2)/2}\max|f_\xi|)$, *no matter how large* λ *and* f.

This says that the smallest scale fields are extremely weakly coupled. The fields $X^{(N-1)}$ can be regarded as external fields of size that will be called B_{N-1}, of order 1 or even allowed to grow with a power of N, see [6]. Their presence in V_N does not affect the size of the couplings, as far as the analysis of the integral over $z^{(N)}$ is concerned, because the couplings remain exponentially small in N, see [20] and [21], being at worst multiplied by a power of B_{N-1}, i.e., changed by a factor which is a power of N.

The smallness of the coupling at small scale is a property called "asymptotic freedom." Once fields and coordinates are "correctly scaled," the real size of the coupling becomes manifest, that is, it is extremely small and the addends in V_N proportional to the "counter-terms" μ_N, ν_N, which looked divergent when the fields were not properly scaled, are in fact of the same order or much smaller than the main φ^4-term.

Therefore, the integration over $z^{(N)}$ can be, heuristically, performed by techniques well established in statistical mechanics (i.e., by straightforward perturbation expansions): at least if the field $X_\xi^{(\le N-1)}$ is smooth and bounded, as prescribed by [6], with $B=B_{N-1}$ growing as a power of N. In this case, denoting symbolically the integration over $z^{(N)}$ by P or by $\langle\ldots\rangle$, it can be expected that it should give

$$\int e^{V_N}dP\left(z^{(N)}\right)\equiv e^{V_{j;N-1}+\overline{\mathcal{R}}(j,N)|\Delta|} \qquad [22]$$

where $V_{j;N-1}$ is the Taylor expansion of $\log\int e^{V_N}dP(z^{(N)})$ in powers of λ (hence essentially in the very small parameter $\lambda\gamma^{-(4-d)N}$) truncated at order j, that is,

$$V_{1;N-1}=[\langle V_N\rangle]^{\le 1}$$
$$V_{2;N-1}=\left[\langle V_N\rangle+\frac{(\langle V_N^2\rangle-\langle V_N\rangle^2)}{2!}\right]^{\le 2}$$
$$V_{3;N-1}=\left[\langle V_N\rangle+\frac{(\langle V_N^2\rangle-\langle V_N\rangle^2)}{2!}+\frac{\left(\langle V_N(\langle V_N^2\rangle-\langle V_N\rangle^2)\rangle-\langle V_N(\langle V_N^2\rangle-\langle V_N\rangle^2)\right)}{3!}\right]^{\le 3},\ldots \qquad [23]$$

where $[\cdot]^{\le j}$ denotes truncation to order j in λ, and $\overline{\mathcal{R}}(j,N)$ is a remainder (depending on $\varphi_\xi^{(\le N-1)}$) which can be expected to be estimated, for $d=2,3$, by

$$|\overline{\mathcal{R}}(j,N)|\le\mathcal{R}(j,N) \stackrel{\text{def}}{=} C_j B_N^{4j}(\lambda N^2\gamma^{-(4-d)N})^{j+1}\gamma^{dN} \qquad [24]$$

for suitable constants C_j, that is, a remainder estimated by the $(j+1)$th power of the coupling times the number of boxes of scale N in Λ. The relations [22]–[24] result from a naive Taylor expansion (in λ of the $\log\int e^{V_N}dP(z^{(N)})$, taking into account that, in V_N as a function of $z^{(N)}$, the $z^{(N)}$'s appear multiplied by quantities at most of size $\le\lambda\gamma^{4-d}N^2B_N^3$, by [20] and [21] if $|X^{(N-1)}|\le B_{N-1}$). In a statistical mechanics model for a lattice spin system, such a calculation of Z_N would lead to a mean-field equation of state once the remainder was neglected.

The peculiarity of field theory is that a relation like [22] and [24] has to be applied again to $V_{j;N-1}$ to perform the integration over $z^{(N-1)}$ and define $V_{j;N-2}$ and, then, again to $V_{j;N-2}\ldots$. Therefore, it will be essential to perform the integral in [22] to an order (in λ) high enough so that the bound $\mathcal{R}(j,N)$ can be

summed over N: this requires (see [24]) an explicit calculation of [23] pushed at least to order $j=1$ if $d=2$ or to order $j=3$ if $d=3$; furthermore it is also necessary to check that the resulting $V_{j;N-1}$ can still be interpreted as low-coupling spin model so that [22] can be iterated with $N-1$ replacing N and then with $N-2$ replacing $N-1, \ldots$.

The first necessary check towards a proof of the discussed heuristic "expectations" is that, defining recursively $V_{j;h}$ from $V_{j,h+1}$ for $h=N-1,\ldots,1,0$ by [23] with V_N replaced by $V_{j;h+1}$ and $V_{j;N-1}$ replaced by $V_{j;h}$, the couplings between the variables $z^{(h)}$ do not become "worse" than those discussed in the case $h=N$. Furthermore, the field $\varphi_\xi^{(\leq N-1)}$ has a high probability of satisfying [6], but fluctuations are possible: hence the $\mathcal{R}$-estimate has to be combined with another one dealing with the large fluctuations of $X_\xi^{(N-1)}$ which has to be shown to be "not worse."

For more details, the reader is referred to Gallavotti (1978, 1985) and Benfatto and Gallavotti (1995).

Effective Potentials and Their Scale (In)Dependence

To analyze the first problem mentioned at the end of the previous section, define $V_{j;h}$ by [23] with V_N replaced by $V_{j;h+1}$ for $h=N-1, N-2, \ldots, 0$. The quantities $V_{j;h}$, which are called "effective potentials" on scale h (and order j), turn out to be in a natural sense scale independent: this is a consequence of renormalizability, realized by Wilson as a much more general property which can be checked, in the very special cases considered here with $d=2,3$, at fixed j by induction, and in the superrenormalizable models considered here it requires only an elementary computation of a few Gaussian integrals as the case $j=3$ (or even $j=1$ if $d=2$) is already sufficient for our purposes.

It can also be (more easily) proved for general j by a dimensional argument parallel to the one presented earlier to check finiteness of the renormalized series. The derivation is elementary but it should be stressed that, again, it is possible only because of the special choice of the counter-terms μ_N, ν_N. If $d=3$, the boundedness and smoothness of the fields $\varphi^{(\leq h)}$ and $z^{(h)}$ expressed by the second of [6] and of [10] is essential; while if $d=2$ the smoothness is not necessary.

The structure of $V_{j;h}$ is conveniently expressed in terms of the fields $X_\xi^{(h)}$, as a sum of three terms $V_h^{(\mathrm{rel})}$ (standing for "relevant" part), $V_h^{(\mathrm{irr})}$ (standing for "irrelevant" part), and a "field independent" part $E(j,h)|\Lambda|$.

The relevant part in $d=2$ is simply of the form [21] with h replacing N: call it $V_h^{(\mathrm{rel},1)}$. If $d=3$, it is given by [20] with h replacing N plus, for $h<N$, a second "nonlocal" term

$$V_h^{(\mathrm{rel},2)} \stackrel{\mathrm{def}}{=} \frac{4^2\,3!}{2!\,2!}\lambda^2 \int \left(C_{\eta\eta'}^{(\leq h)\,3} - C_{\eta\eta'}^{(\leq N)\,3}\right) \times \left(\varphi_\eta^{(\leq h)} - \varphi_{\eta'}^{(\leq h)}\right)^2 \mathrm{d}\eta\mathrm{d}\eta'$$

which is conveniently expressed in terms of a "nonlocal" field

$$Y_{\eta\eta'}^{(h)} \stackrel{\mathrm{def}}{=} \frac{\varphi_\eta^{(\leq h)} - \varphi_{\eta'}^{(\leq h)}}{(\gamma^h|\eta-\eta'|)^{\frac{1}{4}}}$$

as $V_h^{(\mathrm{rel})} = V_h^{(\mathrm{rel},1)} + V_h^{(\mathrm{rel},2)}$ with

$$V_h^{(\mathrm{rel},2)} \stackrel{\mathrm{def}}{=} -\lambda^2\gamma^{-2h} \sum_{\Delta,\Delta'\in\mathcal{Q}_h} \int_{\Delta\times\Delta'} Y_{\eta\eta'}^{(h)2}\, A_{\eta\eta'}^{(h)} \times \mathrm{e}^{-c'\gamma^h|\eta-\eta'|}\, \frac{\mathrm{d}\eta\mathrm{d}\eta'}{|\Delta|\,|\Delta'|} \qquad [25]$$

where

$$0 < a \leq \left(\frac{A_{\eta\eta'}^{(h)}}{(\gamma^h|\eta-\eta'|)^{3-(1/2)}}\right)_N < a'$$

with $a, a', c' > 0$ and the subscript N means that the expression in parenthesis "saturates at scale N", i.e., its denoninator becomes $\gamma^{(3-(1/2))(h-N)}$ as $|\eta-\eta'|\to 0$.

The expression [25] is not the full part of the potential $V_{j;h}$ which is of second order in the fields: there are several other contributions which are collected below as "irrelevant."

It should be stressed that "irrelevant" is a traditional technical term: by no means it should suggest "negligibility." On the contrary, it could be maintained that the whole purpose of the theory is to study the irrelevant terms. The irrelevant part of the potential can be better designated as the "driven part," as its behavior is "controlled" by the relevant part: although initially $V_{j;h}$, $h=N$, contains no irrelevant terms, it eventually contains them for $h<N$ and they keep getting generated as h diminishes. Furthermore, the part of the irrelevant terms generated at scale $h_0 \leq N$ becomes very small at scales $h \ll h_0$ so that the irrelevant part of $V_{j;h}$ at small h (e.g., at $h=0$, i.e., on the "physical scale" of the observer) only depend on the relevant terms in a few scales near h.

It also turns out that the Schwinger functions are simply related to the irrelevant terms.

The irrelevant part of the effective potential can be expressed as a finite sum of integrals of

monomials in the fields $X_\xi^{(h)}$ if $d=2$, or in the fields $X_\xi^{(h)}$ and $Y_{\eta\eta'}^{(h)}$ if $d=3$, which can be written as $V_{j;h}^{(\mathrm{irr})}$ given by

$$\int \left(\prod_{k=1}^{p} X_{\xi_k}^{(h)\,n_k} \prod_{k'=1}^{q} Y_{\eta_{k'}\eta'_{k'}}^{(h)\,n'_{k'}} \right) e^{-\gamma^h c' d(\xi_1,\ldots,\eta'_q)} \lambda^n \gamma^{-ht} \times W(\xi_1,\ldots,\eta'_q) \prod_{k=1}^{p} \frac{d\xi_k}{|\Delta_k|} \prod_{k'=1}^{q} \frac{d\eta_{k'} d\eta'_{k'}}{|\Delta_{k'}^1|\,|\Delta_{k'}^2|} \qquad [26]$$

with the integral extended to products $\Delta_1 \times \cdots \times \Delta_p \times \cdots \times (\Delta_q^1 \times \Delta_q^2)$ of boxes $\Delta \in \mathcal{Q}_h$, and $d(\xi_1,\ldots,\eta'_q)$ is the length of the shortest tree graph that connects all the $p+2q>0$ points, the exponents n,t are ≥ 2, and t is ≥ 3 if $q>0$; the kernel W depends on all coordinates $\xi_1,\ldots,\eta'_q$ and it is bounded above by $C_j \prod_{k'=1}^{q} A_{\eta_{k'}\eta'_{k'}}$ for some C_j; the sums $\sum n_k + \sum n'_{k'}$ cannot exceed $4j$. The test functions f do not appear in [26] because by assumption they are bounded by 1: but W depends on the f's as well.

The field-independent part is simply the value of $\log Z_N(\Lambda,f)$ computed by the perturbation analysis in the section "Perturbation theory" up to order j in λ but using as propagator $(C^{(\leq N)} - C^{(\leq h)})$: thus, $E(j,h)$ is a constant depending on N but uniformly bounded as $N \to \infty$ (because of the renormalizability proved in the section "Perturbation theory").

If $d=2$, there is no need to introduce the nonlocal fields $Y^{(h)}$ and in [26] one can simply take $q=0$, and the relevant part also can be expressed by omitting the term $V_h^{(\mathrm{rel},2)}$ in [25]: unlike the $d=3$ case, the estimate on the kernels W by an N-independent C_j holds uniformly in h without having to introduce Y. For $d=2$, it will therefore be supposed that $V_h^{(\mathrm{rel},2)} \equiv 0$ in [25] and $q=0$ in [26].

It is not necessary to have more information on the structure of $V_{j;h}$ even though one can find simple graphical rules, closely related to the ones in the section "Perturbation theory," to construct the coefficients W in full detail. The W depend, of course, on h but the uniformity of the bound on W is the only relevant property and in this sense the effective potentials are said to be (almost) "scale independent."

The above bounds on the irrelevant part can be checked by an elementary direct computation if $j \leq 3$: in spite of its "elementary character," the uniformity in $h \leq N$ is a result ultimately playing an essential role in the theory together with the dominance of the relevant part over the irrelevant one which, once the fields are properly scaled, is "much smaller" (by a factor of order γ^{-h}, see [26]), at least if h is large.

Remarks

(i) Checking scale independence for $j=1$ is just checking that $\int P(dz^{(h)}) V_{1;h} = V_{1;h-1}$. Note that

$$V_{1;h} \stackrel{\mathrm{def}}{=} \int_\Lambda \lambda \left(\varphi_\xi^{(\leq h)4} - 6C_{00}^{(\leq h)} \varphi_\xi^{(\leq h)2} + 3C_{00}^{(\leq h)2} \right) d\xi$$

hence, calling $:\varphi_\xi^{(\leq h)4}:$ the polynomial in the integral (Wick's monomial of order 4), we have here an elementary Gaussian integral ("martingale property of Wick monomials"). Note the essential role of the counter-terms. For $j>1$, the computation is similar but it involves higher-order polynomials (up to $4j$) and the distinction between $d=2$ and $d=3$ becomes important.

(ii) $V_{j;0}$ contains only the field-independent part $E(j,0)|\Lambda|$ (see above) which is just a number (as there are no fields of scale 0): by the above definitions, it is identical to the perturbative expansion truncated to jth order in λ of $\log Z_N(\Lambda,f)$, well defined as discussed earlier.

Nonperturbative Renormalization: Small Fields

Having introduced the notion of effective potential $V_{j;h}$, of order j and scale h, satisfying the bounds (described after [26]) on the kernels W representing it, the problem is to estimate the remainder in [22] and find its relation with the value [24] given by the heuristic Taylor expansion. Assume $\lambda < 1$ to avoid distinguishing this case from that with $\lambda \geq 1$ which would lead to very similar estimates but to different λ-dependence on some constants.

Define $\chi_B(z^{(h)}) = 1$ if $\|z^{(h)}\|_\Delta \leq Bh^2$ for all $\Delta \in \mathcal{Q}_h$, see [8], and 0 otherwise; then the following lemma holds:

Lemma 1 *Let $\|X^{(h)}\|_\Delta$ be defined as [8] with z replaced by X and suppose $\|X^{(h)}\|_\Delta \leq Bh^4$ for all Δ then, for all $j \geq 1$, it is*

$$\int e^{V_{j,h+1}} \chi_B(z^{(h+1)}) dP(z^{(h+1)}) = e^{V_{j,h} + \mathcal{R}'(j,h+1)|\Lambda|} \qquad [27]$$

with, for suitable constants c_-, c'_-,

$$|\mathcal{R}'_-(j;h+1)| \leq \mathcal{R}_-(j;h+1) \stackrel{\mathrm{def}}{=} \mathcal{R}(j;h+1) + c_- e^{-c'_- B^2 (h+1)^2}$$

and $\mathcal{R}(j;h+1)$ given by [24] with $h+1$ in place of N.

Since $Z_N(\Lambda,f) \geq \int e^{V_N} \prod_{h=1}^{N} \chi_B(z^{(h)}) P(dz^{(h)})$ this immediately gives a lower bound on $E=(1/|\Lambda|) \log Z_N(\Lambda,f)$: in fact if $\chi_B(\|z^{(h')}\|)=1$ for

$h' = 1, \ldots, h$, then $\|X^{(h)}\|_\Delta \leq c\,Bh'^4$ for some c so that, by recursive application of Lemma 1, $Z_N(\Lambda, f) \geq e^{V_{j,0} - \sum_{h=1}^{N} \mathcal{R}_-(j,h)|\Lambda|}$. By the remark at the end of the previous section, given j the lower bound on E just described agrees with the perturbation expansion of $E = (1/|\Lambda|) \log Z_N(\Lambda, f)$ truncated to order j (in λ) up to an error bounded by $\sum_{h=1}^{\infty} \mathcal{R}_-(j,h)$.

Remark The problem solved by Lemma 1 is usually referred to as the small-field problem, to contrast it with the large-field problem discussed later. The proof of the lemma is a simple Taylor expansion in $\lambda\gamma^{-h}$ if $d=3$ or in $\lambda h^2 \gamma^{-2h}$ if $d=2$ to order j (in λ). The constraint on $z^{(h+1)}$ makes the integrations over $z^{(h+1)}$, necessary to compute $V_{j;h}$ from $V_{j;h+1}$, not Gaussian. But the tail estimates [9], together with the Markov property of the distribution of $z^{(h)}$ can be used to estimate the difference with respect to the Gaussian unconstrained integrations of $z^{(h+1)}$: and the result is the addition of the small "tail error" changing $\mathcal{R}$ into $\mathcal{R}_-$ in [27]. The estimate of the main part of the remainder $\mathcal{R}$ would be obvious if the fields $z^{(h)}$ were independent on boxes of scale γ^{-h}: they are not independent but they are Markovian and the estimate can be done by taking into account the Markov property.

For more details, the reader is referred to Wilson (1970, 1972), Gallavotti (1978, 1981, 1985), and Benfatto *et al.* (1978).

Nonperturbative Renormalization: Large Fields, Ultraviolet Stability

The small-field estimates are not sufficient to obtain ultraviolet stability: to control the cases in which $|X_\xi^{(h)}| > Bh^4$ for some ξ or some h, or $|Y_{\xi\eta}^{(h)}| > Bh^4$ for some $|\xi - \eta| < \gamma^{-h}$, a further idea is necessary and it rests on making use of the assumption that $\lambda > 0$ which, in a sense to be determined, should suppress the contribution to the integral defining $Z_N(\Lambda, f)$ coming from very large values of the field. Assume also $\lambda < 1$ for the same reasons advanced in the section "Effective potentials and their scale (in)dependence."

Consider first $d=2$. Let $\mathcal{D}_N$ be the "large-field region" where $|X_\xi^{(N)}| > BN^4$ and let $V_N(\Lambda/\mathcal{D}_N)$ be the integral defining the potential in [21] extended to the region $\Lambda/\mathcal{D}_N$, complement of $\mathcal{D}_N$. This region is typically very irregular (and random as X itself is random with distribution P_N).

An upper bound on the integral defining $Z_N(\Lambda, f)$ is obtained by simply replacing e^{V_N} by $e^{V_N(\Lambda/\mathcal{D}_N)}$ because in $\mathcal{D}_N$ the first term in the integrand in [21] is $\leq -\lambda N^2 \gamma^{2N}(BN^4) < 0$ and it *overwhelmingly* dominates on the remaining terms whose value is bounded by a similar expression with a smaller power of N. Then if $\mathcal{E}^c \stackrel{\text{def}}{=} \Lambda/\mathcal{E}$ denotes the complement in Λ of a set $\mathcal{E} \subset \Lambda$:

Lemma 2 *Let $d=2$. Define $V_h(\mathcal{D}_h^c)$ to be given by the expression* [22] *with the integrals extending over $\Delta_j/\mathcal{D}_h$ and define $\mathcal{R}(j, h+1)$ by* [24]. *Then*

$$\int e^{V_{h+1}(\mathcal{D}_{h+1}^c)}\, dP\left(z^{(h+1)}\right) = e^{V_h(\mathcal{D}_h^c) + \overline{\mathcal{R}}_+(j,h+1)|\Lambda|} \qquad [28]$$

where $|\overline{\mathcal{R}}_+(j, h+1| \leq \mathcal{R}_+(j, h+1 \stackrel{\text{def}}{=} \mathcal{R}(j; h+1) + c_+ e^{-c'_+ B^2 (h+1)^2}$ *with suitable* c_+, c'_+.

Remark Lemma 2 is genuinely not perturbative and making essential use of the positivity of λ. Below the analysis of the proof of the lemma, which consists essentially in its reduction to Lemma 1, is described in detail. It is perhaps the most interesting part and the core of the theory of the proof that truncating the expansion in λ of $(1/|\Lambda|) \log Z_N(\Lambda, f)$ to order j gives as a result an estimate exact to order λ^{j+1} of $(1/|\Lambda|) \log Z_N(\Lambda, f)$.

Let R_N be the cubes $\Delta \in \mathcal{Q}_N$ in which there is at least one point ξ where $|z_\xi^{(N)}| \geq BN^2$. By definition, the region $\mathcal{D}_N/\mathcal{D}_{N-1}$ is covered by R_N.

Remark that in the region $\mathcal{D}_{N-1}/R_N$ the field $X^{(N-1)}$ is large but z_N is not large so that $X^{(N)}$ is still very large: this is so because the bounds set to define the regions $\mathcal{D}$ and R are quite different being BN^4 and BN^2, respectively. Hence, if a point is in $\mathcal{D}_{N-1}$ and not in R_N, then the field $X^{(N)}$ must be of the order $\gg BN^3$. Therefore, by positivity of the $\lambda\varphi_\xi^{(\leq N)4}$ term (which dominates all other terms so that $V^{(N)}(\varphi_\xi^{(\leq N)}) < 0$ for $\xi \in \mathcal{D}_N \cup (\mathcal{D}_{N-1}/R_N)$) we can replace $V_N(\mathcal{D}_N^c)$ by $V((\mathcal{D}_N \cup (\mathcal{D}_{N-1}/R_N))^c)$, for the purpose of obtaining an upper bound.

Furthermore, modulo a suitable correction, it is possible to replace $V((\mathcal{D}_N \cup (\mathcal{D}_{N-1}/R_N))^c)$ by $V((\mathcal{D}_{N-1} \cup R_N)^c)$: because the integrand in V_N is bounded below by

$$-b\lambda\gamma^{-2N}N^2$$

if $d=2$ (by $-b\lambda\gamma^{-N}$ if $d=3$), for some b, so that the points in R_N can at most lower $V((\mathcal{D}_N \cup (\mathcal{D}_{N-1}/R_N))^c)$ by $-b\lambda N^2\, \gamma^{-(4-d)N}\, \#(R_N)$ if $\#R_N$ is the number of boxes of $\mathcal{Q}_N$ in R_N and $V(\varphi_\xi)$ is bounded below by its minimum: thus,

$$V((\mathcal{D}_{N-1} \cup R_N)^c) + b\lambda N^2 \gamma^{(4-d)N} \#(R_N)$$

is an upper bound to $V((\mathcal{D}_N \cup (\mathcal{D}_{N-1}/R_N))^c)$.

In the complement of $\mathcal{D}_{N-1} \cup R_N$, all fields are "small"; if $X^{(N-1)}$ and R_N are fixed this region is not random (as a function of $z^{(N)}$) any more. Therefore,

if $X^{(N-1)}, R_N$ are fixed the integration over $z^{(N)}$, conditioned to having $z^{(N)}$ fixed (and large) in the region R_N, is performed by means of the same argument necessary to prove Lemma 1 (essentially a Taylor expansion in $\lambda\gamma^{-(4-d)N}$). The large size of $z^{(N)}$ in R_N does not affect too much the result because on the boundary of R_N the field $z^{(N)}$ is $\leq BN^2$ (recalling that $z^{(N)}$ is continuous) and since the variable $z^{(N)}$ is Markovian, the boundary effect decays exponentially from the boundary ∂R_N: it adds a quantity that can be shown to be bounded by the number of boxes in R_N on the boundary of R_N, hence by $\#R_N$, times $b'(N-1)^2\gamma^{-(4-d)}(B(N-1)^4)^4$ for some b'.

The result of the integration over $z^{(N)}$ of $e^{V_N((\mathcal{D}_N \cup (\mathcal{D}_{N-1}/R_N))^c)}$ conditioned to the large-field values of $z^{(N)}$ in R_N leads to an upper bound on $\int e^{V_N} P(dz^{(N)})$ as

$$\sum_{R_N} e^{V_{j;N-1}(\mathcal{D}^c_{N-1})+\mathcal{R}(j,N)|\Lambda|} \times \prod_{\Delta\in R_N} \left(c e^{-c'(BN^2)^2} e^{+c''\lambda\gamma^{-(4-d)N}N^2(BN^4)^4} \right)^{\#R_N} \qquad [29]$$

where c, c', c'' are suitable constants: this is explained as follows.

1. Taylor expansion (in λ) of the integral $e^{V_N((\mathcal{D}_{N-1}\cup R_N)^c)+b\lambda N^2\gamma^{-(4-d)N}\#(R_N)}$ (which, by construction, is an upper bound on $e^{V_N(\mathcal{D}^c_N)}$) with respect to the field $z^{(N)}$, conditioned to be fixed and large in R_N, would lead to an upper bound as

$$e^{V_{j;N-1}((\mathcal{D}_{N-1}\cup R_N)^c)+\mathcal{R}'(j,N)|\Lambda|+b''\lambda(BN^4)^4\gamma^{(4-d)N}\#(R_N)}$$

with $\mathcal{R}'$ equal to [24] possibly with some C'_j replacing C_j. The second exponential on the RHS of [29] arises partly from the above correction $b''\lambda(BN^4)^4\gamma^{-(4-d)N}\#(R_N)$ and partly from a contribution of similar form explained in (3) below.
2. Integration over the large conditioning fields fixed in R_N is controlled by the second estimate in [9] (the tail estimate): the first factors in parentheses in [29] is the tail estimate just mentioned, i.e., the probability that $z^{(N)}$ is large in the region R_N. The second factor is only partly explained in (1) above.
3. Without further estimates, the bound [29] would contain $V_{j;N-1}((\mathcal{D}_{N-1}\cup R_N)^c)$ rather than $V_{j;N-1}(\mathcal{D}^c_{N-1})$. Hence, there is the need to change the potential $V_{j;N-1}((\mathcal{D}_{N-1}\cup R_N)^c)$ by "reintroducing" the contribution due to the fields in $R_N/\mathcal{D}_{N-1}$ in order to reconstruct $V_{j;N-1}(\mathcal{D}^c_{N-1})$. Reintroducing this part of the potential costs a quantity like $b'\lambda N^2\gamma^{(4-d)N}(BN^4)^4\#(R_N)$ (because the reintroduction occurs in the region $R_N/\mathcal{D}_{N-1}$ which is covered by R_N and in such points the field $X^{(N-1)}_\xi$ is *not large*, being bounded by $B(N-1)^4$); so that their contribution to the effective potential is still dominated by the φ^4-term and therefore by $\gamma^{-(4-d)N}$ times a power of BN^4 times the volume of R_N (in units γ^{-N}, i.e., $\#R_N$). All this is taken care of by suitably fixing c''.

Note that the sum over R_N of [29] is

$$(1 + c e^{-c'B^2N^4} e^{+c''\lambda\gamma^{-(4-d)N}N^2(BN^4)^4})^{\gamma^{dN}|\Lambda|}$$

(because Λ contains $|\Lambda|\gamma^{dN}$ cubes of $\mathcal{Q}_N$); hence, it is bounded above by $e^{c_+ e^{-c'_+ B^2N^2}}$ for suitably defined c_+, c'_+.

The same argument can be repeated for $V_{j;h}(\mathcal{D}^c_h)$ with any h if $V_{j;h}(\mathcal{D}^c_h)$ is defined by the sum over Δ's in $\mathcal{Q}_h$ of the same integrals as those in [25] and [26] with $\Delta_j/\mathcal{D}_h$ replacing Δ_j in the integration domains.

Applying Lemma 1 recursively with $j \geq 1$ (if $d=3$ it would become necessary to take $j \geq 3$), it follows that there exist N-independent upper and lower bounds $E_\pm|\Lambda|$ on $\log Z(\Lambda, f)$ of the form $V_{j;0} \pm \sum_{h=1}^{\infty}(\mathcal{R}(j,h) + c_\pm e^{-c'_\pm B^2h^2})|\Lambda|$ for $c_\pm, c'_\pm > 0$ suitably chosen and λ-independent for $\lambda < 1$. By the remark at the end of Sec.6, given j, the bounds just described agree with the perturbation expansion $E(j,0)|\Lambda| \equiv V_{j;0}$ of $\log Z(\Lambda, f)$ truncated to order j (in λ) up to the remainders $\pm\sum_{h=1}^{\infty}\mathcal{R}_\pm(j,h)$. Hence, if B is chosen proportional to $\log_+ \lambda^{-1} \stackrel{\text{def}}{=} \log(e+\lambda^{-1})$, the upper and lower bounds coincide to order j in λ with the value obtained by truncating to order j the perturbative series.

The latter remark is important as it implies not only that the bounds are finite (by the section "Perturbation theory") but also that the function $(1/|\Lambda|)\log Z(\Lambda, f)$ is *not* quadratic in f: already to order 1 in λ it is quartic in f (containing a term equal to $-\lambda(\int C_{\xi,0} f_\xi d\xi)^4$).

The latter property is important as it excludes that the result is a "Gaussian" generating function. Thus, the outline of the proof of Lemma 2, which together with Lemma 1 forms the core of the analysis of the ultraviolet stability for $d=2$, is completed.

If $d=3$, more care is needed because (very mild) smoothness, like the considered Hölder continuity with exponent 1/4, of z, X is necessary to obtain the key scale independence property discussed in earlier: therefore, the natural measure of the size of $z^{(h)}$ and $X^{(h)}$ in a box $\Delta \in \mathcal{Q}_h$ is no longer the maximum of $|z^{(h)}_\xi|$ or of $|X^{(h)}_\xi|$. The region $\mathcal{D}_h$ becomes more

involved as it has to consist of the points ξ where $|X_\xi^{(h)}| > Bh^4$ *and* of the pairs η, η' where

$$|Y_{\eta,\eta'}| \equiv \frac{\left|X_\eta^{(h)} - X_{\eta'}^{(h)}\right|}{(\gamma^h|\eta - \eta'|)^{\frac{1}{2}}} > Bh^4$$

i.e., it is not just a subset of Λ.

However, if $d = 3$, the relevant part also contains the negative term $V^{(\mathrm{rel},2)}$, see [25]: and since it dominates over all other terms which contain a Y-field (because their couplings [25] are smaller by about γ^{-h}), the argument given for $d = 2$ can be adapted to the new situation. Two regions $\mathcal{D}_h^1, \mathcal{D}_h^2$ will be defined: the first consists of all the points ξ where $|X_\xi^{(h)}| > Bh^4$ and the second of all the pairs η, η' where $|Y_{\eta,\eta'}^{(h)}| > Bh^4$. The region R_h will be the collection of all $\Delta \in \mathcal{Q}_h$, where $\|z^{(h)}\|_\Delta > Bh^2$, see [8] with $\tau = 0$. Then $V(\mathcal{D}_h^c)$ will be defined as the sum of the integrals in [25] and [26] with the integrals over ξ_i further restricted to $\xi_i \notin \mathcal{D}_h^1$ and those over the pairs η_i, η_i' are further restricted to $(\eta_i, \eta_i') \notin \mathcal{D}_h^2$. With the new settings, Lemma 2 can be proved also for $d = 3$ along the same lines as in the $d = 2$ case.

For more details, the reader is referred to Wilson (1970, 1972), Benfatto *et al.* (1978), and Gallavotti (1981).

Ultraviolet Limit, Infrared Behavior, and Other Applications

The results on the ultraviolet stability are nonperturbative, as no assumption is made on the size of λ (the assumption $\lambda < 1$ has been imposed in the last two sections only to obtain simpler expressions for the λ-dependence of various constants): nevertheless the multiscale analysis has allowed us to use perturbative techniques (i.e., the Taylor expansion in Lemmata 1, 2) to find the solution. The latter procedure is the essence of the renormalization group methods: they aim at reducing a difficult multiscale problem to a sequence of simple single-scale problems. Of course, in most cases, it is difficult to implement the approach and the scalar quantum fields in dimensions 2, 3 are among the simplest examples. The analysis of the beta function and of the running couplings, which appear in essentially all renormalization group applications, does not play a role here (or, better, their role is so inessential that it has even been possible to avoid mentioning them). This makes the models somewhat special from the renormalization group viewpoint: the running couplings at length scale h, if introduced, would tend exponentially to 0 as $h \to \infty$; unlike what happens in the most interesting renormalization group applications in which they either tend to zero only as powers of h or do not tend to zero at all.

The multiscale analysis method, i.e., the renormalization group method, in a form close to the one discussed here has been applied very often since its introduction in physics and it has led to the solution of several important problems. The following is not an exhaustive list and includes a few open questions.

1. The arguments just discussed imply, with minor extra work that $Z_N(\Lambda, f)$ as $N \to \infty$ not only admit uniform upper and lower bounds but also that the limit as $N \to \infty$ actually exists and it is a C^∞ function of λ, f. Its λ and f-derivatives at $\lambda = 0$ and $f = 0$ are given by the formal perturbation calculation. In some cases; it is even possible to show that the formal series for $Z_N(\Lambda, f)$ in powers of λ is Borel summable.
2. The problem of removing the infrared cutoff (i.e., $\Lambda \to \infty$) is in a sense more a problem of statistical mechanics. In fact, it can be solved for $d = 2, 3$ by a typical technique used in statistical mechanics, the "cluster expansion." This is not intended to mean that it is technically an easy task: understanding its connection with the low-density expansions and the possibility of using such techniques has been a major achievement that is not discussed here.
3. The third problem mentioned in the introduction, that is, checking the axioms so that the theory could be interpreted as a quantum field theory is a difficult problem which required important efforts to control and which is not analyzed here. An introduction to it can be its analysis in the $d = 2$ case.
4. Also the problem of keeping the ultraviolet cutoff and removing the infrared cutoff while the parameter m^2 in the propagator approaches 0 is a very interesting problem related to many questions in statistical mechanics at the critical point.
5. Field theory methods can be applied to various statistical mechanics problems away from criticality: particularly interesting is the theory of the neutral Coulomb gas and of the dipole gas in two dimensions.
6. The methods can be applied to Fermi systems in field theory as well as in equilibrium statistical mechanics. The understanding of the ground state in not exactly soluble models of spinless fermions in one dimension at small coupling is one of the results. And via the transfer matrix theory it has led to the understanding of nontrivial critical behavior in two-dimensional models that are not exactly soluble (like Ising next-nearest-neighbor or Ashkin–Teller model). Fermi systems are of particular interest also because in their analysis the large-fields problem is absent, but this great

technical advantage is somewhat offset by the anticommutation properties of the fermionic fields, which do not allow us to employ probabilistic techniques in the estimates.

7. An outstanding open problem is whether the scalar φ^4-theory is possible and nontrivial in dimension $d=4$: this is a case of a renormalizable not asymptotically free theory. The conjecture that many support is that the theory is necessarily trivial (i.e., the function $Z_N(\Lambda, f)$ becomes necessarily a Gaussian in the limit $N \to \infty$). One of the main problems is the choice of the ultraviolet cut-off; unlike the $d=2,3$ cases in which the choice is a matter of convenience it does not seem that the issue of triviality can be settled without a careful analysis of the choice and of the role of the ultraviolet cut-off.
8. Very interesting problems can be found in the study of highly symmetric quantum fields: gauge invariance presents serious difficulties to be studied (rigorously or even heuristically) because in its naive forms it is incompatible with regularizations. Rigorous treatments have been in some cases possible and in few cases it has been shown that the naive treatment is not only not rigorous but it leads to incorrect results.
9. In connection with item (8) an outstanding problem is to understand relativistic pure gauge Higgs fields in dimension $d=4$: the latter have been shown to be ultraviolet stable but the result has not been followed by the study of the infrared limit.
10. The classical gauge theory problem is quantum electrodynamics, QED, in dimension 4: it is a renormalizable theory (taking into account gauge invariance) and its perturbative series truncated after the first few orders give results that can be directly confronted with experience, giving very accurate predictions. Nevertheless, the model is widely believed to be incomplete: in the sense that, if treated rigorously, the result would be a field describing free noninteracting assemblies of photons and electrons. It is believed that QED can make sense only if embedded in a model with more fields, representing other particles (e.g., the standard model), which would influence the behavior of the electromagnetic field by providing an effective ultraviolet cutoff high enough for not altering the predictions on the observations on the time and energy scales on which present (and, possibly, future over a long time span) experiments are performed. In dimension $d=3$, QED is super-renormalizable, once the gauge symmetry is properly taken into account, and it can be studied with the techniques described above for the scalar fields in the corresponding dimension.

In general, constructive quantum field theory seems to be in a deep crisis: the few solutions that have been found concern very special problems and are very demanding technically; the results obtained have often not been considered to contribute appreciably to any "progress." And many consider that the work dedicated to the subject is not worth the results that one can even hope to obtain. Therefore, in recent years, attempts have been made to follow other paths: an attitude that in the past usually did not lead, in general to great achievements but that is always tempting and worth pursuing because the rare major progresses made in physics resulted precisely by such changes of attitude, leaving aside developments requiring work which was too technical and possibly hopeless: just to mention an important case, one can recall quantum mechanics which disposed of all attempts at understanding the observed atomic levels quantization on the basis of refined developments of classical electromagnetism.

For more details, the reader is referred to Nelson (1966), Guerra (1972), Glimm *et al.* (1973), Glimm and Jaffe (1981), Simon (1974), Benfatto *et al.* (1978, 2003), Aizenman (1982), Gawedzky and Kupiainen (1983, 1985a, b), Balaban (1983), and Giuliani and Mastropietro (2005).

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Euclidean Field Theory; Integrability and Quantum Field Theory; Perturbation Theory and its Techniques; Quantum Field Theory: A Brief Introduction; Scattering, Asymptotic Completeness and Bound States.

Further Reading

Aizenman M (1982) Geometric analysis of φ^4-fields and Ising models. *Communications in Mathematical Physics* 86: 1–48.

Balaban T (1983) $(\text{Higgs})_{3,2}$ quantum fields in a finite volume. III. Renormalization. *Communications in Mathematical Physics* 88: 411–445.

Benfatto G, Cassandro M, Gallavotti G *et al.* (1978) Some probabilistic techniques in field theory. *Communications in Mathematical Physics* 59: 143–166.

Benfatto G, Cassandro M, Gallavotti G *et al.* (1980) Ultraviolet stability in Euclidean scalar field theories. *Communications in Mathematical Physics* 71: 95–130.

Benfatto G and Gallavotti G (1995) *Renormalization Group*, pp. 1–143. Princeton: Princeton University Press.

Benfatto G, Giuliani A, and Mastropietro V (2003) Low temperature analysis of two dimensional Fermi systems with symmetric Fermi surface. *Annales Henry Poincaré* 4: 137–193.

De Calan C and Rivasseau V (1981) Local existence of the Borel transform in euclidean ϕ_4^4. *Communications in Mathematical Physics* 82: 69–100.

Fröhlich J (1982) On the triviality of $\lambda\phi_d^4$ theories and the approach to the critical point in $d \geq 4$ dimensions. *Nuclear Physics B* 200: 281–296.

Gallavotti G (1978) Some aspects of renormalization problems in statistical mechanics. *Memorie dell' Accademia dei Lincei* 15: 23–59.

Gallavotti G (1981) Elliptic operators and Gaussian processes. In: *Aspects Statistiques et Aspects Physiques des Processus Gaussiens*, pp. 349–360. Colloques Internat. C.N.R.S, St. Flour. Publications du CNRS, Paris.

Gallavotti G (1985) Renormalization theory and ultraviolet stability via renormalization group methods. *Reviews of Modern Physics* 57: 471–569.

Gawedzky K and Kupiainen A (1983) Block spin renormalization group for dipole gas and $(\partial\phi)^4$. *Annals of Physics* 147: 198–243.

Gawedzky K and Kupiainen A (1985a) Gross–Neveu model through convergent perturbation expansion. *Communications in Mathematical Physics* 102: 1–30.

Gawedzky K and Kupiainen A (1985b) Massless lattice ϕ_4^4 theory: rigorous control of a renormalizable asymptotically free model. *Communications in Mathematical Physics* 99: 197–252.

Giuliani A and Mastropietro V (2005) Anomalous universality in the anisotropic Ashkin–Teller model. *Communications in Mathematical Physics* 256: 681–735.

Glimm J, Jaffe A, and Spencer T (1973) Velo G and Wightman A (eds.) *Constructive Field theory*, Lecture Notes in Physics, vol. 25, pp. 132–242. New York: Springer.

Glimm J and Jaffe A (1981) *Quantum Physics*. Springer.

Guerra F (1972) Uniqueness of the vacuum energy density and Van Hove phenomena in the infinite volume limit for two-dimensional self-coupled Bose fields. *Physical Review Letters* 28: 1213–1215.

Hepp K (1966) *Théorie de la rénormalization*. Lecture Notes in Physics, vol. 2. Heidelberg: Springer.

Nelson E (1966) A quartic interaction in two dimensions. In: Goodman R and Segal I (eds.) *Mathematical Theory of Elementary Particles*, pp. 69–73. Cambridge: M.I.T.

Osterwalder K and Schrader R (1973) Axioms for Euclidean Green's functions. *Communications in Mathematical Physics* 31: 83–112.

Simon B (1974) *The $P(\varphi)_2$ Euclidean (Quantum) Field Theory*. Princeton: Princeton University Press.

Streater RF and Wightman AS (1964) *PCT, Spin, Statistics and All That*. Benjamin-Cummings (reprinted Princeton University Press, 2000).

Wightman AS and Gärding L (1965) Fields as operator-valued distributions in relativistic quantum theory. *Arkiv för Fysik* 28: 129–189.

Wilson KG (1970) Model of coupling constant renormalization. *Physical Review D* 2: 1438–1472.

Wilson KG (1972) Renormalization of a scalar field in strong coupling. *Physical Review D* 6: 419–426.

Current Algebra

G A Goldin, Rutgers University, Piscataway, NJ, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Certain commutation relations among the current density operators in quantum field theories define an infinite-dimensional Lie algebra. The original current algebra of Gell-Mann described weak and electromagnetic currents of the strongly interacting particles (hadrons), leading to the Adler–Weisberger formula and other important physical results. This helped inspire mathematical and quantum-theoretic developments such as the Sugawara model, light cone currents, Virasoro algebra, the mathematical theory of affine Kac–Moody algebras, and non-relativistic current algebra in quantum and statistical physics. Lie algebras of local currents may be the infinitesimal representations of loop groups, local current groups or gauge groups, diffeomorphism groups, and their semidirect products or other extensions. Broadly construed, current algebra thus leads directly into the representation theory of infinite-dimensional groups and algebras. Applications have ranged across conformally invariant field theory, vertex operator algebras, exactly solvable lattice and continuum models in statistical physics, exotic particle statistics and q-commutation relations, hydrodynamics and quantized vortex motion. This brief survey describes but a few highlights.

Relativistic Local Current Algebra for Hadrons

To 'model superfluidity, Landau had proposed in 1941 a quantum hydrodynamics fundamentally based on local fluid densities and currents as (operator) dynamical variables. However, current algebra came into its own in theoretical physics with the ideas of Gell-Mann in the early 1960s. The basic concept, in the era just preceding quantum chromodynamics (QCD), was that even without knowing the Lagrangian governing hadron dynamics in detail, exact kinematical information – the local symmetry – could still be encoded in an algebra of currents. The local (vector and axial vector) current density operators, expressed where possible in terms of underlying quantized field operators in Hilbert space, were to form two octets of Lorentz 4-vectors, with each octet corresponding to the eight generators of the compact Lie group SU(3).

More specifically (Adler and Dashen 1968), let $\mathcal{F}_a^\mu(x), a=1,2,\ldots,8$, $\mu=0,1,2,3$, be an octet of hadronic vector currents, where as usual $x=(x^\nu)=(x^0,\boldsymbol{x})$ denotes a point in four-dimensional spacetime. Likewise, introduce an axial vector octet $\mathcal{F}_a^{5\mu}(x)$. Unless otherwise specified, we use natural units, where $\hbar=1$ and $c=1$. Define the corresponding charges F_a and F_a^5 to be the space integrals of the time components of these currents, that is,

$$\begin{aligned} F_a(x^0) &= \int \mathrm{d}^3x\,\mathcal{F}_a^0(x^0,\boldsymbol{x}) \\ F_a^5(x^0) &= \int \mathrm{d}^3x\,\mathcal{F}_a^{50}(x^0,\boldsymbol{x}) \end{aligned} \qquad [1]$$

where $\mathrm{d}^3x=\mathrm{d}x^1\,\mathrm{d}x^2\,\mathrm{d}x^3$. Then F_1,F_2,F_3 are the three components I_1,I_2,I_3 of the isotopic spin, and $Y=(2\sqrt{3}/3)F_8$ is the hypercharge. The usual electromagnetic current $J_{\mathrm{em}}^\mu(x^0,\boldsymbol{x})$ is given by

$$J_{\mathrm{em}}^\mu = q\left(\mathcal{F}_3^\mu + \frac{\sqrt{3}}{3}\mathcal{F}_8^\mu\right) \qquad [2]$$

where q is the unit elementary charge, and the total charge is given by $Q=\int \mathrm{d}^3x\,J_{\mathrm{em}}^0(x^0,\boldsymbol{x})=q(I_3+Y/2)$. The hadronic part of the weak current entering an effective Lagrangian can be written as

$$\begin{aligned} J_{\mathrm{w}}^\mu = &\left[\left(\mathcal{F}_1^\mu-\mathcal{F}_1^{5\mu}\right)+\mathrm{i}\left(\mathcal{F}_2^\mu-\mathcal{F}_2^{5\mu}\right)\right]\cos\theta_{\mathrm{C}} \\ &+\left[\left(\mathcal{F}_4^\mu-\mathcal{F}_4^{5\mu}\right)+\mathrm{i}\left(\mathcal{F}_5^\mu-\mathcal{F}_5^{5\mu}\right)\right]\sin\theta_{\mathrm{C}} \end{aligned} \qquad [3]$$

where θ_{C} is the Cabibbo angle (determined experimentally to be ~ 0.27 rad). The terms with $\mathcal{F}_1-\mathcal{F}_1^5$ and $\mathcal{F}_2-\mathcal{F}_2^5$ are strangeness conserving, those with $\mathcal{F}_4-\mathcal{F}_4^5$ and $\mathcal{F}_5-\mathcal{F}_5^5$ are not.

The main current algebra hypothesis is that the time components $\mathcal{F}^0$ and $\mathcal{F}^{50}$ of these octets satisfy the equal-time commutation relations:

$$\begin{aligned} &\left[\mathcal{F}_a^0(x^0,\boldsymbol{x}),\mathcal{F}_b^0(y^0,\boldsymbol{y})\right]_{x^0=y^0} \\ &\quad=\mathrm{i}\delta^{(3)}(\boldsymbol{x}-\boldsymbol{y})\sum_d c_{abd}\mathcal{F}_d^0(x^0,\boldsymbol{x}) \\ &\left[\mathcal{F}_a^0(x^0,\boldsymbol{x}),\mathcal{F}_b^{50}(y^0,\boldsymbol{y})\right]_{x^0=y^0} \\ &\quad=\mathrm{i}\delta^{(3)}(\boldsymbol{x}-\boldsymbol{y})\sum_d c_{abd}\mathcal{F}_d^{50}(x^0,\boldsymbol{x}) \\ &\left[\mathcal{F}_a^{50}(x^0,\boldsymbol{x}),\mathcal{F}_b^{50}(y^0,\boldsymbol{y})\right]_{x^0=y^0} \\ &\quad=\mathrm{i}\delta^{(3)}(\boldsymbol{x}-\boldsymbol{y})\sum_d c_{abd}\mathcal{F}_d^0(x^0,\boldsymbol{x}) \end{aligned} \qquad [4]$$

where the c_{abd} are structure constants of the Lie algebra of SU(3), antisymmetric in the indices. Since current commutators relate bilinear expressions to

linear ones, they fix the normalizations of the currents. The chiral currents $\mathcal{F}_a^{L\mu} = (1/2)(\mathcal{F}_a^{\mu} - \mathcal{F}_a^{5\mu})$ and $\mathcal{F}_a^{R\mu} = (1/2)(\mathcal{F}_a^{\mu} + \mathcal{F}_a^{5\mu})$ commute with each other, so that the local current algebra decomposes into two independent pieces.

The Dirac δ-functions in eqns [4] require that $\mathcal{F}_a^0$ and $\mathcal{F}_a^{50}$ be interpreted as (unbounded) operator-valued distributions; while the fixed-time condition suggests these should make mathematical sense as three-dimensional distributions, with x^0 held constant. Such distributions may be modeled on the test-function space $\mathcal{D}$ of real-valued, compactly supported, C^∞ functions on the spacelike hyperplane $\boldsymbol{R}^3$. For functions $f_a, f_a^5 \in \mathcal{D}$, one has formally the "smeared currents" that are expected to be bona fide (unbounded) operators in Hilbert space; suppressing x^0,

$$\begin{aligned} \mathcal{F}_a^0(f_a) &= \int_{R^3} \mathrm{d}^3x f_a(\boldsymbol{x}) \mathcal{F}_a^0(x^0, \boldsymbol{x}) \\ \mathcal{F}_a^{50}(f_a^5) &= \int_{R^3} \mathrm{d}^3x f_a^5(\boldsymbol{x}) \mathcal{F}_a^{50}(x^0, \boldsymbol{x}) \end{aligned} \qquad [5]$$

Equations [4] then become

$$\begin{aligned} [\mathcal{F}_a^0(f_a), \mathcal{F}_b^0(f_b)] &= [\mathcal{F}_a^{50}(f_a), \mathcal{F}_b^{50}(f_b)] \\ &= \mathrm{i} \sum_d \mathcal{F}_d^0(c_{abd} f_a f_b) \end{aligned} \qquad [6]$$

$$[\mathcal{F}_a^0(f_a), \mathcal{F}_b^{50}(f_b)] = \mathrm{i} \sum_d \mathcal{F}_d^{50}(c_{abd} f_a f_b)$$

Let $g(\boldsymbol{x})$ be a C^∞ map from $\boldsymbol{R}^3$ to the Lie algebra $\mathcal{G}$ of chiral SU(3) × SU(3), equal to zero outside a compact set. The set of all such $\mathcal{G}$-valued functions forms an infinite-dimensional Lie algebra under the pointwise bracket, $[g, g'](\boldsymbol{x}) = [g(\boldsymbol{x}), g'(\boldsymbol{x})]$. Let us call this Lie algebra $\mathrm{map}_0(\boldsymbol{R}^3, \mathcal{G})$, where the subscript 0 indicates the condition of compact support when that is applicable (on compact manifolds, we omit the subscript). Expanding $g(\boldsymbol{x})$ with respect to a fixed basis of $\mathcal{G}$, we straightforwardly identify the map g with the two octets of test functions f_a and f_a^5. Then, defining $\mathcal{F}(g) = \sum_a \mathcal{F}_a^0(f_a) + \sum_a \mathcal{F}_a^{50}(f_a^5)$, eqns [6] are interpreted (for fixed x^0) as a representation $\mathcal{F}$ of $\mathrm{map}_0(\boldsymbol{R}^3, \mathcal{G})$.

Integrating out the spatial variables entirely using eqns [1] leads to a representation at x^0 of $\mathcal{G}$ by the charges F_a and F_a^5. The Adler–Weisberger sum rule was first derived (in 1965) from the commutation relations of these charges, together with the assumption of a partially conserved axial-vector current (PCAC). It connected nucleon β-decay coupling with pion–nucleon scattering cross sections, agreeing well with experiment. Various low-energy theorems followed, also in accord with experiment. Shortly thereafter, Adler was able to eliminate the PCAC assumption, and derived a further sum rule going beyond an experimental test of the algebra of charges to test the actual local current algebra. Here, the prediction pertained to structure functions in the deep inelastic scattering of neutrinos. This was elaborated by Bjorken to inelastic electron scattering. On the theoretical side, the study of the chiral current in perturbation theory led into the theory of anomalies. All these ideas were highly influential in subsequent theoretical work (Treiman *et al.* 1985, Mickelsson 1989).

It is a natural idea to try to extend eqns [4] or [6], which elegantly express the combined ideas of locality and symmetry, to an equal-time commutator algebra that would also include the space components of the local currents $\mathcal{F}_a^k, k = 1, 2, 3$. One may write without difficulty the commutators of the charges in [1] with these space components:

$$\begin{aligned} [F_a(x^0), \mathcal{F}_b^k(x^0, \boldsymbol{x})] &= [F_a^5(x^0), \mathcal{F}_b^{5k}(x^0, \boldsymbol{x})] \\ &= \mathrm{i} \sum_d c_{abd} \mathcal{F}_d^k(x^0, \boldsymbol{x}) \\ [F_a(x^0), \mathcal{F}_b^{5k}(x^0, \boldsymbol{x})] &= [F_a^5(x^0), \mathcal{F}_b^k(x^0, \boldsymbol{x})] \\ &= \mathrm{i} \sum_d c_{abd} \mathcal{F}_d^{5k}(x^0, \boldsymbol{x}) \end{aligned} \qquad [7]$$

But the commutator of the local time component with the local space component of the current cannot be merely the obvious extrapolation from eqns [4] and [7], that is, it cannot be

$$\begin{aligned} &[\mathcal{F}_a(x^0, \boldsymbol{x}), \mathcal{F}_b^k(y^0, \boldsymbol{y})]_{x^0 = y^0} \\ &\quad = \mathrm{i}\delta^{(3)}(\boldsymbol{x} - \boldsymbol{y}) \sum_d c_{abd} \mathcal{F}_d^k(x^0, \boldsymbol{x}) \end{aligned}$$

and so forth. Under very general conditions, for a relativistic theory based on local quantum fields or local observables, additional "Schwinger terms" are required on the right-hand sides of such commutators (Renner 1968).

Well-known difficulties in specifying the Schwinger terms are associated with the fact that operator-valued distributions are singular when regarded as if they were functions of spacetime points. Thus, the product of two distributions at a point is often singular or undefined. When the currents forming a local current algebra are written as normal-ordered products of field operator distributions and their derivatives, the Schwinger terms in their commutation relations may be calculated, for example, by "splitting points" in the arguments of the underlying fields, and subsequently letting the separation tend toward zero. The general form of a Schwinger term typically involves the derivative of a δ-function times an operator. This may be a multiple of the identity (i.e., a c-number) or not, depending on the underlying

field-theoretic model. Furthermore, when the number of spacetime dimensions is greater than $1+1$, the c-number Schwinger terms turn out to be infinite. Hence, we do not obtain this way a bona- fide infinite-dimensional, equal-time commutator algebra comprising all the components of the local currents.

Sugawara, Kac–Moody, and Virasoro Algebras

Since equations such as [4] and [6] are not explicitly dependent on how the currents are constructed from underlying canonical fields, one has the possibility of writing a theory entirely in terms of self-adjoint currents as the dynamical variables, bypassing the field operators entirely, and expressing a Hamiltonian operator directly in terms of such local currents. This is in the spirit of approaches to quantum field theory based on local algebras of observables. It suggests consideration of relativistic current algebras with finite c-number or operator Schwinger terms in $s+1$ dimensions, $s \geq 1$.

The Sugawara model, which is of this type, turned out to be one of the most influential of those proposed in the late 1960s and early 1970s. Henceforth, let G be a compact Lie group, and $\mathcal{G}$ its Lie algebra; let $F_a, a = 1, \ldots, \dim \mathcal{G}$, be a basis for $\mathcal{G}$, with $[F_a, F_b] = i\Sigma_d c_{abd} F_d$. The Sugawara current algebra, at the fixed time $x^0 = y^0$ (which, from here on, we suppress in the notation), is given by

$$\begin{aligned}
[J_a^0(x), J_b^0(y)] &= i\delta^{(3)}(x-y)\sum_d c_{abd} J_d^0(x) \\
[J_a^0(x), J_b^k(y)] &= i\delta^{(3)}(x-y)\sum_d c_{abd} J_d^k(x) \\
&\quad + ic\delta_{ab}\frac{\partial}{\partial x^k}\delta^{(3)}(x-y)I \\
[J_a^k(x), J_b^\ell(y)] &= 0 \quad (k,\ell = 1,2,3)
\end{aligned} \tag{8}$$

where $J_a^\mu = (J_a^0, J_a^k)$, $k = 1, 2, 3$, is again a 4-vector, c is a finite constant, and I is the identity operator. The time components in eqns [8] behave like the local currents in eqns [4]. The Schwinger term is a c-number, while setting the commutators of the space components to zero is the simplest choice consistent with the Jacobi identity. The Sugawara Hamiltonian is given in terms of the local currents by the formal expression:

$$H = \frac{1}{2c}\sum_a \int_{\mathbf{R}^3} d^3x \left[J_a^0(x)^2 + \sum_{k=1}^{3} J_a^k(x)^2 \right] \tag{9}$$

where the pointwise products of the currents require interpretation in the particular representation. This Hamiltonian leads to current conservation equations for the J_a^μ.

Related to the Sugawara current algebra, with $s = 1$ and the spatial dimension compactified, are affine Kac–Moody and Virasoro algebras (Goddard and Olive 1986, Kac 1990). Consider the infinite-dimensional Lie algebra map$(S^1, \mathcal{G})$ of smooth functions from the circle to $\mathcal{G}$ under the pointwise bracket. This is also called a loop algebra. Referring to the basis F_a, define $T_a^{(m)}$ for integer m to be the Fourier function $\theta \to F_a \exp[-im\theta]$. The pointwise bracket in map$(S^1, \mathcal{G})$ gives $[T_a^{(m)}, T_b^{(n)}] = i\Sigma_d c_{abd} T_d^{(m+n)}$ for these generators. The corresponding (untwisted) affine Kac–Moody algebra is a (uniquely defined, nontrivial) one-dimensional central extension of this loop algebra – that is, the new generator commutes with all elements of the Lie algebra and, in an irreducible representation, must be a multiple of the identity. In such a representation, the new bracket can be written as

$$[T_a^{(m)}, T_b^{(n)}] = i\sum_d c_{abd} T_d^{(m+n)} + km\delta_{ab}\delta_{m,-n} I \tag{10}$$

where k is a constant. Here, $T_a^{(m=0)}$ is again a representation of $\mathcal{G}$. Self-adjointness of the local currents in the representation imposes the condition $T_a^{(m)^*} = T_a^{(-m)}$.

Now the compactly supported C^∞ (tangent) vector fields on a C^∞ manifold M form a natural Lie algebra under the Lie bracket, denoted by vect$_0(M)$. In local Euclidean coordinates, for $g_1, g_2 \in$ vect$_0(M)$, one can write this bracket as

$$[g_1, g_2] = g_1 \cdot \nabla g_2 - g_2 \cdot \nabla g_1 \tag{11}$$

As the affine Kac–Moody algebras are central extensions of the algebra of $\mathcal{G}$-valued functions on S^1, so are Virasoro algebras central extensions of the algebra of vector fields on S^1. Let $L^{(m)}$ denote the (complexified) vector field described by $\exp[-im\theta](1/i)\partial/\partial\theta$, for integer m. These generators then satisfy $[L^{(m)}, L^{(n)}] = (m-n)L^{(m+n)}$. Adjoining to the Lie algebra of vector fields a new central element (commuting with all the $L^{(m)}$), the Virasoro bracket in an irreducible representation is given by the formula

$$\begin{aligned}
[L^{(m)}, L^{(n)}] &= (m-n)L^{(m+n)} \\
&\quad + c\frac{(m+1)m(m-1)}{12}\delta_{m,-n} I
\end{aligned} \tag{12}$$

where the numerical coefficient c is called the Virasoro central charge; self-adjointness of the currents imposes $L^{(m)^*} = L^{(-m)}$. It is straightforward to verify that eqn [12] satisfies the Jacobi identity. The special form of the central term in the Virasoro current algebra results from the Gelfand–Fuks cohomology on the algebra of vector fields.

The Kac–Moody and Virasoro algebras, both modeled on S^1, may be combined to form a natural semidirect sum of Lie algebras, with the additional bracket

$$[T_a^{(m)}, L^{(n)}] = mT_a^{(m+n)} \qquad [13]$$

Roughly speaking, the Kac–Moody generators correspond to Fourier transforms of charge densities on S^1, whereas the Virasoro generators correspond to Fourier transforms of infinitesimal motions in S^1. The central extensions provide the finite, c-number Schwinger terms. These structures have important application to light cone current algebra, conformally invariant quantum field theories in $(1+1)$-dimensional spacetime, the quantum theory of strings, exactly solvable models in statistical mechanics, and many other domains.

Of greatest physical importance, both in quantum field theory and statistical mechanics, are those irreducible, self-adjoint representations of the Virasoro algebra known as highest weight representations, where the spectrum of the operator $L^{(m=0)}$ is bounded below. In these applications, one represents a pair of Virasoro algebras by mutually commuting sets of operators $L^{(m)}$ and $\bar{L}^{(m)}$. In the quantum theory, for example, one takes the total energy $H \propto \bar{L}^{(0)} + L^{(0)}$, and the total momentum $P \propto \bar{L}^{(0)} - L^{(0)}$. In a highest weight representation, there is a unique eigenstate of $L^{(0)}$ having the lowest eigenvalue h; for this "vacuum" $|h\rangle$, $L^{(m)}|h\rangle = 0$, $m > 0$.

Friedan, Qiu, and Shenker showed in 1984 that highest weight representations are characterized by a class of specific, non-negative values of the central charge c and, correspondingly, of h: either $c \geq 1$ (and $h \geq 0$) or $c = 1 - 6(\ell+2)^{-1}(\ell+3)^{-1}$, $\ell = 1, 2, 3, \ldots$ (and h assumes a corresponding, specified set of values for each value of ℓ). In a beautiful application to the study of the critical behavior of well-known statistical systems, in which the generator of dilations is proportional to $\bar{L}^{(0)} + L^{(0)}$, they discovered a direct correspondence with permitted values of the central charge; thus, $c = 1/2$ for the Ising model, $c = 7/10$ for the tricritical Ising model, $c = 4/5$ for the three-state Potts model, and $c = 6/7$ for the tricritical three-state Potts model.

Current Algebras and Groups

Local current algebras may be exponentiated to obtain corresponding infinite-dimensional topological groups (Pressley and Segal 1986, Mickelsson 1989, Kac 1990). Let G be a Lie group whose Lie algebra is $\mathcal{G}$. The algebra $\mathrm{map}_0(M, \mathcal{G})$, consisting of smooth, compactly supported $\mathcal{G}$-valued functions on M under the pointwise bracket, exponentiates to the local current group $\mathrm{Map}_0(M, G)$, consisting of smooth maps from M to G that are the identity outside a compact set in M, under the pointwise group operation. When M is taken to be the four-dimensional spacetime manifold (rather than a spacelike hyperplane), the local current group modeled on M is mathematically a gauge group for nonabelian gauge field theory.

Likewise, the algebra $\mathrm{vect}_0(M)$ exponentiates to the group $\mathrm{Diff}_0(M)$ of compactly supported C^∞ diffeomorphisms of M (under composition). The Kac–Moody and Virasoro algebras exponentiate to central extensions of the loop group $\mathrm{Map}(S^1, G)$ and the diffeomorphism group $\mathrm{Diff}(S^1)$, respectively. The semidirect sums of the Lie algebras are the infinitesimal generators of semidirect products of the groups.

Under appropriate technical conditions, self-adjoint representations of current algebras generate (and may be obtained from) continuous unitary representations of the corresponding groups. The needed technical conditions have to do with the existence of a dense set of analytic vectors belonging to a common, dense invariant domain of essential self-adjointness for the currents.

Nonrelativistic Current Algebra

In nonrelativistic local current algebra, Schwinger terms do not appear. In 1968, Dashen and Sharp defined (at fixed time t, suppressed in the present notation) a mass density $\rho(\boldsymbol{x}) = m\psi^*(\boldsymbol{x})\psi(\boldsymbol{x})$ and a momentum density $\boldsymbol{J}(\boldsymbol{x}) = (\hbar/2\mathrm{i})\{\psi^*(\boldsymbol{x})\nabla\psi(\boldsymbol{x}) - [\nabla\psi^*(\boldsymbol{x})]\psi(\boldsymbol{x})\}$, where ψ is a second-quantized canonical field; here we keep $\hbar$ in the notation. The resulting equal-time algebra is the semidirect sum:

$$\begin{aligned}
[\rho(\boldsymbol{x}), \rho(\boldsymbol{y})] &= 0 \\
[\rho(\boldsymbol{x}), J^k(\boldsymbol{y})] &= -\mathrm{i}\hbar \frac{\partial}{\partial x^k}[\delta^{(3)}(\boldsymbol{x}-\boldsymbol{y})\rho(\boldsymbol{x})] \\
[J^k(\boldsymbol{x}), J^\ell(\boldsymbol{y})] &= \mathrm{i}\hbar\left\{\frac{\partial}{\partial y^k}[\delta^{(3)}(\boldsymbol{x}-\boldsymbol{y})J^\ell(\boldsymbol{y})]\right. \\
&\qquad \left. - \frac{\partial}{\partial x^\ell}[\delta^{(3)}(\boldsymbol{x}-\boldsymbol{y})J^k(\boldsymbol{x})]\right\}
\end{aligned} \qquad [14]$$

Since this current algebra is independent of whether ψ obeys commutation or anticommutation relations, the information as to particle statistics (Bose or Fermi) is not encoded in the Lie algebra itself but in the choice of its representation (up to unitary equivalence). Again interpreting ρ and J^k as operator-valued distributions, define $\rho(f) = \int_{R^3} \mathrm{d}^3x\, f(\boldsymbol{x})\rho(\boldsymbol{x})$ and $J(\boldsymbol{g}) = \int_{R^3} \mathrm{d}^3x \sum_{k=1}^3 g^k(\boldsymbol{x})J^k(\boldsymbol{x})$, where f and the

components g^k of the vector field g belong to the function-space $\mathcal{D}$. Then the Lie algebra becomes

$$\begin{aligned}[\rho(f_1),\rho(f_2)] &= 0\\ [\rho(f),J(g)] &= i\hbar\rho(g\cdot\nabla f)\\ [J(g_1),J(g_2)] &= -i\hbar J([g_1,g_2])\end{aligned} \qquad [15]$$

Equations [15] are a representation by self-adjoint operators of the semidirect sum of the abelian Lie algebra $\mathcal{D}$ with $\mathrm{vect}_0(\boldsymbol{R}^3)$. The corresponding group is the natural semidirect product of the space $\mathcal{D}$ (regarded as an abelian topological group under addition) with $\mathrm{Diff}_0(\boldsymbol{R}^3)$.

The construction generalizes to a general manifold M or manifold with boundary (in place of $\boldsymbol{R}^3$), and to a general set of charge densities that generate the local Lie algebra $\mathrm{map}_0(M,\mathcal{G})$. When $M=S^1$, we have the Kac–Moody and Virasoro algebras with central charge zero. However, $L^{(0)}$ in the nonrelativistic $(1+1)$-dimensional quantum theories is proportional to the total momentum P, and thus is unbounded above and below.

The continuous unitary representations of $\mathrm{Diff}_0(M)$, or its semidirect product with a local current group at fixed time, thus describe nonrelativistic quantum systems (Albeverio *et al.* 1999, Goldin 2004). The unitary representation $V(\phi), \phi \in \mathrm{Diff}_0(M)$, satisfies $V(\phi_r^g) = \exp[i(r/\hbar)J(g)]$, where $r\in\boldsymbol{R}$ and ϕ_r^g is the one-parameter flow in $\mathrm{Diff}_0(M)$ generated by the vector field g. Such a representation may be described very generally by means of a measure μ on a configuration space Δ, quasi-invariant under a group action of $\mathrm{Diff}_0(M)$ on Δ, together with a unitary 1-cocycle χ on $\mathrm{Diff}_0(M)\times\Delta$. The Hilbert space for the representation is $\mathcal{H}=L^2_{d\mu}(\Delta,\mathcal{W})$, which is the space of measurable functions $\Psi(\gamma), \gamma\in\Delta$, taking values in an inner product space $\mathcal{W}$, and square integrable with respect to μ. The unitary representation V is given by

$$[V(\phi)\Psi](\gamma) = \chi_\phi(\gamma)\Psi(\phi\gamma)\sqrt{\frac{d\mu_\phi}{d\mu}(\gamma)} \qquad [16]$$

where $\phi\gamma$ denotes the group action $\mathrm{Diff}_0(M)\times\Delta\to\Delta$; μ_ϕ is the measure on Δ transformed by ϕ (which, by the quasi-invariance of μ, is absolutely continuous with respect to μ); $d\mu_\phi/d\mu$ is the Radon–Nikodym derivative; and $\chi_\phi(\gamma):\mathcal{W}\to\mathcal{W}$ is a system of unitary operators in $\mathcal{W}$ obeying the cocycle equation

$$\chi_{\phi_1\phi_2}(\gamma) = \chi_{\phi_1}(\gamma)\chi_{\phi_2}(\phi_1\gamma) \qquad [17]$$

Equations [16] and [17] hold outside sets of μ-measure zero in Δ. Given the quasi-invariant measure μ on Δ, one may always choose $\mathcal{W}=\mathbf{C}$ and $\chi_\phi(\gamma)\equiv 1$ to obtain a unitary group representation on complex-valued wave functions; but inequivalent cocycles describe unitarily inequivalent representations.

The configuration space $\Delta^{(N)}, N=1,2,3,\ldots$, consists of N-point subsets of $\boldsymbol{R}^s$, and $\mu^{(N)}$ is the (local) Lebesgue measure on $\Delta^{(N)}$. The corresponding diffeomorphism group and local current algebra representations describe N identical quantum particles in s-dimensional space. When $\chi\equiv 1$, we have bosonic exchange symmetry. Inequivalent cocycles on $\Delta^{(N)}$ are obtained (for $s\geq 2$) by inducing (generalizing Mackey's method) from inequivalent unitary representations of the fundamental group $\pi_1[\Delta^{(N)}]$. For $s\geq 3$, this fundamental group is the symmetric group S_N of particle permutations; the odd representation of $S_N, N\geq 2$, gives fermionic exchange symmetry, while the higher-dimensional representations are associated with particles satisfying the parastatistics of Greenberg and Messiah.

When $s=2$, however, $\pi_1[\Delta^{(N)}]$ is the braid group B_N. Goldin, Menikoff, and Sharp obtained induced representations of the current algebra describing the intermediate statistics proposed by Leinaas and Myrheim for identical particles in 2-space. Such excitations, subsequently termed "anyons" by Wilczek and characterized as charge-flux tube composites, are important constructs in the theory of surface phenomena such as the quantum Hall effect, and anyonic statistics has also been applied to the study of high-T_c superconductivity. Current algebra representations induced by higher-dimensional representations of B_N describe the statistics of "plektons." Similarly, current algebra in nonsimply connected space describes the Aharonov–Bohm effect.

Let $\psi^*(h)=\int_{\boldsymbol{R}^s} d^s x\, h(x)\psi^*(x)$ denote the smeared creation field. Let the indexed set of representations $\rho_N, J_N, N=0,1,2,\ldots$, satisfying the current algebra [15], act in Hilbert spaces $\mathcal{H}_N$, where $\psi^*(h):\mathcal{H}_N\to\mathcal{H}_{N+1}, \psi(h):\mathcal{H}_{N+1}\to\mathcal{H}_N, \psi(h)|\mathcal{H}_0=0$, so that ψ^* and ψ intertwine the N-particle diffeomorphism group representations. Let $\rho(f)$ and $J(g)$ act on $\bigoplus_{N=0}^\infty\mathcal{H}_N$, so that $\rho(f)\Psi_N=\rho_N(f)\Psi_N, J(g)\Psi_N = J_N(g)\Psi_N$. Then conditions for a Fock space hierarchy are specified by commutator brackets of the fields with the currents:

$$\begin{aligned}[\rho(f),\psi^*(h)] &= \psi^*(\rho_{N-1}(f)h)\\ [J(g),\psi^*(h)] &= \psi^*(J_{N-1}(g)h)\end{aligned} \qquad [18]$$

The local creation and annihilation fields for anyons in $\boldsymbol{R}^2$, obeying [18], satisfy q-commutation relations, where q is the relative phase change associated with a single counterclockwise exchange of two anyons,

and the q-commutator $[A, B]_q = AB - qBA$. These relations generalize the canonical commutation ($q = 1$) and anticommutation ($q = -1$) relations of quantum field theory.

When Δ is the configuration space of infinite but locally finite subsets of $\boldsymbol{R}^s$, nonrelativistic current algebra describes the physics of infinite gases in continuum classical or quantum statistical mechanics. Here, the most important kinds of measures μ are Poisson measures (associated with gases of noninteracting particles at fixed average density) or Gibbsian measures (associated with translation-invariant two-body interactions). These measures describe equilibrium states and correlation functions in the classical case, and specify the current algebra representations in the quantum theory.

The group of volume-preserving diffeomorphisms was taken by Arnold as the symmetry group of an ideal, classical, incompressible fluid, and Marsden and Weinstein described the hydrodynamics of such a fluid using the Lie–Poisson bracket associated with the nonrelativistic current algebra of divergenceless vector fields. The idea of using this algebra to study quantized fluid motion, included in the program proposed by Rasetti and Regge, formed the basis of the subsequent study of quantized vortex structures in superfluids from the point of view of geometric quantization on coadjoint orbits of the diffeomorphism group. This leads to quantum configuration spaces whose elements are no longer sets of points – for example, spaces of vortex filaments in $\boldsymbol{R}^2$, or ribbons and tubes in $\boldsymbol{R}^3$.

See also: Algebraic Approach to Quantum Field Theory; Electroweak Theory; Quantum Chromodynamics; Solitons and Kac–Moody Lie Algebras; Symmetries in Quantum Field Theory: Algebraic Aspects; Toda Lattices; Two-Dimensional Conformal Field Theory and Vertex Operator Algebras.

Further Reading

Adler SL and Dashen RF (1968) *Current Algebras and Applications to Particle Physics*. New York: Benjamin.

Albeverio S, Kondratiev YuG, and Röckner M (1999) Diffeomorphism groups and current algebras: configuration space analysis in quantum theory. *Reviews in Mathematical Physics* 11: 1–23.

Arnol'd VI and Khesin BA (1998) *Topological Methods in Hydrodynamics*. Applied Mathematical Sciences, vol. 125. Berlin: Springer.

Gell-Mann M and Ne'eman Y (2000) *The Eightfold Way (1964) (reissued)*. Cambridge, MA: Perseus Publishing.

Goddard P and Olive D (1986) Kac–Moody and Virasoro algebras in relation to quantum physics. *International Journal of Modern Physics A* 1: 303–414.

Goldin GA (1996) Quantum vortex configurations. *Acta Physica Polonica B* 27: 2341–2355.

Goldin GA (2004) Lectures on diffeomorphism groups in quantum physics. In: Govaerts J, Hounkonnou N, and Msezane AZ (eds.) *Contemporary Problems in Mathematical Physics: Proceedings of the Third International Workshop*, pp. 3–93. Singapore: World Scientific.

Goldin GA and Sharp DH (1991) The diffeomorphism group approach to anyons. *International Journal of Modern Physics B* 5: 2625–2640.

Ismagilov RS (1996) *Representations of Infinite-Dimensional Groups*. Translations of Mathematical Monographs, vol. 152. Providence, RI: American Mathematical Society.

Kac V (1990) *Infinite Dimensional Lie Algebras*. Cambridge: Cambridge University Press.

Marsden J and Weinstein A (1983) Coadjoint orbits, vortices, and Clebsch variables for incompressible fluids. *Physica D* 7: 305–323.

Mickelsson J (1989) *Current Algebras and Groups*. New York: Plenum.

Ottesen JT (1995) *Infinite Dimensional Groups and Algebras in Quantum Physics*. Berlin: Springer.

Pressley A and Segal G (1986) *Loop Groups*. Oxford: Oxford University Press.

Renner B (1968) *Current Algebras and Their Applications*. Oxford: Pergamon.

Sharp DH and Wightman AS (eds.) (1974) *Local Currents and Their Applications*. New York: Elsevier.

Treiman SB, Jackiw R, Zumino B, and Witten E (1985) *Current Algebra and Anomalies*. Singapore: World Scientific.

Dirac Operator and Dirac Field

S N M Ruijsenaars, Centre for Mathematics and Computer Science, Amsterdam, The Netherlands

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The Dirac equation arose in the early days of quantum mechanics, inspired by the problem of taking special relativity into account in the quantum mechanical description of a freely moving electron. From the outset, however, Dirac looked for an equation that also accomodated the electron spin and that could be modified to include interaction with an external electromagnetic field. The equation he discovered satisfies all of these requirements. On the other hand, when it is rewritten in Hamiltonian form, the spectrum of the resulting Dirac operator includes not only the desired interval $[mc^2, \infty)$ (where m is the electron mass and c the speed of light), but also an interval $(-\infty, -mc^2]$.

Dirac himself already considered this negative part of the spectrum as unphysical, since no such negative energies had been observed and their presence would entail instability of the electron. This physical flaw of the "first-quantized" description of a relativistic electron led to the introduction of "second quantization," as encoded in quantum field theory. In the field-theoretic version of the Dirac theory, the unphysical negative energies are obviated by a prescription that originated in Dirac's hole theory.

Specifically, Dirac postulated that the negative-energy states of his equation were occupied by a sea of unobservable particles, the Pauli principle forbidding an occupancy greater than one. In this heuristic picture, the annihilation of a negative-energy electron yields a hole in the sea, observable as a new type of positive-energy particle with the same mass, but opposite charge. This led Dirac to predict that the electron should have an oppositely charged partner.

His prediction was soon confirmed experimentally, the partner of the negatively charged electron showing up as the positively charged positron. More generally, all electrically charged particles (not only spin-1/2 particles described by the Dirac equation) have turned out to have oppositely charged antiparticles. Furthermore, some electrically neutral particles also have distinct antiparticles.

Returning to the second-quantized Dirac theory, this involves a Dirac quantum field in which the creation/annihilation operators of negative-energy states are replaced by annihilation/creation operators of positive-energy holes, resp. The hole theory substitution therefore leads to a Hilbert space (called Fock space) that accomodates an arbitrary number of particles and antiparticles with the same mass and opposite charge.

Soon after the introduction of the Dirac equation (which dates from 1928), it turned out that the number of particles and antiparticles is not conserved in a high-energy collision. Such creation and annihilation processes admit a natural description in the Fock spaces associated with relativistic quantum field theories. The very comprehensive mathematical description of real-world elementary particle phenomena that is now called the standard model arose some 30 years ago, and has been abundantly confirmed by experiment ever since. It involves

Notes:

various relativistic quantum fields with nonlinear interactions. The Dirac quantum field is an essential ingredient, inasmuch as it is used to describe all spin-1/2 particles and antiparticles in the model (including quarks, electrons, neutrinos etc.).

After this survey (which is not only very brief, but also biased toward the physical concepts at issue), the contents of this article will be sketched. The free Dirac equation associated with the physical Minkowski spacetime $\mathbb{R}^4$ is first detailed. The exposition and notation are slightly unconventional in some respects. This is because we are partly preparing the ground for a mathematically precise account of the second-quantized version of the free Dirac theory. For example, momentum space (as opposed to position space) is emphasized, since the variable x in the Dirac equation does not have a clear physical significance and should be discarded in the Hilbert space formulation of the second-quantized Dirac field. The latter acts on a Fock space of multi-particle and -antiparticle wave functions depending on momentum and spin variables, and the spacetime dependence of the Dirac field is solely a consequence of relativistic covariance. (In particular, the variable x in the Dirac field $\Psi(t, x)$ should not be viewed as the position of particles and antiparticles created and annihilated by the field.)

To be sure, there is much more to the Dirac theory than its free first- and second-quantized versions for Minkowski spacetime $\mathbb{R}^4$. The primary purpose here is, however, to present these foundational versions in some detail. A much more sketchy account of further developments can be found in subsequent sections. First, the one-particle theory is reconsidered. Generalizations of the free theory to arbitrary dimensions and Euclidean settings are sketched and interactions with external fields are described, touching on various aspects and applications.

The next focus is on relations with index theory that arise when the massless Euclidean Dirac operator is generalized to geometric settings, namely l-dimensional Riemannian manifolds allowing a spin structure. We illustrate the general Atiyah–Singer index theory for the Dirac framework with some simple examples for $l=1$ (Toeplitz operators) and $l=2$ (the manifold $S^1 \times S^1$).

More information on the many-particle Dirac theory appears in the final section. Brief remarks on the Dirac field in interaction with other quantized fields are followed by an elaboration of the far simpler situation of the Dirac field interacting with external fields. Among the S-operators corresponding to such fields there is a special class of unitary matrix multipliers; the external field then vanishes for $t < 0$ and equals the pure gauge field corresponding to the unitary matrix for $t \geq 0$. Specializing to an even spacetime dimension and choosing special "kink" type unitaries, the associated Fock-space quadratic forms can be made to converge to the free Dirac field.

As mentioned already, Dirac's second quantization procedure was invented to get rid of the unphysical negative energies of the first-quantized (one-particle) theory. It is an amazing fact that the resulting formalism for the simplest case (namely the massless Dirac operator in a two-dimensional spacetime) can be exploited for quite different purposes. In particular, this setting can be tied in with various soliton equations and the representation theory of certain infinite-dimensional groups and Lie algebras. In conclusion, some of these applications are briefly sketched, namely the construction of special solutions to the Kadomtsev–Petviashvili (KP) equation (including the KP solitons and finite-gap solutions) and special representations of Kac–Moody and Virasoro algebras.

The Free One-Particle Dirac Equation in $\mathbb{R}^4$

The free time-dependent Dirac equation is a linear hyperbolic evolution equation for a function $\Psi(t, x)$ on spacetime $\mathbb{R}^4$ with values in $\mathbb{C}^4$. It involves four 4×4 matrices γ^μ, $\mu = 0, 1, 2, 3$, satisfying the γ-algebra

$$\gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu = 2g^{\mu\nu} 1_4, \quad g = \mathrm{diag}(1, -1, -1, -1) \quad [1]$$

Using the Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}$$
$$\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad [2]$$

one can choose for example

$$\gamma^0 = \begin{pmatrix} 0 & 1_2 \\ 1_2 & 0 \end{pmatrix}, \quad \gamma^k = \begin{pmatrix} 0 & -\sigma_k \\ \sigma_k & 0 \end{pmatrix}$$
$$k = 1, 2, 3 \quad [3]$$

Now the free Dirac equation reads

$$(\mathrm{i}\hbar\gamma^0 \partial_t + \mathrm{i}\hbar c \gamma \cdot \nabla - mc^2 1_4)\Psi(t, x) = 0 \quad [4]$$

where $\hbar$ is Planck's constant, c the speed of light, and m the particle mass. Using from now on units so that $\hbar = c = 1$, this can be abbreviated as

$$\left(\mathrm{i}\sum_{\mu=0}^{3}\gamma^{\mu}\partial_{\mu} - m\right)\Psi(x) = 0, \quad x = (x^0, x^1, x^2, x^3)$$
$$\partial_{\mu} = \partial/\partial x^{\mu}, \quad \mu = 0, 1, 2, 3 \qquad [5]$$

The relativistic invariance of this equation can be understood as follows. First, since the equation does not explicitly involve the spacetime coordinates, it is invariant under spacetime translations. (If $\Psi(t, \boldsymbol{x})$ solves [5], then also $\Psi(t - a_0, \boldsymbol{x} - \boldsymbol{a})$ is a solution for all $(a_0, \boldsymbol{a}) \in \mathbb{R}^4$.) Second, it is invariant under Lorentz transformations (rotations and boosts). Indeed, if $\Psi(x)$ is a solution and $L \in \mathrm{SO}(1,3)$, then $S(L)\Psi(L^{-1}x)$ solves [5] too, where $S(L)$ denotes a (suitably normalized) matrix satisfying

$$S(L)^{-1}\gamma^{\mu}S(L) = \sum_{\nu=0}^{3} L^{\mu}{}_{\nu}\gamma^{\nu} \qquad [6]$$

(The matrices $\gamma^{\mu\prime}$ on the right-hand side of [6] satisfy the γ-algebra [1]. From this, the existence of a representation $S(L)$ of $\mathrm{SO}(1,3)$ satisfying [6] is readily deduced.)

As a consequence, the Poincaré group (inhomogeneous Lorentz group) acts in a natural way on the space of solutions to the time-dependent Dirac equation, expressing its independence of the choice of inertial frame. For quantum mechanical purposes, however, one needs to choose a frame and use the associated time variable to rewrite the equation as a Hilbert space evolution equation.

The relevant Hilbert space $\check{\mathcal{H}}$ is the space of four-component functions that are square integrable over space,

$$\check{\mathcal{H}} = L^2(\mathbb{R}^3, \mathrm{d}\boldsymbol{x}) \otimes \mathbb{C}^4 \qquad [7]$$

To obtain a self-adjoint Hamiltonian on $\check{\mathcal{H}}$, one multiplies [5] by γ^0 and introduces the Hermitian matrices

$$\beta = \gamma^0, \qquad \alpha^k = \gamma^0\gamma^k, \quad k = 1, 2, 3 \qquad [8]$$

Then, one obtains the Schrödinger type equation

$$\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t}\psi = \check{H}\psi \qquad [9]$$

where $\check{H}$ is the Dirac operator,

$$\check{H} = -\mathrm{i}\boldsymbol{\alpha}\cdot\nabla + \beta m \qquad [10]$$

Under Fourier transformation,

$$\mathcal{F}: \check{\mathcal{H}} \to L^2(\mathbb{R}^3, \mathrm{d}\boldsymbol{p}) \otimes \mathbb{C}^4$$
$$\psi(\boldsymbol{x}) \mapsto \phi(\boldsymbol{p}) = (2\pi)^{-3/2}\int_{\mathbb{R}^3} \mathrm{d}\boldsymbol{x}\ \exp(-\mathrm{i}\boldsymbol{x}\cdot\boldsymbol{p})\psi(\boldsymbol{x}) \qquad [11]$$

eqn [9] turns into

$$\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t}\phi = D(\boldsymbol{p})\phi, \quad D(\boldsymbol{p}) = \boldsymbol{\alpha}\cdot\boldsymbol{p} + \beta m \qquad [12]$$

The matrix $D(\boldsymbol{p})$ is Hermitian and has square $E_p^2\mathbf{1}_4$, where E_p is the relativistic energy,

$$E_p = (\boldsymbol{p}\cdot\boldsymbol{p} + m^2)^{1/2} \qquad [13]$$

corresponding to a momentum $\boldsymbol{p}$. Now, we have

$$U_C\overline{D(-\boldsymbol{p})} = -D(\boldsymbol{p})U_C \qquad [14]$$

where U_C is the charge conjugation matrix,

$$U_C = \mathrm{i}\gamma^2 \qquad [15]$$

Hence, the four eigenvalues of $D(\boldsymbol{p})$ are given by E_p, E_p, $-E_p$, and $-E_p$. Therefore, the matrices

$$P_{\pm}(\boldsymbol{p}) = \frac{1}{2}\left(\mathbf{1}_4 \pm \frac{D(\boldsymbol{p})}{E_p}\right) \qquad [16]$$

are projections on the positive and negative spectral subspaces of $D(\boldsymbol{p})$.

As orthonormal base for the positive-energy subspace, we can now choose

$$w_{+,j}(\boldsymbol{p}) = \left(\frac{2E_p}{E_p + m}\right)^{1/2} P_+(\boldsymbol{p})b_j, \quad j = 1, 2 \qquad [17]$$

where

$$b_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\1\\0\end{pmatrix}, \quad b_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\1\\0\\1\end{pmatrix} \qquad [18]$$

Next, setting

$$w_{-,j}(\boldsymbol{p}) = U_C\overline{w_{+,j}}(\boldsymbol{p}), \quad j = 1, 2 \qquad [19]$$

an orthonormal base $w_{-,1}(-\boldsymbol{p})$, $w_{-,2}(-\boldsymbol{p})$ for the negative-energy subspace of $D(\boldsymbol{p})$ is obtained; cf. [14].

The upshot is that the time-independent Dirac equation

$$\check{H}\psi = E\psi \qquad [20]$$

gives rise to bounded eigenfunctions

$$\begin{aligned} &e_{+,j}(\boldsymbol{x},\boldsymbol{p}) \\ &\quad = (2\pi)^{-3/2}\exp(\mathrm{i}\boldsymbol{x}\cdot\boldsymbol{p})w_{+,j}(\boldsymbol{p}), \quad j=1,2 \\ &e_{-,j}(\boldsymbol{x},\boldsymbol{p}) \\ &\quad = (2\pi)^{-3/2}\exp(-\mathrm{i}\boldsymbol{x}\cdot\boldsymbol{p})w_{-,j}(\boldsymbol{p}), \quad j=1,2 \end{aligned} \tag{21}$$

with eigenvalues $E=E_p$ and $E=-E_p$, resp. Clearly, they are not square-integrable, but they can be used as the kernel of a unitary transformation between $\check{\mathcal{H}}$ (7) and the Hilbert space

$$\begin{aligned} \mathcal{H} &= \mathcal{H}_+ \oplus \mathcal{H}_- = P_+\mathcal{H} \oplus P_-\mathcal{H} \\ \mathcal{H}_+, \mathcal{H}_- &= L^2(\mathbb{R}^3, \mathrm{d}\boldsymbol{p}) \otimes \mathbb{C}^2 \end{aligned} \tag{22}$$

Specifically, we have

$$\begin{aligned} &W : \mathcal{H} \to \check{\mathcal{H}} \\ &f(\boldsymbol{p}) = (f_+(\boldsymbol{p}), f_-(\boldsymbol{p})) \\ &\quad \mapsto \psi(\boldsymbol{x}) = \sum_{\delta=+,-}\sum_{j=1,2}\int_{\mathbb{R}^3} \mathrm{d}\boldsymbol{p}\, e_{\delta,j}(\boldsymbol{x},\boldsymbol{p}) f_{\delta,j}(\boldsymbol{p}) \end{aligned} \tag{23}$$

which entails

$$(W^{-1}\psi)_{\delta,j}(\boldsymbol{p}) = \int_{\mathbb{R}^3} \mathrm{d}\boldsymbol{x}\, \overline{e_{\delta,j}}(\boldsymbol{x},\boldsymbol{p}) \cdot \psi(\boldsymbol{x}) \tag{24}$$

(Here and throughout this article, a bar denotes complex conjugation.)

From the above, it is clear that the Dirac Hamiltonian $\check{H}$ acting on the Hilbert space $\check{\mathcal{H}}$ is unitarily equivalent to the multiplication operator on $\mathcal{H}$ [22] given by

$$(Hf)_\delta(\boldsymbol{p}) = \delta E_{\boldsymbol{p}} f_\delta(\boldsymbol{p}), \quad \delta = +,- \tag{25}$$

Indeed, W is a diagonalizing transformation for $\check{H}$, the relation

$$H = W^{-1}\check{H}W \tag{26}$$

yielding an explicit realization of the spectral theorem.

Using the same notational convention, the momentum, charge conjugation, parity, and time-reversal operators on $\check{\mathcal{H}}$, given by

$$(\check{P}_k\psi)(\boldsymbol{x}) = -\mathrm{i}\partial_k\psi(\boldsymbol{x}) \tag{27}$$

$$(\check{C}\psi)(\boldsymbol{x}) = U_C\bar{\psi}(\boldsymbol{x}) \tag{28}$$

$$(\check{P}\psi)(\boldsymbol{x}) = U_P\psi(-\boldsymbol{x}), \quad U_P = \gamma^0 \tag{29}$$

$$(\check{T}\psi)(\boldsymbol{x}) = U_T\bar{\psi}(\boldsymbol{x}), \quad U_T = \gamma^1\gamma^3 \tag{30}$$

transform into the operators

$$(P_kf)_\delta(\boldsymbol{p}) = \delta p_k f_\delta(\boldsymbol{p}), \quad \delta = +,- \tag{31}$$

$$(Cf)_\delta(\boldsymbol{p}) = \bar{f}_{-\delta}(\boldsymbol{p}), \quad \delta = +,- \tag{32}$$

$$(Pf)_\delta(\boldsymbol{p}) = \delta f_\delta(-\boldsymbol{p}), \quad \delta = +,- \tag{33}$$

$$(Tf)_\delta(\boldsymbol{p}) = \mathrm{i}\sigma_2\bar{f}_\delta(-\boldsymbol{p}), \quad \delta = +,- \tag{34}$$

Note that P_k, P, and T leave the positive- and negative-energy subspaces $\mathcal{H}_+$ and $\mathcal{H}_-$ invariant, whereas C interchanges them.

To conclude this section, we describe some salient features of the unitary representation of the (identity component of the) Poincaré group on $\mathcal{H}$, which follows from the representation on solutions to [5] already sketched. The spacetime translations over $a \in \mathbb{R}^4$ are represented by the unitary operator $\exp(-\mathrm{i}a_0H + \mathrm{i}\boldsymbol{a}\cdot\boldsymbol{P})$; explicitly,

$$\begin{aligned} &(\exp(-\mathrm{i}a_0H + \mathrm{i}\boldsymbol{a}\cdot\boldsymbol{P})f)_\delta(\boldsymbol{p}) \\ &\quad = \exp(-\mathrm{i}\delta(a_0E_{\boldsymbol{p}} - \boldsymbol{a}\cdot\boldsymbol{p}))f_\delta(\boldsymbol{p}), \quad \delta = +,- \end{aligned} \tag{35}$$

The representation of the Lorentz group involves unitary 2×2 matrices $U(k, A)$, where k is an arbitrary 4-vector satisfying $k^\mu k_\mu = 1$ and A the matrix in SL(2, C) representing $L \in \mathrm{SO}(1,3)$. (Recall that SL(2, C) can be viewed as a 2-fold cover of SO(1, 3).) In particular, $U(k, A)$ does not depend on k for rotations,

$$U(k, A) = A^*, \quad \forall A \in \mathrm{SU}(2) \tag{36}$$

(Here and henceforth, we use $*$ to denote the Hermitian adjoint of matrices and operators.) For boosts, however, there is dependence on the vector k, which is the image of the vector (1, 0) under the boost. We refrain from a more detailed description of $U(k, A)$, as this would carry us too far afield.

The unitary SO(1, 3) representation leaves the decomposition $\mathcal{H} = P_+\mathcal{H} \oplus P_-\mathcal{H}$ invariant. On the positive-energy subspace $\mathcal{H}_+$, it is given by

$$(U(L)f)_+(\boldsymbol{p}) = \left(\frac{p_0^L}{p_0}\right)^{1/2} U\left(\frac{\boldsymbol{p}}{m}, A\right)^* f_+(\boldsymbol{p}^L) \tag{37}$$

where

$$p = (E_{\boldsymbol{p}}, \boldsymbol{p}), \quad p^L = L^{-1}p \tag{38}$$

On $\mathcal{H}_-$, it is given by the complex-conjugate representation,

$$(U(L)f)_-(\boldsymbol{p}) = \left(\frac{p_0^L}{p_0}\right)^{1/2} U\left(\frac{\boldsymbol{p}}{m}, A\right)^{\mathrm{t}} f_-(\boldsymbol{p}^L) \tag{39}$$

just as for the spacetime translations, cf. [35]. (The superscript t is used to denote the transpose matrix.) This feature is crucial for the second-quantized Dirac theory, which is discussed next.

The Free Dirac Field in $\mathbb{R}^4$

The free Dirac field is an operator-valued distribution on a Fock space that describes an arbitrary number of spin-1/2 particles and antiparticles in terms of momentum space wave functions. Since spin-1/2 particles are fermions (which encodes the Pauli exclusion principle), an M-particle wave function $F^{+,\ldots,+}_{j_1,\ldots,j_M}(p_1,\ldots,p_M)$ (where $j_l \in \{1, 2\}$ is the spin index) is antisymmetric under any interchange of a pair (j_i, p_i) and (j_k, p_k). Likewise, N-antiparticle wave functions $F^{-,\ldots,-}_{k_1,\ldots,k_N}(q_1,\ldots,q_N)$ are antisymmetric. But a wave function $F^{+-}_{j,k}(p, q)$ describing a particle–antiparticle pair need not have any symmetry property, since a particle and an antiparticle can be distinguished by their charge.

The relevant Fock space is therefore the tensor product of two antisymmetric Fock spaces built over the one-particle and one-antiparticle spaces $L^2(\mathbb{R}^3, \mathrm{d}p) \otimes \mathbb{C}^2$. For later purposes, it is important to view these spaces as the summands $\mathcal{H}_+$ and $\mathcal{H}_-$ of the space $\mathcal{H}$ from the previous section. Thus, the arena for the free Dirac field is the Hilbert space

$$\mathcal{F}_a(\mathcal{H}) \simeq \mathcal{F}_a(\mathcal{H}_+) \otimes \mathcal{F}_a(\mathcal{H}_-) \qquad [40]$$

where, for example,

$$\mathcal{F}_a(\mathcal{H}) = (\mathbb{C} \oplus \mathcal{H} \oplus (\mathcal{H} \otimes \mathcal{H})_a \oplus \cdots)^- \qquad [41]$$

where the bar denotes the completion of the infinite direct sum in the obvious inner product. The tensor $(1, 0, 0, \ldots)$ is viewed as the vacuum (the "filled Dirac sea") and denoted by Ω.

To get around in Fock space, one employs the creation and annihilation operators $c^{(*)}(f)$, $f \in \mathcal{H}$. The creation operator $c^*(f)$, $f \in \mathcal{H}$, is defined by linear and continuous extension of its action on the vacuum Ω and on elementary antisymmetric tensors, recursively given by

$$\begin{aligned} &c^*(f)\Omega = f, \quad c^*(f)f_1 = f \wedge f_1, \ \ldots \\ &c^*(f)f_1 \wedge \cdots \wedge f_N = f \wedge f_1 \wedge \cdots \wedge f_N, \ldots \end{aligned} \qquad [42]$$

Its adjoint, the annihilation operator $c(f)$, satisfies

$$\begin{aligned} &c(f)\Omega = 0, \quad c(f)f_1 = (f, f_1)\Omega, \ \ldots \\ &c(f)f_1 \wedge \cdots \wedge f_N = \sum_{j=1}^{N} (-)^{j-1}(f, f_j) \\ &\qquad \times f_1 \wedge \cdots \wedge \widehat{f_j} \wedge \cdots \wedge f_N, \ldots \end{aligned} \qquad [43]$$

Accordingly, the operators $c^{(*)}(f)$ satisfy the canonical anticommutation relations (CARs) over $\mathcal{H}$,

$$\begin{aligned} &\{c(f), c(g)\} = 0, \\ &\{c(f), c^*(g)\} = (f, g), \quad \forall f, g \in \mathcal{H} \end{aligned} \qquad [44]$$

where $\{A, B\}$ denotes the anticommutator $AB + BA$. (From this, one readily deduces that $c^{(*)}(f)$ is bounded with norm $\|f\|$.)

Next, recalling the direct sum decomposition [22], a notation change

$$c^{(*)}(P_+f) \to a^{(*)}(P_+f), \quad c^{(*)}(P_-f) \to b^{(*)}(P_-f) \qquad [45]$$

is made, thus indicating that $a^{(*)}$ and $b^{(*)}$ should be viewed as the creation/annihilation operators of particles and antiparticles, resp. Since $\mathcal{H}_+$ and $\mathcal{H}_-$ are copies of $L^2(\mathbb{R}^3, \mathrm{d}p) \otimes \mathbb{C}^2$, a given function $(f_1(p), f_2(p))$ in the latter space can occur both as an argument of $a^{(*)}(\cdot)$ and of $b^{(*)}(\cdot)$; it can also be viewed as a smearing function for unsmeared quantities $a_j^{(*)}(p)$ and $b_j^{(*)}(p)$, $j = 1, 2$, that are often referred to as operators as well (even though they are only quadratic forms). Thus, one has, for example,

$$\begin{aligned} b^*(f) &= \sum_{j=1}^{2} \int_{\mathbb{R}^3} \mathrm{d}p\, b_j^*(p) f_j(p) \\ b(f) &= \sum_{j=1}^{2} \int_{\mathbb{R}^3} \mathrm{d}p\, b_j(p) \bar{f}_j(p) \end{aligned} \qquad [46]$$

As explained shortly, the smeared time-zero Dirac field takes the form

$$\Phi(f) = a(P_+f) + b^*(KP_-f), \quad f \in \mathcal{H} \qquad [47]$$

Here and below, K denotes complex conjugation on $\mathcal{H}$, $\mathcal{H}_+$, and $\mathcal{H}_-$. Just as the operators $c^{(*)}(f)$, the operators $\Phi^{(*)}(f)$ satisfy the CARs over $\mathcal{H}$,

$$\begin{aligned} &\{\Phi(f), \Phi(g)\} = 0 \\ &\{\Phi(f), \Phi^*(g)\} = (f, g), \quad \forall f, g \in \mathcal{H} \end{aligned} \qquad [48]$$

as is readily verified using [44]–[45]. But this Φ-representation is not unitarily equivalent to the c-representation [44]. This becomes clear in particular from the consideration of a crucial type of CAR automorphism that is considered next.

To this end, we fix a unitary operator U on $\mathcal{H}$. Then it is plain that the operators

$$\tilde{c}^{(*)}(f) = c^{(*)}(Uf) \qquad [49]$$

$$\tilde{\Phi}^{(*)}(f) = \Phi^{(*)}(Uf) \qquad [50]$$

also satisfy the CARs. The CAR-algebra automorphism $c^{(*)}(f) \mapsto \tilde{c}^{(*)}(f)$ can be unitarily implemented in $\mathcal{F}_a(\mathcal{H})$, since one has

$$\tilde{c}^{(*)}(f) = \Gamma(U)c^{(*)}(f)\Gamma(U^*) \qquad [51]$$

where $\Gamma(U)$ denotes the Fock-space product operator corresponding to U. Thus, for example,

$$\begin{aligned} &\Gamma(U)\Omega = \Omega, \quad \Gamma(U)f = Uf, \dots \\ &\Gamma(U)f_1 \wedge \cdots \wedge f_N = Uf_1 \wedge \cdots \wedge Uf_N, \dots \end{aligned} \qquad [52]$$

For the CAR automorphism $\Phi^{(*)}(f) \mapsto \tilde{\Phi}^{(*)}(f)$ this is not true, however. Rewriting it in terms of the annihilation and creation operators $a^{(*)}$ and $b^{(*)}$ via [47], it amounts to a linear transformation (Bogoliubov transformation), whose unitary implementability has been clarified several decades ago. To be specific, the necessary and sufficient condition for unitary implementability is that the off-diagonal parts

$$U_{+-} = P_+ U P_-, \qquad U_{-+} = P_- U P_+ \qquad [53]$$

in the 2×2 matrix decomposition of operators on $\mathcal{H}$ be Hilbert–Schmidt operators. Therefore, no problem arises when U is diagonal with respect to this decomposition. Indeed, in that case one can choose as unitary implementer the product operator

$$\tilde{\Gamma}(U) = \Gamma(U_{++}) \otimes \Gamma(KU_{--}K) \qquad [54]$$

(cf. the tensor product structure [40] of $\mathcal{F}_a(\mathcal{H})$).

In particular, the automorphism

$$\Phi(f) \mapsto \Phi(e^{itH}f) \qquad [55]$$

where H is the free diagonalized Dirac Hamiltonian [25], is implemented by the operator

$$\tilde{\Gamma}(e^{itH}) = \Gamma(e^{itE}) \otimes \Gamma(e^{itE}) \qquad [56]$$

where E denotes multiplication by E_p on $\mathcal{H}_+$ and $\mathcal{H}_-$. The change of CAR representation, therefore, entails that the unphysical negative energies of the one-particle theory are replaced by positive energies of antiparticles. Hence, we obtain a mathematically precise version of Dirac's hole theory substitution $b_j(\boldsymbol{p}) \to b_j^*(\boldsymbol{p})$, $b_j^*(\boldsymbol{p}) \to b_j(\boldsymbol{p})$.

More generally, if one chooses for U the Poincaré group representation (given by [35] and [37]–[39]), then the Fock-space implementer [54] is the tensor product of two product operators with the same action on $\mathcal{F}_a(L^2(\mathbb{R}^3, d\boldsymbol{p}) \otimes \mathbb{C}^2)$. Observe that this is also true for the Fock-space version $\tilde{\Gamma}(T) = \Gamma(T)$ of the time-reversal operator [34]. By contrast, the Fock-space parity operator $\tilde{\Gamma}(P) = \Gamma(P)$ gives rise to two product operators with slightly different actions, cf. [33]. Accordingly, particles and antiparticles have opposite parity.

The map

$$\Phi(f) \mapsto \Phi(Cf)^* \qquad [57]$$

also yields a CAR automorphism. It is unitarily implemented by the Fock-space charge-conjugation operator

$$\mathcal{C} = \Gamma\left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right) \qquad [58]$$

which interchanges particles and antiparticles. Notice that $\mathcal{C}$ is unitary, whereas C is antiunitary.

It remains to establish the precise relation of the above to the customary free Dirac field $\Psi(t, \boldsymbol{x})$. This is a quadratic form on $\mathcal{F}_a(\mathcal{H})$ given by

$$\begin{aligned} \Psi(t,\boldsymbol{x}) = (2\pi)^{-3/2} \int_{\mathbb{R}^3} d\boldsymbol{p} \sum_{j=1,2} \Big(& a_j(\boldsymbol{p}) w_{+,j}(\boldsymbol{p}) e^{-iE_p t + i\boldsymbol{p}\cdot\boldsymbol{x}} \\ & + b_j^*(\boldsymbol{p}) w_{-,j}(\boldsymbol{p}) e^{iE_p t - i\boldsymbol{p}\cdot\boldsymbol{x}} \Big) \end{aligned} \qquad [59]$$

(Its expectation $\langle F_1, \Psi(t, \boldsymbol{x})F_2 \rangle$ is, for example, well defined for F_1, F_2 in the dense subspace of $\mathcal{F}_a(\mathcal{H})$ that consists of vectors with finitely many particles and antiparticles and wave functions in Schwartz space.) It satisfies the time-dependent Dirac equation

$$i\partial_t \Psi = (-i\alpha \cdot \nabla + \beta m)\Psi \qquad [60]$$

in the sense of quadratic forms. Furthermore, smearing it with a function $\overline{\psi}(\boldsymbol{x})$ in the Hilbert space $\check{\mathcal{H}}$ (7), we obtain

$$\begin{aligned} &\int_{\mathbb{R}^3} d\boldsymbol{x} \overline{\psi}(\boldsymbol{x}) \cdot \Psi(t,\boldsymbol{x}) \\ &\quad = \Phi(e^{itH} W^{-1} \psi) \\ &\quad = \tilde{\Gamma}(e^{itH}) \Phi(W^{-1}\psi) \tilde{\Gamma}(e^{-itH}) \quad \psi \in \check{\mathcal{H}} \end{aligned} \qquad [61]$$

As announced, the time evolution of the free Dirac field is, therefore, given by the unitary one-parameter group [56], whose generator (the second-quantized Dirac Hamiltonian) has spectrum $\{0\} \cup [m, \infty)$.

The Dirac field $\Psi(t, \boldsymbol{x})$ can also be smeared with a test function $F(t, \boldsymbol{x})$ in the Schwartz space $S(\mathbb{R}^4)^4$, yielding a bounded operator

$$\Psi(F) = \int_{\mathbb{R}^4} dx F(x) \cdot \Psi(x) \qquad [62]$$

Then one obtains the relativistic covariance relation

$$\tilde{\Gamma}(U(a,L))\Psi(F)\tilde{\Gamma}(U(a,L))^* = \Psi(F^{a,L}) \qquad [63]$$

where

$$F^{a,L}(x) = S(L^{-1})^t F(L^{-1}(x-a)) \qquad [64]$$

and $U(a, L)$ denotes the Poincaré group representation on $\mathcal{H}$, cf. [35] and [37]–[39]. Likewise, one gets the inversion formulas

$$\tilde{\Gamma}(I)\Psi(F)\tilde{\Gamma}(I)^* = \Psi(F_I), \quad I = P, T \qquad [65]$$

with

$$F_P(t,\boldsymbol{x}) = U_P^t F(t,-\boldsymbol{x}), \quad F_T(t,\boldsymbol{x}) = U_T^t \bar{F}(-t,\boldsymbol{x}) \qquad [66]$$

while the Fock-space charge-conjugation operator [58] transforms the Dirac field as

$$\mathcal{C}\Psi(F)\mathcal{C} = \Psi(F_C)^* \qquad [67]$$

with

$$F_C(x) = U_C^t \bar{F}(x) \qquad [68]$$

Finally, let us consider the global U(1) gauge transformations $f \mapsto e^{i\phi} f$, where $\phi \in \mathbb{R}$ and $f \in \mathcal{H}$. They can be implemented by

$$\tilde{\Gamma}(e^{i\phi}) = \Gamma(e^{i\phi}) \otimes \Gamma(e^{-i\phi}) \qquad [69]$$

and one has

$$\tilde{\Gamma}(e^{i\phi})\Psi(F)\tilde{\Gamma}(e^{i\phi})^* = \Psi(F_\phi) \qquad [70]$$

with

$$F_\phi(x) = e^{-i\phi} F(x) \qquad [71]$$

The generator Q of the one-parameter group $\phi \mapsto \tilde{\Gamma}(e^{i\phi})$ is the charge operator: on wave functions describing N_+ particles and N_- antiparticles, it has eigenvalue $N_+ - N_-$.

More on the One-Particle Dirac Theory

Even for the free one-particle setting, the account given earlier is far from complete. To begin with, the free Dirac equation admits a specialization to massless particles. In the Weyl representation of the γ-algebra adopted above, the choice $m = 0$ entails that the p-space equation [12] decouples into two 2×2 equations for spinors that can be labeled by their chirality ("handedness"). This refers to their eigenvalue with respect to the chirality matrix

$$\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} \mathbf{1}_2 & 0 \\ 0 & -\mathbf{1}_2 \end{pmatrix} \qquad [72]$$

and this notion derives from the noninvariance of the separate 2×2 equations under parity. (A positive-chirality spinor is mapped to a negative-chirality spinor under the parity operator $\check{P}$ (33) and vice versa.) Since the weak interaction breaks parity symmetry, the two 2×2 equations (often called Weyl equations) do have physical relevance. Indeed, the associated quantum fields are a crucial ingredient of the standard model.

Next, we point out that it is possible to switch to a representation in which the gamma matrices are real. This so-called Majorana representation is convenient (but not indispensable) in the description of neutral spin-1/2 particles. By definition, such particles are equal to their antiparticles, so that the second-quantized formalism of the previous section must be adapted: one needs the neutral CAR algebra over $\mathcal{H}$ (also known as self-dual CAR).

For various purposes, it is important to formulate the free Dirac equation for a spacetime whose spatial dimension is arbitrary. Then one needs, first of all, gamma matrices satisfying the (Minkowski) Clifford algebra relations

$$\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2g^{\mu\nu}\mathbf{1}_\Lambda, \quad g = \mathrm{diag}(1, -\mathbf{1}_n) \qquad [73]$$

where n is the space dimension and the minimal size $\Lambda \times \Lambda$ of the gamma matrices is to be determined.

Clearly, for $n = 1$ and $n = 2$, one can take $\Lambda = 2$, choosing, for example,

$$\gamma^0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \gamma^1 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$
$$\gamma^2 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \qquad [74]$$

to fulfill [73]. For $n = 4$, one can take $\Lambda = 4$, just as for $n = 3$, supplementing [1] with the matrix $i\gamma^5$, cf. [72].

More generally, for $n = 2N - 1$ and $n = 2N$, one can take $\Lambda = 2^N$ in [73]. Indeed, a representation on the 2^N-dimensional fermion Fock space $\mathcal{F}_a(\mathbb{C}^N)$ (cf. [41]) is readily constructed using the creation and annihilation operators described in the previous section. Once this has been taken care of, most of the discussion on the free one-particle Dirac equation in $\mathbb{R}^4$ can be easily generalized. Of special importance in this regard is the straightforward adaptation of the formulas [7]–[26], which form the foundation for the second-quantized version. Indeed, the discussion of the last section applies nearly verbatim for arbitrary spacetime dimension.

In several applications, the so-called Euclidean version of the free Dirac theory in spacetime dimension $n + 1$ is important. Basically, this version is obtained upon replacing $i\partial_0$ by ∂_{n+1} in the Dirac equation, a substitution that changes the character of the equation from hyperbolic to elliptic. Provided that the mass vanishes, the Euclidean Dirac equation admits a reinterpretation as a time-independent zero-eigenvalue Weyl equation in a Minkowski spacetime of dimension $n + 2$. (This equation is often called the zero-mode equation.)

Let us now turn to the description of the interaction with an external electromagnetic potential $A_\mu(t, x)$. This can be taken into account via the minimal substitution,

$$\partial_\mu \to \partial_\mu + ieA_\mu \qquad [75]$$

also known as the covariant derivative, in the time-dependent Dirac equation [5].

For the electron in the Coulomb field of a nucleus of charge Ze, one has

$$A_k = 0,\ k = 1,2,3, \quad A_0 = -\frac{Ze}{4\pi|x|} \qquad [76]$$

and the time-independent equation

$$\left(-i\alpha \cdot \nabla + \beta m - \frac{Ze^2}{4\pi|x|}\right)\psi = E\psi \qquad [77]$$

can be solved explicitly. This leads to a bound-state spectrum that is more accurate than its nonrelativistic counterpart. In particular, one finds that energy levels that are degenerate in the nonrelativistic theory split up into slightly different levels. The resulting fine structure of the Dirac levels can be understood as a consequence of the coupling between the spin of the electron and its orbital motion.

In spite of this better agreement with the experimental levels, the physical interpretation of the Dirac electron in a Coulomb field is enigmatic. This is not only because of the persistence of the negative-energy states of the free theory (which turn into scattering states), but also because of unphysical properties of the position operator. More general time-independent external fields (such as step potentials $A_0(x)$ with a step height larger than $2m$) can cause transitions between positive- and negative-energy states (Klein paradox). This phenomenon is enhanced when time dependence is allowed. In particular, any external field that is given by functions in $C_0^\infty(\mathbb{R}^4)$ leads to a scattering operator S on the one-particle space $\mathcal{H}$ [22] that has nonzero off-diagonal parts $S_{\pm\mp}$. Hence, a positive-energy wave packet scattering at such a time- and space-localized field has a nonzero probability to show up as a negative-energy wave packet.

When one tensors the one-particle space $\check{\mathcal{H}}$ with an internal symmetry space $\mathbb{C}^k$, one can also couple external Yang–Mills fields A_μ taking values in the $k \times k$ matrices via the substitution [75]. (From a geometric viewpoint, this can be rephrased as tensoring the spinor bundle with a vector bundle equipped with a connection A.) The generalization of this external gauge field coupling to a Minkowski spacetime or Euclidean space of arbitrary dimension is straightforward. An adaptation of the resulting interacting one-particle Dirac theory in arbitrary dimension to quite general geometric settings also yields a crucial starting point for index theory.

Before turning to the latter area, we conclude this section with another striking application of the one-particle framework, namely the massless Dirac equation in two spacetime dimensions with special external fields. Specifically, the relevant Dirac operator is of the form

$$\begin{pmatrix} i\frac{d}{dx} & -iq(x) \\ ir(x) & -i\frac{d}{dx} \end{pmatrix} \qquad [78]$$

where $r(x)$ and $q(x)$ are not necessarily real valued. (Note that this operator is in general not self-adjoint.) With suitable restrictions on r and q, the direct and inverse scattering theory associated with the Dirac operator [78] can be applied to various nonlinear PDEs in two spacetime dimensions to solve their Cauchy problems in considerable detail. As a crucial special case, initial conditions yielding vanishing reflection give rise to soliton solutions for the pertinent equation.

The first example in this framework was found by Zakharov and Shabat (the nonlinear Schrödinger equation); with other choices of r and q several other soliton PDEs (including the sine-Gordon and modified Korteweg–de Vries equations) were handled by Ablowitz, Kaup, Newell; and Segur, who studied a quite general class of external fields r and q.

The Dirac Operator and Index Theory

Thus far, we have considered various versions of the Dirac operator associated with the spaces $\mathbb{R}^l$ for some $l \geq 1$. For applications in the area of index theory, however, one needs to generalize this base manifold. Indeed, one can define a Dirac operator for any l-dimensional oriented Riemannian manifold $\mathcal{M}$ that admits a spin structure. This is a lifting of the transition functions of the tangent bundle $T\mathcal{M}$ (which may be assumed to take values in $SO(l)$) to the simply connected twofold cover $Spin(l)$ (taking $l \geq 3$).

Choosing first $l = 2N + 1$, the spin group has a faithful irreducible representation on $\mathbb{C}^{2^N}$. Hence, one obtains a $\mathbb{C}^{2^N}$-bundle over $\mathcal{M}$, the spinor bundle. The Levi-Civita connection on $\mathcal{M}$ derived from the metric can now be lifted to a connection on the spinor bundle. From the covariant derivative corresponding to the spin connection and the

Clifford algebra generators $\gamma^1, \ldots, \gamma^l$, one can then construct a first-order elliptic differential operator that acts on sections of the spinor bundle. (For the case $\mathcal{M} = \mathbb{R}^{2N+1}$ with its Euclidean metric, this construction yields the massless positive-chirality Dirac operator acting on wave functions with 2^N components, as considered above.)

The massless Dirac operator thus obtained is self-adjoint as an operator on the L^2-space $\mathcal{H}$ associated with the spinor bundle, and it has infinite-dimensional positive and negative spectral subspaces $\mathcal{H}_+$ and $\mathcal{H}_-$. (In this section the check accent on position-space quantities is omitted.) Specializing to the case of compact $\mathcal{M}$, a continuous map from $\mathcal{M}$ to $\mathbb{C}^*$ gives rise to a Fredholm operator on $\mathcal{H}_+$, and more generally a continuous map from $\mathcal{M}$ to $\mathrm{GL}(k, \mathbb{C})$ yields a Fredholm operator on $\mathcal{H}_+ \otimes \mathbb{C}^k$.

For a smooth map, the Fredholm index of this operator can be written in terms of an integral over $\mathcal{M}$ involving certain closed differential forms. The value of this integral does not change when exact forms are added, since $\mathcal{M}$ has no boundary. Hence, one is dealing with de Rham cohomology classes. In this context, the class involved ("characteristic class") is determined by the Riemann curvature tensor of $\mathcal{M}$ and the topological ("winding") characteristics of the map.

The simplest example of this state of affairs arises for $l = 1$ and $\mathcal{M} = S^1$ with its obvious spin structure (periodic boundary conditions). Writing $\psi \in \mathcal{H} = L^2(S^1)$ as

$$\psi(z) = \sum_{n \in \mathbb{Z}} a_n z^n, \quad z \in S^1 \qquad [79]$$

the Dirac operator H on $\mathcal{H}$ reads

$$H = z \frac{\mathrm{d}}{\mathrm{d}z} \qquad [80]$$

It has eigenfunctions z^n, $n \in \mathbb{Z}$. Thus, we may choose

$$(P_+\psi)(z) = \sum_{n \geq 0} a_n z^n, \quad (P_-\psi)(z) = \sum_{n<0} a_n z^n \qquad [81]$$

As a consequence, the functions in $\mathcal{H}_+(\mathcal{H}_-)$ are L^2-boundary values of holomorphic functions in $|z| < 1$ ($|z| > 1$). Operators of the form

$$T_\psi = P_+ M_\psi P_+ \qquad [82]$$

where ψ is a continuous function on S^1 and M_ψ denotes multiplication by ψ, are called Toeplitz operators. It is not hard to see that they are Fredholm (viewed as operators on $\mathcal{H}_+$), provided that ψ does not vanish on S^1. (Recall a bounded operator B is Fredholm if it has finite-dimensional kernel K and cokernel C. Its Fredholm index is given by

$$\mathrm{index}(B) = \dim K - \dim C \qquad [83]$$

and is norm continuous and invariant under addition of a compact operator.) Assuming $\psi(S^1) \subset \mathbb{C}^*$ from now on, the curve $\psi(S^1)$ has a well-defined winding number $w(\psi)$ with respect to the origin. The equality

$$\mathrm{index}(T_\psi) = -w(\psi) \qquad [84]$$

between objects from the area of analysis on the left-hand side and from the areas of topology and geometry on the right-hand side is the simplest example of an Atiyah–Singer type index formula. When ψ is not only continuous but also smooth, the index formula can be rewritten as

$$\mathrm{index}(T_\psi) = -\frac{1}{2\pi \mathrm{i}} \int_{S^1} \frac{\mathrm{d}\psi}{\psi} \qquad [85]$$

yielding a characteristic class version.

It should be noted that the operator M_ψ on $\mathcal{H}$ has a bounded inverse $M_{1/\psi}$ when $0 \notin \psi(S^1)$, hence a trivially vanishing index. Therefore, the compression [82] involving the spectral projection of the Dirac operator is needed to get a nonzero index. Observe also that the equality [84] is quite easily verified for the case $\psi(z) = z^n$, since T_ψ yields a power of the right ($n > 0$) or left ($n < 0$) shift on $P_+\mathcal{H} \simeq l^2(\mathbb{N})$.

We proceed to the case of even-dimensional manifolds, $l = 2N$. Then the fiber $\mathbb{C}^{2^N}$ of the spinor bundle splits into a direct sum of even and odd spinors, corresponding to two distinct representations of $\mathrm{Spin}(2N)$ on $\mathbb{C}^{2^{N-1}}$. (Here it is assumed that $N > 3$; recall the Lie algebra isomorphisms $\mathrm{so}(4) \simeq \mathrm{so}(3) \oplus \mathrm{so}(3)$ and $\mathrm{so}(6) \simeq \mathrm{su}(4)$.) With respect to this decomposition, the Dirac operator can be written as

$$H = \begin{pmatrix} 0 & \mathcal{D}^* \\ \mathcal{D} & 0 \end{pmatrix} \qquad [86]$$

where $\mathcal{D}$ and $\mathcal{D}^*$ are again first-order elliptic differential operators expressed in terms of Clifford algebra generators and the spin connection. Tensoring the spinor bundle with a vector bundle equipped with a connection A, one can define a Dirac operator on the tensor product which involves A and takes the form

$$H_A = \begin{pmatrix} 0 & \mathcal{D}_A^* \\ \mathcal{D}_A & 0 \end{pmatrix} \qquad [87]$$

with respect to the even/odd spinor decomposition. Once more, the index of $\mathcal{D}_A$ (viewed as a Fredholm operator between two different Hilbert spaces) can be expressed as an integral over $\mathcal{M}$ involving characteristic classes that depend on the curvatures of the two connections.

Probably the simplest example of the constructions just sketched is given by the torus $\mathcal{M}=S^1\times S^1$ with its flat metric. Employing the above coordinate and spin structure on S^1, one can take

$$\mathcal{H}=L^2(S^1\times S^1)\otimes\mathbb{C}^2,\quad \mathcal{D}=z_1\frac{\partial}{\partial z_1}+\mathrm{i}z_2\frac{\partial}{\partial z_2} \qquad [88]$$

Since the curvature vanishes, the index theorem for this situation implies $\mathrm{index}(\mathcal{D})=0$. (Note that this is also plain from [88]: both kernel and cokernel of $\mathcal{D}$ are spanned by the constant sections.) On the other hand, when one tensors the spinor bundle with a line bundle with connection A, the index formula reads

$$\mathrm{index}(\mathcal{D}_A)=-\frac{1}{2\pi}\int_{S^1\times S^1}F \qquad [89]$$

where F is the curvature 2-form corresponding to A.

The Atiyah–Singer index theorem for Dirac operators has far-reaching applications. It can be used to derive other results in this area, such as the Gauss–Bonnet–Chern theorem, the Hirzebruch signature theorem, and (when $\mathcal{M}$ is a Kähler manifold) Riemann–Roch type theorems. From this, one can obtain information on various questions, such as the existence of positive scalar curvature metrics or zeros of vector fields on $\mathcal{M}$. Other applications include insights on topological invariants of manifolds obtained from "simple" manifolds (such as spheres and tori) by glueing or covering operations. This hinges on the additive properties of the index that are clear from its being given by an integral over the manifold. Conversely, the integrality of Fredholm indices can be used to deduce that certain rational cohomology classes are actually integral on manifolds that admit the structure that is required for the pertinent index theorem to apply, that certain manifolds do not admit such structures, since one knows that the relevant class is not integral, etc.

More on the Dirac Field

As mentioned earlier, the free-field formalism can be easily generalized to an arbitrary spacetime dimension d. For $d>4$, however, no renormalizable interacting quantum field models involving the Dirac field are known. For the physical case $d=4$ the standard model involves various Dirac fields interacting with quantized gauge fields and Klein–Gordon fields. Although its perturbation theory is renormalizable, its mathematical existence is to date wide open.

It is far beyond the scope of this article to elaborate on the analytical difficulties of relativistic quantum field theories, let alone those associated with the standard model. Even for $d=2$ and 3, a nonperturbative construction of interacting quantum field models involving the Dirac field is an extremely difficult enterprise. Apart from some rigorous results on certain self-interacting Dirac field models, the only interacting model that is reasonably well understood from the constructive field theory viewpoint is the Yukawa model for $d=2$ and 3. This describes the interaction between the Dirac field Ψ and a Klein–Gordon field ϕ, the interaction term being formally given by $g(\Psi^*\gamma^0\Psi)\phi$.

On the other hand, the interaction of the quantized Dirac field with external classical fields is much more easily understood and analytically controlled. As a bonus, within this context, one can make contact with various issues of physical and mathematical relevance. We now proceed to sketch the external-field framework and some of its applications.

Let us first consider the addition of an external field term $gV(t,\boldsymbol{x})$ to the free Dirac operator $\check{H}$ on

$$\check{\mathcal{H}}=L^2(\mathbb{R}^n,\mathrm{d}\boldsymbol{x})\otimes\mathbb{C}^\Lambda\otimes\mathbb{C}^k \qquad [90]$$

We assume from now on that the coupling g is real and that V is a self-adjoint $k\Lambda\times k\Lambda$ matrix-valued function on spacetime $\mathbb{R}^{n+1}$ with matrix elements that are in $C_0^\infty(\mathbb{R}^{n+1})$. Then the (interaction picture) *scattering operator S exists. It is unitary and has off-*diagonal Hilbert–Schmidt parts $S_{\pm\mp}$, so that a unitary Fock-space S-operator $\tilde{\Gamma}(S)$ implementing the Bogoliubov transformation generated by S *exists:*

$$\tilde{\Gamma}(S)\Phi(f)\tilde{\Gamma}(S)^*=\Phi(Sf),\quad \forall f\in\mathcal{H} \qquad [91]$$

The arbitrary phase in $\tilde{\Gamma}(S)$ can be fixed by requiring that the vacuum expectation value of $\tilde{\Gamma}(S)$ be positive. More precisely, this number is generically nonzero and satisfies

$$|(\Omega,\tilde{\Gamma}(S)\Omega)|=\det(1+T_S)^{-1/2} \qquad [92]$$

where T_S is a positive trace class operator determined by S.

The vector $\tilde{\Gamma}(S)\Omega$ is a superposition of wave functions with an equal and arbitrary number of particles and antiparticles. More generally, the Fock-space S-operator $\tilde{\Gamma}(S)$ leaves the subspaces of $\mathcal{F}_a(\mathcal{H})$ with a fixed eigenvalue $q\in\mathbb{Z}$ of the charge operator Q invariant, and can create and

annihilate an arbitrary number of particle–antiparticle pairs.

The unitary propagator $U(T_1, T_2)$ corresponding to $V(t, \boldsymbol{x})$ does not have Hilbert–Schmidt off-diagonal parts (unless the spacetime dimension is sufficiently small and special external fields are chosen). Even so, the diagonal parts are Fredholm with vanishing index, and the off-diagonal parts are compact. Omitting the ill-defined determinantal factor, these properties imply that one obtains a renormalized quadratic form $\tilde{\Gamma}_{\text{rcn}}(U(T_1, T_2))$ satisfying the implementing relation

$$\tilde{\Gamma}_{\text{rcn}}(U(T_1, T_2))\Phi(f) = \Phi(U(T_1, T_2)f)\tilde{\Gamma}_{\text{rcn}}(U(T_1, T_2)), \quad \forall f \in \mathcal{H} \qquad [93]$$

in the quadratic form sense.

The above unitary operators on $\mathcal{H}$ yield Fredholm diagonal parts whose indices vanish. (They are norm continuous in g and reduce to the identity for $g = 0$.) This is why their Fock-space implementers leave the charge sectors invariant. Indeed, for a unitary operator U on $\mathcal{H}$ with compact off-diagonal parts the implementer maps the charge-q sector to the charge-$(q + q(U))$ sector, where

$$q(U) = \text{index}(U_{--}) \qquad [94]$$

Specializing to the case

$$n = 2N - 1, \quad \Lambda = 2\lambda, \quad \lambda = 2^{N-1} \qquad [95]$$

a unitary $(k\Lambda \times k\Lambda)$-matrix multiplier $\check{U}$ on $\check{\mathcal{H}}$ does not have compact off-diagonal parts in general. But when it is of the form

$$\check{U} = \begin{pmatrix} \mathbf{1}_\lambda \otimes u_+(\boldsymbol{x}) & 0 \\ 0 & \mathbf{1}_\lambda \otimes u_-(\boldsymbol{x}) \end{pmatrix} \qquad [96]$$

with respect to the chiral decomposition (the generalization of the γ^5-decomposition [72] to even spacetime dimension), then it suffices for compactness of the off-diagonal parts that the matrices $u_\pm(\boldsymbol{x}) \in \text{U}(k)$ are continuous and converge to $\mathbf{1}_k$ for $|\boldsymbol{x}| \to \infty$.

Viewing $\mathbb{R}^{2N-1}$ as arising from S^{2N-1} via stereographic projection, the latter unitaries can be viewed as continuous maps from S^{2N-1} to $\text{U}(k)$, reducing to $\mathbf{1}_k$ at the north pole. As such, they yield elements of the homotopy group $\pi_{2N-1}(\text{U}(k))$. By virtue of Bott's periodicity theorem, the latter group equals $\mathbb{Z}$ for $k \geq N$. Thus, the maps $u_\pm$ have a well-defined "winding number" $w(u_\pm) \in \mathbb{Z}$ for $k \geq N$. From the index formula

$$\text{index}(U_{--}) = w(u_+) - w(u_-) \qquad [97]$$

and [94] one now deduces that one can obtain implementers $\tilde{\Gamma}_{\text{rcn}}(U)$ effecting a nonzero charge change from unitary maps with nonzero winding number.

In particular, choosing $k = \lambda = 2^{N-1} \geq N$, there exist quite special "kink maps"

$$u_{\epsilon,a}(\boldsymbol{x}) \in \text{U}(\lambda), \quad \epsilon > 0, \ a \in \mathbb{R}^{2N-1} \qquad [98]$$

with winding number 1 and such that the quadratic form implementers of the unitary multiplication operators

$$\check{U}_{+,\epsilon,a} = -\begin{pmatrix} \mathbf{1}_\lambda \otimes u_{\epsilon,a}(\boldsymbol{x}) & 0 \\ 0 & \mathbf{1}_\lambda \otimes \mathbf{1}_\lambda \end{pmatrix}$$
$$\check{U}_{-,\epsilon,a} = -\begin{pmatrix} \mathbf{1}_\lambda \otimes \mathbf{1}_\lambda & 0 \\ 0 & \mathbf{1}_\lambda \otimes u_{\epsilon,\text{a}}(-\boldsymbol{x}) \end{pmatrix} \qquad [99]$$

converge to (a linear combination of the chiral components of) the free Dirac field $\Psi(0, \boldsymbol{a})$ as the kink size parameter ϵ goes to 0.

For the special case $N = 1$, one can take

$$u_{\epsilon,a}(x) = \frac{x - a - \text{i}\epsilon}{x - a + \text{i}\epsilon} \qquad [100]$$

and the off-diagonal parts of $U_{\pm,\epsilon,a}$ are actually Hilbert–Schmidt. Thus, the implementers can be chosen to be unitary operators. But to get convergence to the Dirac field components $\Psi(0, a)_\pm$ as $\epsilon \to 0$, the unitary implementers $\tilde{\Gamma}(U_{\pm,\epsilon,a})$ should be renormalized by a multiplicative factor.

For the $N = 1$ case, the unitary multipliers [96] give rise to loop groups. Indeed, requiring

$$\lim_{x \to \pm\infty} u_\delta(x) = 1_k, \quad \delta = +, - \qquad [101]$$

we are dealing with continuous maps $S^1 \to \text{U}(k)$. From the viewpoint of the Dirac theory, these groups are local gauge groups. The convergence to the Dirac field just sketched can be used to great advantage to clarify the structure of the corresponding Fock-space gauge groups. Their Lie algebras yield representations of Kac–Moody algebras, a topic which is considered shortly.

Before doing so, it should be pointed out that under some mild smoothness assumptions all of the above unitary matrix multipliers can also be viewed as S-operators associated with very special external fields. Indeed, the gauge-transformed Dirac operator

$$\check{H}_U = \check{U}^* \check{H} \check{U} \qquad [102]$$

is of the form

$$\check{H}_U = \check{H} + V(\boldsymbol{x}) \qquad [103]$$

where $V(x)$ is a self-adjoint $k\Lambda \times k\Lambda$ matrix on $\mathbb{R}^{2N-1}$ (a "pure gauge" field). If one now defines a time-dependent external field by

$$V(t,x) = \begin{cases} V(x), & t \geq 0 \\ 0, & t < 0 \end{cases} \qquad [104]$$

then $\check{U}$ equals the S-operator for $V(t, x)$. (Equivalently, $\check{U}$ is the $t \to \infty$ wave operator for the time-independent external field $V(x)$.)

To conclude this section we sketch some applications of the second-quantized Dirac formalism for the special case $N = 1$, $m = 0$, and positive chirality. Even though we could stick to the massless positive-chirality Dirac operator $-\mathrm{id}/\mathrm{d}x$ on the line, it is simpler and more natural to start from its counterpart on the circle already considered in the last section, cf. [80]. (Under the Cayley transform, the positive- and negative-energy subspaces of $-\mathrm{id}/\mathrm{d}x$ on $L^2(\mathbb{R})$ correspond to those of $z\mathrm{d}/\mathrm{d}z$ on $L^2(S^1)$, given by [81].) Letting $z = \mathrm{e}^{\mathrm{i}\theta}$, we then obtain

$$\begin{aligned} &\check{H} = -\mathrm{id}/\mathrm{d}\theta, \quad \check{\mathcal{H}} = L^2([0,2\pi], \mathrm{d}\theta) \\ &\mathcal{H} = l^2(\mathbb{Z}), \quad \mathcal{H}_+ = l^2(\mathbb{N}), \quad \mathcal{H}_- = l^2(\mathbb{Z}_-) \end{aligned} \qquad [105]$$

and a corresponding Dirac field

$$\Psi(t,\theta) = (2\pi)^{-1/2}\left(\sum_{n=0}^{\infty} a_n \mathrm{e}^{-\mathrm{i}nt+\mathrm{i}n\theta} + \sum_{n=1}^{\infty} b^*_{-n}\mathrm{e}^{\mathrm{i}nt-\mathrm{i}n\theta}\right)$$
$$(t,\theta) \in \mathbb{R} \times [0,2\pi] \qquad [106]$$

where

$$a_l = c(e_l),\ l \geq 0, \quad b_l = c(e_l),\ l < 0 \qquad [107]$$

and $\{e_l\}_{l\in\mathbb{Z}}$ is the canonical basis of $l^2(\mathbb{Z})$.

Consider now the group $\mathrm{GL}(\mathcal{H})$ of bounded operators on $\mathcal{H}$ with bounded inverses. The transformation

$$\Phi^*(f) \mapsto \Phi^*(Gf), \quad \Phi(f) \mapsto \Phi(G^{-1*}f)$$
$$f \in \mathcal{H}, \quad G \in GL(\mathcal{H}) \qquad [108]$$

leaves the CAR [48] invariant. Provided that G belongs to the subgroup

$$G_2(\mathcal{H}) = \{G \in \mathrm{GL}(\mathcal{H}) \mid G_{\pm\mp} \text{ Hilbert–Schmidt}\} \qquad [109]$$

there exists an implementer $\tilde{\Gamma}(G)$ on $\mathcal{F}_a(\mathcal{H})$:

$$\begin{aligned} &\tilde{\Gamma}(G)\Phi^*(f) = \Phi^*(Gf)\tilde{\Gamma}(G), \\ &\tilde{\Gamma}(G)\Phi(f) = \Phi(G^{-1*}f)\tilde{\Gamma}(G), \quad \forall f \in \mathcal{H} \end{aligned} \qquad [110]$$

In particular, the multiplication operator

$$\exp(h(x)), \quad h(x) = \sum_{k=1}^{\infty} x_k z^{-k}, \quad z = \mathrm{e}^{\mathrm{i}\theta} \qquad [111]$$

belongs to $G_2(\mathcal{H})$ provided the sequence x_k vanishes sufficiently fast as $k \to \infty$. Thus, one obtains an implementer $\tilde{\Gamma}(\mathrm{e}^{h(x)})$, the so-called KP evolution operator. This designation is justified by the vacuum expectation value

$$\tau(x) = (\Omega, \tilde{\Gamma}(\mathrm{e}^{h(x)})\tilde{\Gamma}(G)\Omega), \quad G \in G_2(\mathcal{H}) \qquad [112]$$

being a tau-function solving the hierarchy of KP evolution equations in Hirota bilinear form, as first shown by Sato and his Kyoto school. For example, the KP equation itself,

$$u_{yy} = \partial_x\left(\tfrac{4}{3}u_t - 2uu_x - \tfrac{1}{3}u_{xxx}\right) \qquad [113]$$

has the bilinear form

$$\left(\frac{\partial^4}{\partial y_1^4} - 4\frac{\partial}{\partial y_1}\frac{\partial}{\partial y_3} + 3\frac{\partial^2}{\partial y_2^2}\right)\tau(x+y)\tau(x-y)\Big|_{y=0} = 0 \qquad [114]$$

the relation being given by

$$x_1 = x, \quad x_2 = y, \quad x_3 = t, \quad u = 2\partial_1^2 \ln \tau \qquad [115]$$

The class of solutions to [113] thus obtained includes not only the rational and soliton solutions (which correspond to choosing $\check{G}$ as multiplication by a rational function of $z = \mathrm{e}^{\mathrm{i}\theta}$ that does not vanish on S^1), but also the finite-gap solutions associated with compact Riemann surfaces. Moreover, for suitable subgroups of $G_2(\mathcal{H})$, one obtains tau-functions for related soliton hierarchies, including the Korteweg–de Vries, Boussinesq and Hirota–Satsuma hierarchies. Even though the class of solutions associated with $G_2(\mathcal{H})$ via the Dirac formalism is large, it should be noted that from the perspective of the Cauchy problem for the pertinent evolution equations the solutions are nongeneric, inasmuch as the initial data are real-analytic functions.

Finally, we consider Lie algebra representations related to the above special starting point [105] for the second-quantized Dirac framework. Assume that $\exp(tA)$ is a one-parameter group of bounded operators on $\mathcal{H}$ with generator A in the Lie algebra of $G_2(\mathcal{H})$,

$$g_2(\mathcal{H}) = \{A \text{ bounded} \mid A_{\pm\mp} \text{ Hilbert–Schmidt}\} \qquad [116]$$

Then one can take

$$\tilde{\Gamma}(\exp(tA)) = \exp(t\mathrm{d}\tilde{\Gamma}(A)) \qquad [117]$$

where $\mathrm{d}\tilde{\Gamma}(A)$ is the Fock-space operator uniquely determined up to an additive constant by its commutation relation

$$[\mathrm{d}\tilde{\Gamma}(A), \Phi^*(f)] = \Phi^*(Af), \quad \forall f \in \mathcal{H} \qquad [118]$$

with the smeared Dirac field $\Phi^*(f)$. Fixing the constant by requiring

$$(\Omega, \mathrm{d}\tilde{\Gamma}(A)\Omega) = 0 \qquad [119]$$

the map $A \mapsto \mathrm{d}\tilde{\Gamma}(A)$ satisfies the Lie algebra relations

$$[\mathrm{d}\tilde{\Gamma}(A), \mathrm{d}\tilde{\Gamma}(B)] = \mathrm{d}\tilde{\Gamma}([A, B]) + C(A, B)\mathbf{1} \qquad [120]$$

so that the term

$$C(A, B) = \mathrm{tr}(A_{-+}B_{+-} - B_{-+}A_{+-}) \qquad [121]$$

encodes a central extension of the Lie algebra $g_2(\mathcal{H})$ [116].

The developments sketched in the previous paragraph are in fact independent of the specific form of the Hilbert space $\mathcal{H}$ and its $\mathcal{H}_+/\mathcal{H}_-$ decomposition. But the special feature of the choice [105] and its $S^1 \to \mathbb{R}$ analog is that the smeared Dirac current

$$\int_0^{2\pi} \mathrm{d}\theta\, \psi(\theta) :\Psi^*(0,\theta)\Psi(0,\theta):, \quad \psi \in C^\infty(S^1) \qquad [122]$$

(where the double dots denote normal ordering – the replacement of terms involving $b_k b_l^*$ by $-b_l^* b_k$) is of the form $\mathrm{d}\tilde{\Gamma}(A_\psi)$ with $A_\psi \in g_2(\mathcal{H})$ determined by ψ. (For spacetime dimension $d > 2$, this is no longer true, as the Hilbert–Schmidt condition is violated.) Moreover, [120] reduces to

$$[\mathrm{d}\tilde{\Gamma}(A_\psi), \mathrm{d}\tilde{\Gamma}(A_\phi)] = C(A_\psi, A_\phi)\mathbf{1} \qquad [123]$$

with the central extension explicitly given by

$$C(A_\psi, A_\phi) = \frac{\mathrm{i}}{2\pi}\int_0^{2\pi} \mathrm{d}\theta\, \psi'(\theta)\phi(\theta) \qquad [124]$$

We have just sketched the details of the (simplest version of the) Dirac current algebra: the term [124] is commonly known as the Schwinger term, so that the central extension featuring in [120]–[121] may be viewed as a generalization. The above setup can also be slightly generalized so as to obtain representations of the Virasoro algebra, which is a central extension of the Lie algebra of polynomial vector fields on S^1. The general framework has a quite similar version for the neutral Dirac field (Majorana field), described in terms of the self-dual CAR algebra. In the neutral setting, one can construct the Neveu–Schwarz and Ramond representations of the Virasoro algebra, which are crucial in string theory.

Tensoring $\check{\mathcal{H}}$ with an internal symmetry space $\mathbb{C}^k$ and starting from the Lie algebra of rational maps $S^1 \to \mathrm{sl}(k, \mathbb{C})$, $z \mapsto M(z)$, with poles occurring solely at $z = 0$ and $z = \infty$ (regarded as multiplication operators on $L^2(S^1)^k$), the Fock-space counterparts obtained via the $\mathrm{d}\tilde{\Gamma}$-operation yield representations of the Kac–Moody Lie algebra $A_{k-1}^{(1)}$. Specifically, on the charge-0 sector of $\mathcal{F}_a(\mathcal{H})$, one obtains the so-called basic representation, whereas the charge-q sectors with $q = 1, \ldots, k-1$, yield the fundamental representations. Using the neutral version of Dirac's second quantization, one can also obtain the basic and a fundamental representation of the Kac–Moody algebras $B_l^{(1)}$ (for $k = 2l+1$) and $D_l^{(1)}$ (for $k = 2l$).

See also: Bosons and Fermions in External Fields; Clifford Algebras and Their Representations; Current Algebra; Dirac Fields in Gravitation and Nonabelian Gauge Theory; Gerbes in Quantum Field Theory; Holonomic Quantum Fields; Index Theorems; Quantum Field Theory in Curved Spacetime; Quantum Chromodynamics; Random Walks in Random Environments; Relativistic Wave Equations Including Higher Spin Fields; Solitons and Kac–Moody Lie Algebras; Spinors and Spin Coefficients; Symmetry Classes in Random Matrix Theory.

Further Reading

Bjorken JD and Drell SD (1964) *Relativistic Quantum Mechanics*. New York: McGraw-Hill.

Carey AL and Ruijsenaars SNM (1987) On fermion gauge groups, current algebras and Kac–Moody algebras. *Acta Applicandae Mathematicae* 10: 1–86.

Date E, Jimbo M, Kashiwara M, and Miwa T (1983) Transformation groups for soliton equations. In: Jimbo M and Miwa T (eds.) *Proceedings of RIMS Symposium, Nonlinear Integrable Systems – Classical Theory and Quantum Theory*, pp. 39–119. Singapore: World Scientific.

Dirac PAM (1928) The quantum theory of the electron. *Proceedings of the Royal Society of London. Series A* 117: 610–624.

Dirac PAM (1928) The quantum theory of the electron, II. *Proceedings of the Royal Society of London. Series A* 118: 351–361.

Glimm J and Jaffe A (1981) *Quantum Physics*. New York: Springer.

Itzykson C and Zuber JB (1980) *Quantum Field Theory*. New York: McGraw-Hill.

Pressley A and Segal G (1986) *Loop Groups*. Oxford: Clarendon.

Rose ME (1961) *Relativistic Electron Theory*. New York: Wiley.

Ruijsenaars SNM (1989) Index formulas for generalized Wiener–Hopf operators and boson–fermion correspondence in $2N$ dimensions. *Communications in Mathematical Physics* 124: 553–593.

Schweber SS (1961) *An Introduction to Relativistic Quantum Field Theory*. Evanston, IL: Row-Peterson.

Streater RF and Wightman AS (1964) *PCT, Spin and Statistics, and All That*. New York: Benjamin.

Thaller B (1992) *The Dirac Equation*. New York: Springer.

Dispersion Relations

J Bros, CEA/DSM/SPhT, CEA/Saclay, Gif-sur-Yvette, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Dispersion relations constitute a basic chapter of mathematical physics which covers various types of classical and quantum scattering phenomena and illustrates in a typical way the importance of general principles in theoretical physics, among which causality plays a major role. Each such phenomenon is described in terms of a scattering amplitude $F(\omega)$, which is a complex-valued function of a frequency variable ω; in quantum physics, this variable becomes an energy variable called E (or s in particle physics), as it follows from the fundamental de Broglie relation $E = \hbar\omega$. The real and imaginary parts of $F(\omega)$, which are called respectively the dispersive part $D(\omega)$ and the absorptive part $A(\omega)$ of F, have well-defined physical interpretations for all these phenomena; they represent quantities which are essentially accessible to measurements. The term dispersion relations refers to linear integral equations which relate the functions $D(\omega)$ and $A(\omega)$; such integral equations are always closely related to the Cauchy integral representation of a subjacent holomorphic function $\hat{F}(\omega^{(c)})$ of the complexified frequency (or energy) variable $\omega^{(c)}$. $\hat{F}(\omega^{(c)})$ is called the holomorphic scattering function or in short the scattering function, and the scattering amplitude appears as the boundary value of the latter, taken at positive real values of ω from the upper half-plane of $\omega^{(c)}$, namely

$$F(\omega) = \lim_{\varepsilon \to 0} \hat{F}(\omega + i\varepsilon), \quad \varepsilon > 0$$

Historically, the first relations of that type to be obtained were the Kramers–Krönig relations (1926), which concern the propagation of light in a dielectric medium. In this basic example, $F(\omega)$ represents the complex refractive index of the medium $n'(\omega) = n(\omega) + i\kappa(\omega)$ for a monochromatic wave with frequency ω. The dispersive part $D(\omega)$ is the real refractive index $n(\omega)$, which is the inverse ratio of the phase velocity of the wave in the medium to its velocity c in the vacuum: the fact that it depends on the frequency ω corresponds precisely to the phenomenon of dispersion of light in a dielectric medium. A slab of the latter thus appears as a prototype of a macroscopic scatterer. The absorptive part $A(\omega)$ is the rate of exponential damping $\kappa(\omega)$ of the wave, caused by the absorption of energy in the medium.

It has appeared much later that for many scattering phenomena, dispersion relations can be derived from an appropriate set of general physical principles. This means that inside a certain axiomatic framework these relations are model independent with respect to the detailed structure of the scatterer or to the detailed type of particle interaction in the quantum case.

In a very short and oversimplifying way, the following logical scheme holds. At first, one can say that any mathematical formulation of a physical principle of causality results in support-type properties with respect to a time variable t of an appropriate "causal structural function" $R(t)$ of the physical system considered: typically, such a causal function should vanish for negative values of t. It follows that its Fourier transform $\tilde{R}$ admits an analytic continuation $\tilde{R}^{(c)}$ in the upper half-plane of the corresponding conjugate variable, interpreted as a frequency (or an energy in the quantum case): here is the general reason for the occurrence of complex frequencies and of holomorphic functions of such variables. In fact, the relevant holomorphic scattering function $\hat{F}(\omega^{(c)})$ always appears as generated by $\tilde{R}^{(c)}$ via some (more or less sophisticated) procedure: in the simplest case, $\hat{F}$ coincides with $\tilde{R}^{(c)}$ itself, but this is not so in general. Finally, the derivation of suitable analyticity and boundedness properties of $\hat{F}(\omega^{(c)})$ in a domain whose typical form is the upper half-plane, allows one to apply a Cauchy-type integral representation to this function; the dispersion relations directly follow from the latter.

The first part of this article aims to describe the most typical dispersion relations and their link with the Cauchy integral. It then presents two basic illustrations of these relations, which are: (1) in classical physics, the Kramers–Krönig relations mentioned above, and (2) in quantum physics, the dispersion relations for the forward scattering of equal-mass particles. The aim of the subsequent parts is to give as complete as possible accounts of the derivation of the relevant analyticity domains inside appropriate axiomatic frameworks which, respectively, contain the previous two examples. The simplest axiomatic framework is the one which governs all the phenomena of linear response: in the latter, the proof of analyticity and dispersion relations most easily follows the logical line sketched above. It will be presented together with its application to the derivation of

the Kramers–Krönig relations. The rest of the article is devoted to the derivation of the so-called crossing analyticity domains which are the relevant background of dispersion relations for the two-particle scattering (or collision) amplitudes in particle physics. This derivation relies on the general axiomatic framework of relativistic quantum field theory (QFT) (*see* Axiomatic Quantum Field Theory) and more specifically on the "analytic program in complex momentum space" of the latter. This framework, whose rigorous mathematical form has been settled around 1960, represents the safest conceptual approach for describing the particle collision processes in a range of energies which covers by far all those that can be produced and will be produced in the accelerators for several decades. A simple account of the field-theoretical axiomatic framework and of the logical line of the derivation of dispersion relations will be presented here for the simplest kinematical situations. A broader presentation of the analytic program including an extended class of analyticity properties for the general structure functions and (two-particle and multiparticle) collision amplitudes in QFT can also be found in this encyclopedia (*see* Scattering in Relativistic Quantum Field Theory: The Analytic Program). For brevity, we shall not treat here the derivation of dispersion relations in the framework of nonrelativistic potential theory. Concerning the latter, the interested reader can refer to the book by Nussenzweig (1972). A collection of old basic papers on field-theoretical dispersion relations can be found in the review book edited by Klein (1961). For a recent and well-documented review of the multiplicity of versions and applications of dispersion relations and their experimental checking, the reader can consult the article by Vernov (1996).

Typical Dispersion Relations

The possibility of defining the scattering function $\hat{F}(\omega^{(c)})$ in the full upper half-plane and of exploiting the corresponding boundary value F of $\hat{F}$ on the negative part as well as on the positive part of the real axis will depend on the framework of considered phenomena. For the moment, we do not consider the more general situations which also occur in particle physics and will be described later ("crossing domains" and "quasi-dispersion-relations").

In the simplest cases, the real and imaginary parts D and A of F are extended to negative values of the variable ω via additional symmetry relations resulting from appropriate "reality conditions." As a typical and basic example, there occurs the symmetry relation $\overline{\hat{F}(\omega^{(c)})} = \hat{F}(-\overline{\omega^{(c)}})$, (with $\omega^{(c)}$ and $-\overline{\omega^{(c)}}$ in the upper half-plane) and correspondingly $D(\omega) = D(-\omega)$, $A(\omega) = -A(-\omega)$ on the reals; we shall call (S) this symmetry relation.

The simplest case of dispersion relations is then obtained when D and A are linked by the reciprocal Hilbert transformations:

$$D(\omega) = \frac{1}{\pi} P \int_{-\infty}^{+\infty} A(\omega') \frac{1}{\omega' - \omega} \, d\omega' \qquad [1a]$$

$$A(\omega) = -\frac{1}{\pi} P \int_{-\infty}^{+\infty} D(\omega') \frac{1}{\omega' - \omega} \, d\omega' \qquad [1b]$$

where P denotes Cauchy's principal value, defined for any differentiable function $\varphi(x)$ (sufficiently regular at infinity) by

$$P \int_{-\infty}^{+\infty} \frac{\varphi(x)}{x} \, dx = \lim_{\varepsilon \to 0} \left[\int_{-\infty}^{-\varepsilon} \varphi(x) \frac{dx}{x} + \int_{\varepsilon}^{+\infty} \varphi(x) \frac{dx}{x} \right] \qquad [2]$$

As a matter of fact, the pair of equations [1a], [1b] is equivalent to the following relation for $F \doteq D + iA$:

$$F(\omega) = \frac{1}{2i\pi} \int_{-\infty}^{+\infty} F(\omega') \lim_{\varepsilon \to 0} \frac{1}{\omega' - \omega - i\varepsilon} \, d\omega' \qquad [3]$$

The latter is obtained as a limiting case of the Cauchy formula

$$\hat{F}(\omega^{(c)}) = \frac{1}{2i\pi} \int_{-\infty}^{+\infty} \frac{F(\omega')}{\omega' - \omega^{(c)}} \, d\omega' \qquad [4]$$

expressing the fact that $\hat{F}$ is holomorphic and sufficiently decreasing at infinity in the upper half-plane $\mathcal{T}_+$ of the complex variable $\omega^{(c)}$ and that $F(\omega)$ is the boundary value of $\hat{F}(\omega^{(c)})$ on all the reals.

Finally, one checks that in view of the symmetry relation (S), the Hilbert integral relations between D and A given above reduce to the following dispersion relations:

$$D(\omega) = \frac{2}{\pi} P \int_{0}^{+\infty} A(\omega') \frac{\omega'}{\omega'^2 - \omega^2} \, d\omega' \qquad [5a]$$

$$A(\omega) = -\frac{2\omega}{\pi} P \int_{0}^{+\infty} D(\omega') \frac{1}{\omega'^2 - \omega^2} \, d\omega' \qquad [5b]$$

Two Basic Examples

1. *The Kramers–Krönig relation in classical optics* It will be shown in the next part that the complex refractive index $n'(\omega) = n(\omega) + i\kappa(\omega)$ of a dielectric medium is the boundary value of a holomorphic

function $\hat{n}'(\omega^{(c)})$ in $\mathcal{I}_+$ satisfying the symmetry relation (S), and such that the integral $\int_{-\infty}^{\infty} |\hat{n}'(\omega + i\eta) - 1|^2 \, d\omega$ is uniformly bounded for all $\eta > 0$.

It follows that all the previous relations are satisfied by the function $\hat{F}(\omega^{(c)}) = \hat{n}'(\omega^{(c)}) - 1$. In particular, the real refractive index $n(\omega)$ and the "extinction coefficient" $\beta(\omega) \doteq 2\omega\kappa(\omega)/c$ (c being the velocity of light in the vacuum) are linked by the following Kramers–Krönig dispersion relation (corresponding to eqn [5a]):

$$n(\omega) - 1 = \frac{c}{2\pi} P \int \frac{\beta(\omega')}{\omega'^2 - \omega^2} \, d\omega' \qquad [6]$$

2. *Dispersion relation for the forward two-particle scattering amplitude in relativistic quantum physics* One considers the following collision phenomenon in particle physics. A particle Π_2 with mass m, called the target and sitting at rest in the laboratory, is collided by an identical particle Π_1 with relativistic energy ω larger than m ($= mc^2$; in high-energy physics, one usually chooses units such that $c = 1$). After the collision, the particle Π_1 is scattered in all possible directions, θ, of space, according to a certain quantum scattering amplitude $T_\theta(\omega)$, whose modulus is essentially the rate of probability for detecting Π_1 in the direction θ. The forward scattering amplitude $T_0(\omega)$ corresponds to the detection of Π_2 in the forward longitudinal direction with respect to its incidence direction towards the target. Let us also assume that the particles carry no charge of any kind, so that each particle coincides with its "antiparticle." In that case, $T_0(\omega)$ is shown to be the boundary value of a scattering function $\hat{T}_0(\omega^{(c)})$ enjoying the following properties:

1. it is a holomorphic function in $\mathcal{I}_+$ satisfying the symmetry relation (S);
2. its behavior at infinity in $\mathcal{I}_+$ is such that the integral

$$\int_{-\infty}^{\infty} \left| \frac{\hat{T}_0(\omega + i\eta)}{(\omega + i\eta)^2} \right|^2 d\omega$$

   is uniformly bounded for all $\eta > 0$; and
3. under more specific assumptions on the mass spectrum of the subjacent theory, the "absorptive part" $A(\omega) \doteq \operatorname{Im} T_0(\omega)$ vanishes for $|\omega| < m$.

Then by applying eqn [5a] to the function $D(\omega) \doteq \operatorname{Re}[(T_0(\omega) - T_0(0))/\omega^2]$ (regular at $\omega = 0$), one obtains the following dispersion relation:

$$\operatorname{Re} T_0(\omega) = T_0(0) + \frac{2\omega^2}{\pi} P \int_m^{+\infty} A(\omega') \frac{1}{\omega'(\omega'^2 - \omega^2)} \, d\omega' \qquad [7]$$

Remark In view of (3), the scattering function $\hat{T}_0(\omega^{(c)})$ admits an analytic continuation as an even function of $\omega^{(c)}$ (still called $\hat{T}_0$) in the cut-plane $C_m^{(\mathrm{cut})} \doteq C \setminus \{\omega \in R; |\omega| \geq m\}$. In fact, in view of (S) and (3), the boundary value T_0 of $\hat{T}_0$ satisfies the relation $T_0(\omega) = T_0(-\omega)$ in the real interval $\delta_m \doteq \{\omega \in R; -m < \omega < m\}$. Let us then introduce the function $\hat{T}_0^-(\omega^{(c)}) \doteq \hat{T}_0(-\omega^{(c)})$ as a holomorphic function of $\omega^{(c)}$ in $\mathcal{I}_-$: one sees that the boundary values of $\hat{T}_0$ and $\hat{T}_0^-$ from the respective domains $\mathcal{I}_+$ and $\mathcal{I}_-$ coincide on δ_m and therefore admit a common analytic continuation throughout this real interval (in view of "Painlevé's lemma" or "one-dimensional edge-of-the-wedge theorem"). One also notes that in view of (S) the extended function $\hat{T}_0$ satisfies the "reality condition" $\hat{T}_0(\omega^{(c)}) = \hat{T}_0(\overline{\omega^{(c)}})$ in $C_m^{(\mathrm{cut})}$. The fact that $\hat{T}_0$ is well defined as an even holomorphic function in the cut-plane $C_m^{(\mathrm{cut})}$ has been established in the general framework of QFT, as explained in the last part of this article.

Phenomena of Linear Response: Causality and Dispersion Relations in the Classical Domain

The subsequent axiomatic framework and results (due to J S Toll (1952, 1956)) concern any physical system which exhibits the following type of phenomena: whenever it receives some excitation signal, called the input and represented by a real-valued function of time $f_{\mathrm{in}}(t)$ with compact support, the system emits a response signal, called the output and represented by a corresponding real-valued function $f_{\mathrm{out}}(t)$, in such a way that the following postulates are satisfied:

(P1) *Linearity.* To every linear combination of inputs $a_1 f_{\mathrm{in},1} + a_2 f_{\mathrm{in},2}$, there corresponds the output $a_1 f_{\mathrm{out},1} + a_2 f_{\mathrm{out},2}$.

(P2) *Reproductibility or time-translation invariance.* Let τ be a time-translation parameter taking arbitrary real values; to every "time-translated input" $f_{\mathrm{in}}^{(\tau)}(t) \doteq f_{\mathrm{in}}(t - \tau)$, there corresponds the output $f_{\mathrm{out}}^{(\tau)}(t) \doteq f_{\mathrm{out}}(t - \tau)$.

(P3) *Causality.* The effect cannot precede the cause, namely if t_{in} and t_{out} denote respectively the lower bounds of the supports of $f_{\mathrm{in}}(t)$ and $f_{\mathrm{out}}(t)$, then there always holds the inequality $t_{\mathrm{in}} \leq t_{\mathrm{out}}$.

(P4) *Continuity of the response.* There exists some continuity inequality which expresses the fact that a certain norm of the output is majorized by a corresponding norm of the input. The case of an L_2-norm inequality of the

form $|f_{out}| \leq |f_{in}|$ is particularly significant: when the norm $|f| \doteq [\int |f(t)|^2 \, dt]^{1/2}$ is interpretable as an energy (for the output as well as for the input), it acquires the meaning of a "dissipation" property of the system.

The postulate of linear dependence (P1) of f_{out} with respect to f_{in} is obviously satisfied if the response is described by any general kernel $K(t, t')$ such that the following formula makes sense:

$$f_{out}(t) = \int_{-\infty}^{+\infty} K(t, t') f_{in}(t') dt' \qquad [8]$$

Conversely, the existence of a distribution kernel K can be established rigorously under the continuity assumption postulated in (P4) by using the Schwartz nuclear theorem. In full generality (see our comment in the next paragraph), the kernel $K(t, t')$ appears to be a tempered distribution in the pair of variables (t, t') and the previous integral formula holds in the sense of distributions, which means that both sides of eqn [8] must be considered as tempered distributions (in t) acting on any smooth test-function $g(t)$ in the Schwartz space $\mathcal{S}$. (Note, for instance, that the trivial linear application $f_{out} = f_{in}$ is represented by the kernel $K(t, t') = \delta(t - t')$).

From the reproductibility postulate (P2), it follows that the distribution K can be identified with a distribution of the single variable $\tau = t - t'$, namely $K(t, t') \doteq R(t - t')$. Moreover, the real-valuedness condition imposed to the pairs (f_{in}, f_{out}) entails that R is real. Finally, the causality postulate (P3) implies that the support of the distribution R is contained in the positive real axis, so that one can write, in the sense of distributions,

$$f_{out}(t) = \int_{-\infty}^{t} R(t - t') f_{in}(t') dt' \qquad [9]$$

The convolution kernel $R(t - t')$ is typically what one calls in physics a "retarded kernel."

If we now introduce the frequency variable ω, which is the conjugate of the time variable t, by the Fourier transformation

$$\tilde{f}(\omega) = \int_{-\infty}^{+\infty} f(t) \, e^{i\omega t} \, dt$$

we see that the convolution equation [9] is equivalent to the following one:

$$\tilde{f}_{out}(\omega) = \tilde{R}(\omega) \tilde{f}_{in}(\omega) \qquad [10]$$

In the latter, the Fourier transform $\tilde{R}(\omega)$ of R is a tempered distribution, which is the boundary value from the upper half-plane $\mathcal{I}_+$ of a holomorphic function $\tilde{R}^{(c)}(\omega^{(c)})$, called the Fourier–Laplace transform of R. $\tilde{R}^{(c)}$ is defined for all $\omega^{(c)} = \omega + i\eta$, with $\eta > 0$, by the following formula in which the exponential is a good test-function for the distribution R (since exponentially decreasing for $t \to +\infty$):

$$\tilde{R}^{(c)}(\omega^{(c)}) = \int_0^{+\infty} R(t) e^{i\omega^{(c)} t} \, dt \qquad [11]$$

More precisely, the tempered-distribution character of R is strictly equivalent to the fact that $\tilde{R}^{(c)}$ is of moderate growth both at infinity and near the reals in $\mathcal{I}_+$, namely that it satisfies a majorization of the following form for some real positive numbers p and q:

$$|\tilde{R}^{(c)}(\omega + i\eta)| \leq C \frac{(1 + |\omega|^2 + \eta^2)^q}{\eta^p} \qquad [12]$$

We thus conclude from eqn [10] that each phenomenon of linear response is represented very simply in the frequency variable by the multiplicative operator $\tilde{R}(\omega)$, whose analytic continuation $\tilde{R}^{(c)}(\omega^{(c)})$ is called the (causal) response function.

A Typical Illustration: The Damped Harmonic Oscillator

We consider the motion $x = x(t)$ of a damped harmonic oscillator of mass m submitted to an external force $F(t)$. The force is the input ($f_{in} = F$) and the resulting motion is the output, namely $f_{out}(t) = x(t)$. All the previous general postulates (P1)–(P4) are then satisfied, but this particular model is, of course, governed by its dynamical equation

$$x''(t) + 2\gamma x'(t) + \omega_0^2 x(t) = \frac{F(t)}{m} \qquad [13]$$

where ω_0 is the eigenfrequency of the oscillator and γ is the damping constant ($\gamma > 0$). The relevant solution of this second-order differential equation with constant coefficients is readily obtained in terms of the Fourier transforms $\tilde{x}(\omega)$ of $x(t)$ and $\tilde{F}(\omega)$ of $F(t)$. One can in fact replace eqn [13] by the equivalent equation

$$(-\omega^2 - 2i\gamma\omega + \omega_0^2) \tilde{x}(\omega) = \frac{\tilde{F}(\omega)}{m} \qquad [14]$$

whose solution is of the form [10], namely $\tilde{x}(\omega) = \tilde{R}(\omega) \tilde{F}(\omega)$, with

$$\tilde{R}(\omega) = -\frac{\tilde{F}(\omega)}{m(\omega^2 + 2i\gamma\omega - \omega_0^2)} = -\frac{\tilde{F}(\omega)}{m(\omega - \omega_1)(\omega - \omega_2)} \qquad [15a]$$

$$\omega_{1,2} = \pm(\omega_0^2 - \gamma^2)^{1/2} - i\gamma \qquad [15b]$$

It is clear that the rational function defined by eqns [15] admits an analytic continuation in the full complex plane of $\omega^{(c)}$ minus the pair of simple poles (ω_1, ω_2) *which lie in the lower half-plane.* In particular, it is holomorphic (and decreasing at infinity) in $\mathcal{I}_+$, as expected from the previous general result. Moreover, this example suggests that for any particular phenomenon of linear response, the details of the dynamics are encoded in the singularities of the holomorphic scattering function $\tilde{R}^{(c)}(\omega^{(c)})$, which all lie in the lower half-plane. The validity of a dispersion relation only expresses the analyticity (and decrease at infinity) of that function in the upper half-plane, which is model independent.

Remark The same mathematical analysis applies to any electric oscillatory circuit, in which the capacitance, inductance, and resistance are involved in place of the parameters m, ω_0 and γ: f_{in} and f_{out} correspond respectively to an external electric potential and to the current induced in the circuit; the response function is the admittance of the circuit.

Application to the Kramers–Krönig Relation

The background of the Kramers–Krönig relation [6], namely the analyticity and boundedness properties of the complex refractive index $\hat{n}'(\omega^{(c)})$ in $\mathcal{I}_+$, is provided by the previous axiomatic framework. However, it is not the quantity $\hat{n}'(\omega^{(c)})$ itself but appropriate functions of the latter which play the role of causal response functions; two phenomena can in fact be exhibited, which both contribute to proving the relevant properties of $\hat{n}'(\omega^{(c)})$.

1. *Propagation of light in a dielectric slab with thickness* δ. One considers the wave front $f_{\text{in}}(t)$ of an incoming wave normally incident upon the slab, with Fourier decomposition

$$f_{\text{in}}(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \tilde{f}(\omega)\, e^{-i\omega t}\, d\omega \qquad [16]$$

After having traveled through the medium, it gives rise to an outgoing wave $f_{\text{out}}(t)$ on the exit face of the slab, whose Fourier decomposition can be written as follows (provided the thickness δ of the slab is very small):

$$f_{\text{out}}(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \tilde{f}(\omega)\, e^{-i\omega(t - n'(\omega)/c\delta)}\, d\omega \qquad [17]$$

In the latter, the real part of $n'(\omega)/c$ is the inverse of the light velocity in the medium, while its imaginary part takes into account the exponential damping of the wave. The output f_{out} thus appears as a causal linear response with respect to f_{in} (since f_{out} "starts after" f_{in}). According to the general formula [10], the corresponding response function $\tilde{R}_\delta^{(c)}$ can be *directly computed from eqns [16] and [17], which* yields:

$$\tilde{R}_\delta^{(c)}(\omega^{(c)}) = e^{i\omega^{(c)}\hat{n}'(\omega^{(c)})\delta/c} \qquad [18]$$

In view of the previous axiomatic analysis, $\tilde{R}_\delta^{(c)}$ has to be holomorphic and of moderate growth in $\mathcal{I}_+$, and since this holds for all δ's sufficiently small, it can be shown that the function $\hat{n}'(\omega^{(c)})$ itself is holomorphic and of moderate growth in $\mathcal{I}_+$ (no logarithmic singularity can be produced).

2. *Polarization of the medium produced by an electric field.* The dielectric polarization signal $P(t)$ produced at a point of a medium by an external electric field $E(t)$ is also a phenomenon of linear response which obeys the postulates (P1)–(P4); the corresponding formula [10] reads

$$\tilde{P}(\omega) = \chi'(\omega)\tilde{E}(\omega) \qquad [19a]$$

where χ' is the complex dielectric susceptibility of the medium, which is related to n' by Maxwell's relation

$$\tilde{\chi}'(\omega) = \frac{[n'^2(\omega) - 1]}{4\pi} \qquad [19b]$$

One thus recovers the fact that χ' admits an analytic continuation in $\mathcal{I}_+$; one can also show by a physical argument that $\tilde{\chi}'(\omega)$, and thereby $n'(\omega) - 1$, tends to zero as a constant divided by ω^2 when ω tends to infinity. This behavior at infinity extends to $\hat{n}'(\omega^{(c)}) - 1$ *in* $\mathcal{I}_+$ *in view of the Phragmen–Lindelöf* theorem, since $\hat{n}'$ is known (from (1)) to be of moderate growth. This justifies the analytic background of Kramers–Krönig's relation.

From Relativistic QFT to the Dispersion Relations of Particle Physics: Historical Considerations and General Survey

In the quantum domain, the derivation of dispersion relations for the two-particle scattering (or collision) amplitudes of particle physics has represented, since 1956 and throughout the 1960s, an important conceptual progress for the theoretical treatment of that branch of physics. These phenomena are described in a quantum-theoretical framework in which the basic kinematical variables are the energies and momenta of the particles involved. These variables play the role of the frequency of light in the optical scattering phenomena. Moreover, since large energies and momenta are involved, which allow the occurrence of particle creation

according to the conservation laws of special relativity, it is necessary to use a relativistic quantum-mechanical framework. Around 1950, the success of the quantum electrodynamics formalism for computing the electron–photon, electron–electron, and electron–positron scattering amplitudes revealed the importance of the concept of relativistic quantum field for the understanding of particle physics. However, the methods of perturbation theory, which had ensured the success of quantum electrodynamics in view of the small value of the coupling parameter of that theory (namely the electric charge of the electron), were at that time inapplicable to the strong nuclear interaction phenomena of high-energy physics. This failure motivated an important school of mathematical physicists for working out a model-independent axiomatic approach of relativistic QFT (e.g., Lehmann, Symanzik, Zimmermann (1954), Wightman (1956), and Bogoliubov (1960); *see* Axiomatic Quantum Field Theory). Their main purpose was to provide a conceptually satisfactory treatment of relativistic quantum collisions, at least for the case of massive particles. Among various postulates expressing the invariance of the theory under the Poincaré group in an appropriate quantum-mechanical Hilbert-space framework, the approach basically includes a certain formulation of the principle of causality, called microcausality or local commutativity. This axiomatic approach of QFT was followed by a conceptually important variant, namely the algebraic approach to QFT (Haag, Kastler, Araki 1960), whose most important developments are presented in the book by Haag (1992) (*see* Algebraic Approach to Quantum Field Theory). From the historical viewpoint, and in view of the analyticity properties that they also generate, one can say that all these (closely related) approaches parallel the axiomatic approach of linear response phenomena with, of course, a much higher degree of complexity. In particular, the characterization of scattering (or collision) amplitudes in terms of appropriate structure functions of the basic quantum fields of the theory is a nontrivial preliminary step which was taken at an early stage of the theory under the name of "asymptotic theory and reduction formulae" (Lehmann, Symanzik, Zimmermann 1954–57, Haag–Ruelle 1962, Hepp 1965). There again, in the field-theoretical axiomatic framework, causality generates analyticity through Fourier–Laplace transformation, but several complex variables now play the role which was played by the complex frequency in the axiomatics of linear response phenomena: they are obtained by complexifying the relativistic energy–momentum variables of the (Fourier transforms of the) quantum fields involved in the high-energy collision processes. In fact, the holomorphic functions which play the role of the causal response function $\tilde{R}(\omega)$ are the QFT structure functions or "Green functions in energy–momentum space." The study of all possible analyticity properties of these functions resulting from the QFT axiomatic framework is called the analytic program (*see* Scattering in Relativistic Quantum Field Theory: The Analytic Program). The primary basic scope of the latter concerns the derivation of analyticity properties for the scattering functions of two-particle collision processes, which appears to be a genuine challenge for the following reason. The basic Einstein relation $E = mc^2$, which applies to all the incoming and outgoing particles of the collisions, operates as a geometrical constraint on the corresponding physical energy–momentum vectors: according to the Minkowskian geometry, the latter have to belong to mass hyperboloids, which define the so-called "mass shell" of the collision considered. It is on the corresponding complexified mass-shell manifold that the scattering functions are required to be defined as holomorphic functions. In the analytic program of QFT, the derivation of such analyticity domains and of corresponding dispersion relations in the complex plane of the squared total energy variable, s, of each given collision process then relies on techniques of complex geometry in several variables. As a matter of fact, the scattering amplitude is a function (or distribution) of two variables $F(s, t)$, where t is a second important variable, called the squared momentum transfer, which plays the role of a fixed parameter for the derivation of dispersion relations in the variable s. The value $t = 0$ corresponds to the special kinematical situation which has been described above (for the case of equal-mass particles Π_1 and Π_2) under the name of forward scattering and the variable s is a simple affine function of the energy ω of the colliding particle Π_1 in the laboratory Lorentz frame, (namely $s = 2m^2 + 2m\omega$ in the equal-mass case). It is for the corresponding scattering amplitude $T_0(\omega) \doteq F_0(s) \doteq F(s, t)_{|t=0}$ that a dispersion relation such as eqn [7] can be derived, although this derivation is far from being as simple as for the phenomena of linear response in classical physics: even in that simplest case, it already necessitates the use of analytic completion techniques in several complex variables. The first proof of this dispersion relation was performed by K Symanzik in 1956. In the case of general kinematical situations of measurements, the direction of observation of the scattered particle includes a nonzero angle with the incidence direction, which always corresponds to a negative value of t. The derivation of dispersion relations at fixed $t = t_0 < 0$, namely for the scattering amplitude

$F_{t_0}(s) \doteq F(s,t)_{|t=t_0}$ requires further arguments of complex geometry, and it is submitted to subtle limitations of the form $t_1 < t_0 \leq 0$, where t_1 depends on the mass spectrum of the particles involved in the theory. The first rigorous proof of dispersion relations at $t < 0$ was performed by N N Bogoliubov in 1960.

Three conceptually important features of the dispersion relations in particle physics deserve to be pointed out.

1. In comparison with the dispersion relations of classical optics, a feature which appears to be new is the so-called "crossing property," which is characteristic of high-energy physics since it relies basically on the relativistic kinematics. According to that property, the boundary values of the analytic scattering function $\hat{F}_t(s)$ at positive and negative values of s from the respective half-planes $\mathrm{Im}\, s > 0$ and $\mathrm{Im}\, s < 0$ are interpreted, respectively, as the scattering amplitudes of two physically different collision processes, which are deduced from each other by replacing the incident particle by the corresponding antiparticle; one also says that "these two collision processes are related by crossing." A typical example is provided by the proton–proton and proton–antiproton collisions, whose scattering amplitudes are therefore mutually related by the property of analytic continuation. This type of relationship between the values of the scattering function at positive and negative values of s generalizes in a nontrivial way the symmetry relation (S) satisfied by the forward scattering function $\hat{T}_0(\omega^{(c)})$ when each particle coincides with its antiparticle (see the second basic example above). No nontrivial crossing property holds in that special case and the fact that $\hat{T}_0$ is an even function of $\omega^{(c)}$ precisely expresses the identity of the two-collision processes related by crossing. In the general case, for $t = 0$ as well as for $t = t_0 < 0$ for any value of t_0, the analyticity domain that one obtains for the scattering function is not the full cut-plane of s: in its general form, a "crossing domain" may exclude some bounded region B_{t_0} from the cut-plane, but it always contains an infinite region which is the exterior of a circle minus cuts along the two infinite parts of the real s-axis (Bros, Epstein, Glaser 1965): these cuts are along the physical regions of the two collision processes related by crossing. In that general case, the scattering function $\hat{F}_t(s)$ still satisfies what can be called a quasi-dispersion-relation, in which the right-hand side contains an additional Cauchy integral, taken along the boundary of B_t.

2. A second important feature concerns the behavior at large values of s of the scattering functions $\hat{F}_t(s)$ in their analyticity domain. As indicated in the presentation of the second basic example, a "precise-increase" property was expected to be satisfied by the forward scattering amplitude $T_0(\omega)$ for ω (or s) tending to infinity. This "precise-increase" property implied the necessity of writing the corresponding dispersion relation [7] for the function $(T_0(\omega) - T_0(0))/\omega^2$: this is what one calls a "dispersion relation with a subtraction." As a matter of fact, the existence of such restrictive bounds on the total cross sections at high energies had been discovered in 1961 by M Froissart: his derivation relied basically on the use of the unitarity of the scattering operator (expressing the quantum principle of conservation of probabilities), but also on a strong analyticity postulate for the scattering function not implied by the general field-theoretical approach (namely the Mandelstam domain of "double dispersion relations"). In the general framework of QFT, Froissart-type bounds appeared to be closely linked to a further nontrivial extension of the range of "admissible" values of t for which $\hat{F}_t(s)$ can be analytically continued in a cut-plane or crossing domain. In fact, the extension of this range to positive (i.e., "unphysical") and even complex values of t, and as a second step the proof of Froissart-type bounds in $s(\log s)^2$ for $F_t(s)$ at all these admissible values of t, were performed in 1966 by A Martin. They rely on a subtle conspiracy of the analyticity properties deduced from the QFT axiomatic framework and of positivity and unitarity properties expressing the basic Hilbertian structure of the quantum collision theory. The consequence of these bounds on the exact form of the dispersion relations is that, as in formula [7] of the case $t = 0$, it is justified to write a (the so-called "subtracted") dispersion relation for $(F_t(s) - F_t(0) - sF_t'(0))/s^2$: for the general case when the crossing property replaces the symmetry (S), such a dispersion relation involves two subtractions (since $F_t'(0) \neq 0$). Detailed information concerning the interplay of analyticity and unitarity on the mass shell and the derivation of refined forms of dispersion relations and various boundedness properties for the scattering functions are given in the book by Martin (1969).

3. Constraints imposed by dispersion relations and experimental checks. The conceptual importance of dispersion relations incorporating the above features (1) and (2) is displayed by such spectacular application as the relationship between the high-energy behaviors of proton–proton and proton–antiproton cross sections. Even though the closest forms of relationship between these cross sections (e.g., the existence of equal high-energy limits) necessitate for their proof some extra assumption concerning, for instance, the behavior

of the ratio between the dispersive and absorptive parts of the forward scattering amplitude, one can speak of an actual model-independent implication of general QFT that imposes nontrivial constraints on phenomena. Otherwise stated, checking experimentally the previous type of relationship up to the limits of high energies imposed by the present technology of accelerators constitutes an indirect, but important test of the validity of the general principles of QFT.

As a matter of fact, it has also appeared frequently in the literature of high-energy physics during the last 40 years that the Froissart bound by itself was considered as a key criterion to be satisfied by any sensible phenomenological model in particle physics. As already stated above, the Froissart bound is one of the deepest consequences of the analytic program of general QFT, since its derivation also incorporates in the most subtle way the quantum principle of probability conservation. Would it be only for the previous basic results, the derivation of dispersion relations (and, more generally, the results of the analytic program) in QFT appear as an important conceptual bridge between a fundamental theoretical framework of relativistic quantum physics and the phenomenology of high-energy particle physics.

Basic Concepts and Main Steps in the QFT Derivation of Dispersion Relations

The rest of this article outlines the derivation of the analytic background of dispersion relations for the forward scattering amplitudes in the framework of axiomatic QFT. After a brief introduction on relativistic scattering processes and the problematics of causality in particle physics, it gives an account of the Wightman axioms and the simplest reduction formula which relates the forward scattering amplitude to a retarded product of the field operators. Then it describes how the latter can be used for justifying a certain type of analyticity domain for the forward scattering functions, namely a crossing domain or in the best cases a cut-plane in the squared energy variable s. This is the basic result that allows one to write dispersion relations (or quasi-dispersion-relations) at $t=0$; the exact form of the latter, including at most two subtractions, relies on the use of Hilbertian positivity and of the unitarity of the scattering operator.

Relativistic Quantum Scattering as a Phenomenon of Linear Response

Collisions of quantum particles may be seen as phenomena of linear response, but in a way which differs greatly from what has been previously described.

Particles in Minkowskian geometry Each state of a relativistic classical particle with mass m is characterized by its energy–momentum vector or 4-momentum $p=(p_0,\boldsymbol{p})$ satisfying the mass-shell condition $p^2 \doteq p_0^2-\boldsymbol{p}^2=m^2$ (in units such that $c=1$). In view of the condition of positivity of the energy $p_0>0$ the "physical mass shell" thus coincides with the positive sheet H_m^+ of the mass hyperboloid H_m with equation $p^2=m^2$.

The set of all energy–momentum configurations characterizing the collisions of two relativistic classical particles with initial (resp. final) 4-momenta p_1,p_2 (resp. p_1',p_2') is the mass-shell manifold $\mathcal{M}$ defined by the conditions

$$p_i^2=m^2,\quad p_i'^2=m^2,\quad p_{i,0}>0,\quad p_{i,0}'>0,\quad i=1,2$$
$$p_1+p_2=p_1'+p_2'$$

where the latter equation expresses the relativistic law of total energy–momentum conservation. $\mathcal{M}$ is an eight-dimensional manifold, invariant under the (six-dimensional) Lorentz group: the orbits of this group that constitute a foliation of $\mathcal{M}$ are parametrized by two variables, namely the squared total energy $s=(p_1+p_2)^2=(p_1'+p_2')^2$ and the squared momentum transfer $t=(p_1-p_1')^2=(p_2-p_2')^2$ (or $u=(p_1-p_2')^2=4m^2-s-t$). In these variables, called the Mandelstam variables, the "physical region" Φ of the collision is represented by the set of pairs (s,t) (or triplets (s,t,u) with $s+t+u=4m^2$) such that $t\le 0, u\le 0$, and therefore $s\ge 4m^2$.

Correspondingly, each state of a relativistic quantum particle with mass m is characterized by a wave packet $\hat{f}(p)$ on H_m^+, which is an element of unit norm of $L_2(H_m^+;\mu_m(p))$, with $\mu_m(p)=d\boldsymbol{p}/(\boldsymbol{p}^2+m^2)^{1/2}$. In Minkowskian spacetime with coordinates $x=(x_0,\boldsymbol{x})$, any such state is represented by a wave function $f(x)$ whose Fourier transform is the tempered distribution (with support in H_m^+) $\hat{f}(p)\times\delta(p^2-m^2)$: $f(x)$ is a positive-energy solution of the Klein–Gordon equation $(\partial^2/\partial x_0^2-\Delta_x+m^2)f(x)=0$. A free two-particle state is a symmetric wave packet $\hat{f}(p_1,p_2)$ on $H_m^+\times H_m^+$ in the Hilbert space $L_2(H_m^+\times H_m^+;\mu_m\otimes\mu_m)$.

Scattering kernels as response kernels: distribution character While the input to be considered is a free wave packet $\hat{f}_{\text{in}}(p_1,p_2)$ on $H_m^+\times H_m^+$, representing the preparation of an initial two-particle state, the output corresponds to the detection of a final two-particle state also characterized by a wave packet $\hat{g}_{\text{out}}(p_1',p_2')$ on $H_m^+\times H_m^+$. In quantum mechanics, linearity is linked to the "superposition principle" of states,

which allows one to state that collisions are described by a certain bilinear form $(\hat{f}_{\text{in}}, \hat{g}_{\text{out}}) \rightarrow S(\hat{f}_{\text{in}}, \hat{g}_{\text{out}})$, called the "scattering matrix." This bilinear form is bicontinuous with respect to the Hilbertian norms of the wave packets, and it then results from the Schwartz nuclear theorem that it is represented by a distribution kernel $S(p_1, p_2; p'_1, p'_2)$, namely a tempered distribution with support contained in $\mathcal{M}$, in such a way that (formally)

$$S(\hat{f}_{\text{in}}, \hat{g}_{\text{out}}) = \int \hat{f}_{\text{in}}(p_1, p_2)\overline{\hat{g}_{\text{out}}}(p'_1, p'_2) S(p_1, p_2; p'_1, p'_2) \times \mu_m(p_1)\mu_m(p_2)\mu_m(p'_1)\mu_m(p'_2) \qquad [20]$$

If there were no interaction, $S(\hat{f}_{\text{in}}, \hat{g}_{\text{out}})$ would reduce to the Hilbertian scalar product $<\hat{g}_{\text{out}}, \hat{f}_{\text{in}}>$ in $L_2(H_m^+ \times H_m^+; \mu_m \otimes \mu_m)$ and the corresponding kernel S would be the identity kernel

$$I(p_1, p_2; p'_1, p'_2) = \frac{1}{2}[\delta(p_1 - p'_1)\delta(p_2 - p'_2) + \delta(p_1 - p'_2)\delta(p_2 - p'_1)]$$

In the general case, the interaction is therefore described by the scattering kernel $T(p_1, p_2; p'_1, p'_2) \doteq S(p_1, p_2; p'_1, p'_2) - I(p_1, p_2; p'_1, p'_2)$. The action of T as a bilinear form (defined in the same way as the action of S in eqn [20]) may be seen as the quantum analog of the classical response formula [10]. Note, however, the difference in the mathematical treatment of the output: instead of being considered as the direct response $(\tilde{f}_{\text{out}})$ to the input, it is now explored by Hilbertian duality in terms of detection wave packets $\hat{g}_{\text{out}}$, in conformity with the principles of quantum theory. Finally, in view of the invariance of the collision process under the Lorentz group, the scattering kernel T is constant along the orbits of this group in $\mathcal{M}$ and it then defines a distribution $F(s, t) \doteq T(p_1, p_2; p'_1, p'_2)$ with support in the physical region Φ: this is what is called the scattering amplitude.

What becomes of causality? One can show that the positive-energy solutions of the Klein–Gordon equation cannot vanish in any open set of Minkowski spacetime; they necessarily spread out in the whole spacetime. This makes it impossible to formulate a causality condition comparable to eqn [9] in terms of the spacetime wave functions f_{in} and g_{out} corresponding to the input and output wave packets $\hat{f}_{\text{in}}, \hat{g}_{\text{out}}$. In this connection, it is, however, appropriate to note that (after various attempts of "weak causality conditions") a certain condition called "macrocausality" (Iagolnitzer and Stapp 1969; see the book by Iagolnitzer (1992)) has been shown to be equivalent to some local properties of analyticity of the scattering kernel T; but it is not our purpose to develop that point here for two reasons: (1) the interpretation of that condition is rather involved, because it integrates a very weak form of causality together with the spatial short-range character of the strong nuclear interactions between the elementary particles; (2) the domains of analyticity obtained are by far too small with respect to those necessary for writing dispersion relations. The reason for this failure is that the scattering kernel only represents an asymptotic quantum observable, in the sense that it is intended to describe observations far apart from the extremely small spacetime region where the particles strongly interact, namely in regions where this interaction is asymptotically small. Although well adapted to what is actually observed in the detection experiments, the concept of scattering kernel is not sufficient for describing the fundamental interactions of physics: it must be enriched by other theoretical concepts which might explicitly take into account the microscopic interactions in spacetime. This motivates the introduction of quantum fields as basic quantities in particle physics.

Relativistic Quantum Fields: Microcausality and the Retarded and Advanced Kernels; Analyticity in Complex Energy–Momentum Space

By an idealization of the concept of quantum electromagnetic field and a generalization to all types of microscopic interactions of matter, one considers that all the phenomena involving such interactions can be described by fields $\Phi_i(x)$, whose amplitude can, in principle, be measured in arbitrarily small regions of Minkowski spacetime. In the quantum framework, one is thus led to the notion of local observable $\mathcal{O}$ (emphasized as a basic concept in the axiomatic approach of Araki, Haag, and Kastler). In the Wightman field-theoretical framework, a local observable corresponds to the measuring process of a ponderated average of a field $\Phi_i(x)$ of the form $\mathcal{O} \doteq \Phi_i[f] = \int \Phi_i(x) f(x)\, dx$. In the latter, $f(x)$ denotes a smooth real-valued test-function with (arbitrary) compact support K in spacetime; the observable $\mathcal{O}$ is then said to be localized in K. Each observable $\mathcal{O} = \Phi_i(f)$ has to be a self-adjoint (unbounded) operator acting in (a dense domain of) the Hilbert space $\mathcal{H}$ generated by all the states of the system of fundamental fields $\{\Phi_i\}$; therefore, the correct mathematical concept of relativistic quantum field $\Phi(x)$ is an "operator-valued tempered distribution on Minkowski spacetime." Here the additional "temperateness assumption" is a convenient technical assumption which in particular allows the passage to the energy–momentum space by making use of the Fourier transformation.

In this QFT framework, it is natural to express a certain form of causality by assuming that two observables $\Phi(f)$ and $\Phi(f')$ commute if the supports of f and f' are spacelike-separated regions in spacetime, which means that no signal with velocity smaller or equal to the velocity of light can propagate from either one of these regions to the other. This expresses the idea that these two observables should be independent, that is, "compatible as quantum observables." This postulate is equivalent to the following condition, called microcausality or local commutativity, and understood in the sense of operator-valued tempered distributions:

$$[\Phi(x_1), \Phi(x_1')] = 0, \quad \text{for } (x_1 - x_1')^2 < 0 \qquad [21]$$

where $(x_1 - x_1')^2$ is the squared Minkowskian pseudonorm of $x = x_1 - x_1' = (x_0, \boldsymbol{x})$, namely $x^2 = x_0^2 - \boldsymbol{x}^2$. It follows that for every admissible pair of states Ψ, Ψ' in $\mathcal{H}$, the tempered distribution

$$C_{\Psi,\Psi'}(x_1, x_1') \doteq <\Psi, [\Phi(x_1), \Phi(x_1')]\Psi'> \qquad [22]$$

has its support contained in the union of the sets $\mathcal{V}^+ : x_1 - x_1' \in \overline{V^+}$ and $\mathcal{V}^- : x_1 - x_1' \in \overline{V^-}$, where $\overline{V^+}$ and $\overline{V^-}$ are, respectively, the closures of the forward and backward cones $V^+ \doteq \{x = (x_0, \boldsymbol{x}); x_0 > |\boldsymbol{x}|\}$, $V^- \doteq -V^+$ in Minkowski spacetime. It is always possible to decompose the previous distribution as

$$C_{\Psi,\Psi'}(x_1, x_1') = R_{\Psi,\Psi'}(x_1, x_1') - A_{\Psi,\Psi'}(x_1, x_1') \qquad [23]$$

in such a way that the supports of the distributions $R_{\Psi,\Psi'}(x_1, x_1')$ and $A_{\Psi,\Psi'}(x_1, x_1')$ belong, respectively, to $\mathcal{V}^+$ and $\mathcal{V}^-$. $R_{\Psi,\Psi'}$ and $A_{\Psi,\Psi'}$ are called, respectively, retarded and advanced kernels and they are often formally expressed (for convenience) as follows:

$$R_{\Psi,\Psi'}(x_1, x_1') = \theta(x_{1,0} - x_{1,0}')C_{\Psi,\Psi'}(x_1, x_1')$$
$$A_{\Psi,\Psi'}(x_1, x_1') = -\theta(x_{1,0} - x_{1,0}')C_{\Psi,\Psi'}(x_1, x_1')$$

in terms of the Heaviside step function $\theta(t)$ of the time-coordinate difference $t = x_{1,0} - x_{1,0}'$. For every pair (Ψ, Ψ'), $R_{\Psi,\Psi'}(x_1, x_1')$ appears as a relativistic generalization of the retarded kernel $R(t - t')$ of eqn [10]: its support property in spacetime, similar to the support property of R in time, expresses a relativistic form of causality, or "Einstein causality."

There exists a several-variable extension of the theory of Fourier–Laplace transforms of tempered distributions which is based on a formula similar to eqn [11]. We introduce the vector variables $X = (x_1 + x_1')/2, x = x_1 - x_1'$ and a complex 4-momentum $k = p + iq = (k_0, \boldsymbol{k})$ as the conjugate vector variable of x with respect to the Minkowskian scalar product $k \doteq k_0 x_0 - \boldsymbol{k} \cdot \boldsymbol{x}$, and we define

$$\tilde{R}^{(c)}_{\Psi,\Psi'}(k, X) = \int_{\overline{V}^+} R_{\psi,\Psi'}\left(X + \frac{x}{2}, X - \frac{x}{2}\right) e^{ik \cdot x}\, dx \qquad [24]$$

Since $q \cdot x > 0$ for all pairs (q, x) such that $q \in V^+$, $x \in \overline{V}^+$, it follows that $\tilde{R}^{(c)}_{\Psi,\Psi'}(k, X)$ is holomorphic with respect to k in the domain $\mathcal{T}^+$ containing all $k = p + iq$ such that q belongs to V^+. Moreover, in the limit $q \to 0$ this holomorphic function tends (in the sense of distributions) to the Fourier transform $\tilde{R}_{\Psi,\Psi'}(p, X)$ of $R_{\Psi,\Psi'}(X + x/2, X - x/2)$ with respect to x. The domain $\mathcal{T}^+$, which is called the "forward tube," is the analog of the domain $\mathcal{T}_+$ of the ω-plane; bounds of moderate type comparable to those of [12] apply to the holomorphic function $\tilde{R}^{(c)}_{\Psi,\Psi'}$ in $\mathcal{T}^+$. Similarly, the advanced kernel $A_{\Psi,\Psi'}(X + x/2, X - x/2)$ admits a Fourier–Laplace transform $\tilde{A}^{(c)}_{\Psi,\Psi'}(k, X)$, which is holomorphic and of moderate growth in the "backward tube" $\mathcal{T}^-$ containing all $k = p + iq$ such that q belongs to V^-. In view of [23], the Fourier transform $\tilde{C}_{\Psi,\Psi'}(p, X)$ of $C_{\Psi,\Psi'}(X + x/2, X - x/2)$ then appears as the difference between the boundary values of $\tilde{R}^{(c)}_{\Psi_1,\Psi_2}$ and $\tilde{A}^{(c)}_{\Psi_1,\Psi_2}$ on the reals (from the respective domains $\mathcal{T}^+$ and $\mathcal{T}^-$).

The Field-Theoretical Axiomatic Framework and the Passage from the Structure Functions of QFT to the Scattering Kernels (Case of Forward Scattering)

The postulates (Wightman axioms) Apart from the causality postulate, which we have already presented above in view of its distinguished role for generating analyticity properties in complex energy–momentum space, the field-theoretical axiomatic approach to collision theory is based on the following postulates (for all the fundamental developments of axiomatic field theory, the interested reader may consult the books by Streater and Wightman (1980) and by Jost (1965); *see* Axiomatic Quantum Field Theory).

1. There exists a unitary representation $g \to \mathcal{U}(g)$ of the Poincaré group G in the Hilbert space of states $\mathcal{H}$; in this representation, the abelian subgroup of translations of space and time has a Lie algebra whose generators are interpreted as the four self-adjoint (commuting) operators P_μ of total energy–momentum of the system.
2. The quantum field operators $\Phi(x)$ transform covariantly under that representation; in the simplest case of scalar fields (considered here), $\Phi(gx) = U(g)\Phi(x)U(g^{-1})$.
3. There exists a unique state Ω, called the vacuum, such that the action of all polynomials of field operators on Ω generates a dense subset of $\mathcal{H}$;

moreover, Ω is assumed to be invariant under the representation $\mathcal{U}$ of G, and thereby such that $P_\mu \Omega = 0$.

4. Spectral condition or positivity of energy in all physical states. The joint spectrum Σ of the operators P_μ is contained in the closed forward cone $\overline{V^+}$ of energy–momentum space. In order to perform the collision theory of massive particles, one needs a more detailed "mass-gap assumption": Σ is the union of the origin O, of one or several positive sheets of hyperboloid $H^+_{m_i}$ and of a region V^+_M defined by the conditions $p^2 \geq M^2, p_0 > 0$, with M larger than all the m_i.

The Hilbert space $\mathcal{H}$ is correspondingly decomposed as the direct sum of the vacuum subspace (or zero-particle subspace) generated by Ω, of subspaces of stable one-particle states with masses m_i isomorphic to $L_2(H^+_{m_i}, \mu_{m_i})$, and of a remaining subspace $\mathcal{H}'$. As a result of the construction of "asymptotic states," $\mathcal{H}'$ can be shown to contain two subspaces $\mathcal{H}'_{\text{in}}$ and $\mathcal{H}'_{\text{out}}$, generated, respectively, by N-particle incoming states (with N arbitrary and ≥ 2) and by N-particle outgoing states. The collision operator S is then defined as the partially isometric operator from $\mathcal{H}'_{\text{out}}$ onto $\mathcal{H}'_{\text{in}}$, which maps a reference basis of outgoing states onto the corresponding basis of incoming states.

An independent postulate: asymptotic completeness (*see* **Scattering, Asymptotic Completeness and Bound States and Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools**) The theory is said to satisfy the property of asymptotic completeness if all the states of $\mathcal{H}$ can be interpreted as superpositions of various N-particle states (either in the incoming or in the outgoing state basis), namely if one has $\mathcal{H}' = \mathcal{H}'_{\text{in}} = \mathcal{H}'_{\text{out}}$. This property is not implied by the previous postulates on quantum fields, but its physical interpretation and its role in the analytic program are of primary importance (see Scattering in Relativistic Quantum Field Theory: The Analytic Program). Let us simply note here that asymptotic completeness implies as a by-product the unitarity property of the collision operator S on the full Hilbert space $\mathcal{H}'$ (i.e., $SS^* = S^*S = I$).

Connection between retarded kernel and scattering kernel for the forward scattering case; a simple "reduction formula" We consider the scattering of a particle Π_1 with mass m_1 on a target consisting of a particle Π_2 with mass m_2 and denote by $T(p_1, p_2; p'_1, p'_2)$ the corresponding scattering kernel (defined similarly as for the case of equal-mass particles considered earlier). Equations [22]–[24] are then applied to the case when Ψ and Ψ' coincide with a one-particle state of Π_2 at rest, namely with 4-momentum $p_2 = p'_2$ along the time axis: $p_2 = ((p_2)_0, 0), (p_2)_0 = m_2$. This describes in a simple way the case of forward scattering, since in view of the energy–momentum conservation law $p_1 + p_2 = p'_1 + p'_2$, the choice $p_2 = p'_2$ also implies that $p_1 = p'_1$. (The possibility of restricting the distribution $T(p_1, p_2; p'_1, p'_2)$ to such fixed values of the energy–momenta is shown to be mathematically well justified). The advantage of this simple case is that the corresponding kernels [22], [23] of (x_1, x'_1) are invariant under spacetime translations and therefore depend only on x (and not on X). We can thus rewrite eqns [22], [23] with simplified notations as follows:

$$C_{p_2}(x) \doteq \langle p_2, \left[\Phi\left(\frac{x}{2}\right), \Phi\left(-\frac{x}{2}\right)\right] p_2 \rangle = R_{p_2}(x) - A_{p_2}(x) \qquad [25]$$

which can be shown to give correspondingly by Fourier transformation

$$\tilde{C}_{p_2}(p) = \langle p_2, \tilde{\Phi}(p)\tilde{\Phi}(-p)p_2 \rangle - \langle p_2, \tilde{\Phi}(-p)\tilde{\Phi}(p)p_2 \rangle = \tilde{R}_{p_2}(p) - \tilde{A}_{p_2}(p) \qquad [26]$$

If the particle Π_1 appears in the asymptotic states of the field Φ, the scattering kernel $T(p_1, p_2; p'_1, p'_2)$ is then given in the forward configurations $p_1 = p'_1 \in H^+_{m_1}, p_2 = p'_2 \in H^+_{m_2}$, by the following reduction formula in which $s = (p_1 + p_2)^2$:

$$F_0(s) = T(p_1, p_2; p_1, p_2) = \left[(p_1^2 - m_1^2)\tilde{R}_{p_2}(p_1)\right]_{|H^+_{m_1}} \qquad [27]$$

Analyticity Domains in Energy–Momentum Space: From the "Primitive Off-Shell Domains" of QFT to the Crossing Manifolds on the Mass Shell

For simplicity, we shall restrict ourselves to the consideration of forward scattering amplitudes, namely to the derivation of crossing analyticity domains and (quasi-)dispersion relations at $t = 0$ for two-particle collision processes of the form $\Pi_1 + \Pi_2 \to \Pi_1 + \Pi_2$, Π_1 and Π_2 being given massive particles with arbitrary spins and charges.

The holomorphic function $H_{p_2}(k)$ and its primitive domain D. Nontriviality of dispersion relations for the scattering amplitudes As suggested by eqn [24], we can exploit the analyticity properties of the Fourier–Laplace transforms of the retarded and advanced kernels R_{p_2} and A_{p_2}: in fact, $\tilde{R}_{p_2}(p)$ and

$\tilde{A}_{p_2}(p)$ are, respectively, the boundary values of the holomorphic functions

$$\tilde{R}^{(c)}_{p_2}(k) = \int_{\overline{V}^+} R_{p_2}(x)\, e^{ik\cdot x}\, dx$$
$$\tilde{A}^{(c)}_{p_2}(k) = \int_{\overline{V}^-} A_{p_2}(x)\, e^{ik\cdot x}\, dx \qquad [28]$$

from the corresponding domains $\mathcal{T}^+$ and $\mathcal{T}^-$. According to the reduction formula [27], it is appropriate to consider correspondingly the functions $H^+_{p_2}(k) \doteq (k^2 - m_1^2)\tilde{R}^{(c)}_{p_2}(k)$ and $H^-_{p_2}(k) \doteq (k^2 - m_1^2)\tilde{A}^{(c)}_{p_2}(k)$, which are also, respectively, holomorphic in $\mathcal{T}^+$ and $\mathcal{T}^-$. Then the forward scattering amplitude $F_0(s) = F(s,0) = T(p_1,p_2;p_1,p_2)$ appears as the restriction to the hyperboloid sheet $p \in H^+_{m_1}$ of the boundary value $H^+_{p_2}(p)$ of $H^+_{p_2}(k)$ on the reals.

Moreover, it can be seen that the two boundary values $H^+_{p_2}(p) = (p^2 - m_1^2)\tilde{R}_{p_2}(p)$ and $H^-_{p_2}(p) = (p^2 - m_1^2)\tilde{A}_{p_2}(p)$ coincide as distributions in the region

$$\mathcal{R} = \{p \in \mathbf{R}^4; (p+p_2)^2 < (m_1+m_2)^2; \quad (p-p_2)^2 < (m_1+m_2)^2\} \qquad [29]$$

This follows from the intermediate expression in eqn [26] and from the fact that a state of the form $(p^2 - m_1^2)\Phi(\pm p)p_2 >$ is a state of energy–momentum $\pm p + p_2$ and therefore vanishes (in view of the spectral condition) if $(\pm p + p_2)^2 < (m_1+m_2)^2$ (here we also use a simplifying assumption according to which no one-particle bound state is present in this channel).

The situation obtained concerning the holomorphic functions $H^+_{p_2}(k)$ and $H^-_{p_2}(k)$ parallels (in complex dimension four) the case of a pair of holomorphic functions in the upper and lower half-planes whose boundary values on the reals coincide on a certain interval playing the role of $\mathcal{R}$. As in this one-dimensional case there is a theorem, called the "edge-of-the-wedge theorem" (see below), which implies that $H^+_{p_2}(k)$ and $H^-_{p_2}(k)$ have a common analytic continuation $H_{p_2}(k)$: this function is holomorphic in a domain D which is the union of $\mathcal{T}^+, \mathcal{T}^-$ and of a complex neighborhood of $\mathcal{R}$; D is called the primitive domain of $H_{p_2}(k)$.

Moreover, it follows from the postulate of invariance of the field $\Phi(x)$ under the action of the Poincaré group (see postulate (2)) that the holomorphic function $H_{p_2}(k)$ only depends of the two complex variables $\zeta = k^2 (= k_0^2 - \mathbf{k}^2)$ and $k \cdot p_2$ or equivalently $s = (k+p_2)^2 = \zeta + m_2^2 + 2k\cdot p_2$; it thus defines a corresponding holomorphic function $\hat{H}_{p_2}(\zeta, s) \doteq H_{p_2}(k)$ in the image of D in these variables.

In view of the reduction formula [27], the scattering function $\hat{F}_0(s)$ should appear as the restriction of the holomorphic function $\hat{H}_{p_2}(\zeta, s)$ to the physical mass-shell value $\zeta = m_1^2$. However, it turns out that the section of D by the complex mass-shell manifold $\mathcal{M}^{(c)}$ with equation $k^2 = m_1^2$ is empty: this geometrical fact is responsible for the nontriviality of the proof of dispersion relations for the physical quantity $\hat{F}_0(s)$ on the mass shell. In fact, the tube $T^+ \cup T^-$ which constitutes the basic part of the domain D and is given by the field-theoretical microcausality postulate, is a "purely off-shell" complex domain, as it can be easily checked: if a complex point $k = p + iq$ is such that $q^2 > 0$, the corresponding squared mass $\zeta = k^2 = p^2 - q^2 + 2ip\cdot q$ is real if and only if $p \cdot q = 0$, which implies $p^2 < 0$ (i.e., p spacelike) and therefore $\zeta = p^2 - q^2 < 0$.

"Off-shell dispersion relations" as a first step The starting point, which is easy to obtain from the domain D, is the analyticity of the holomorphic function $\hat{H}_{p_2}(\zeta, s)$ in a cut-plane of the variable s for all negative values of the squared mass variable ζ. This cut-plane Δ_ζ is always the complement in $\mathbf{C}$ (i.e., the complex s-plane) of the union of the s-cut (s real $\geq (m_1+m_2)^2$) and of the u-cut ($u = 2\zeta + 2m_2^2 - s$ real $\geq (m_1+m_2)^2$). This analyticity property thus justifies "off-shell dispersion relations" at fixed negative values of ζ for the field-theoretical structure function $\hat{H}_{p_2}(\zeta, s)$.

The latter property and the subsequent analysis concerning the process of analytic continuation of $\hat{H}_{p_2}$ to positive values of ζ will be more easily understood geometrically if one reduces the complex space of k to a two-dimensional complex space, which is legitimate in view of the equality $H_{p_2}(k) = \hat{H}_{p_2}(\zeta, s)$.

Having chosen the k_0-axis along p_2, we reduce the orthogonal space coordinates $\mathbf{k}$ of k to the radial variable k_r. One thus gets the following expressions of the variables ζ and s (resp. u):

$$\zeta = k_0^2 - k_r^2, \quad s = \zeta + m_2^2 + 2m_2 k_0$$
$$(\text{resp. } u = \zeta + m_2^2 - 2m_2 k_0)$$

Then we can write $\hat{H}_{p_2}(\zeta, s) \doteq \mathrm{H}_{p_2}(k_0, k_r) = \mathrm{H}_{p_2}(k_0, -k_r)$, and describe the image D_r of the domain D in the variables $k \doteq (k_0, k_r) = p + iq$ as $T^+_r \cup T^-_r \cup \mathcal{N}(\mathcal{R}_r)$, where:

1. $T^\pm_r$ is defined by the condition $q^2 \doteq q_0^2 - q_r^2 > 0$, $q_0 > 0$ or $q_0 < 0$,
2. $\mathcal{N}$ is a complex neighborhood of the real region $\mathcal{R}_r$ defined as follows. Let h^+_s, h^-_u be the two branches of hyperbolae with respective equations:

$$h^+_s : (p_0 + m_2)^2 - p_r^2 = (m_1+m_2)^2, \; p_0 + m_2 > 0$$
$$h^-_u : (p_0 - m_2)^2 - p_r^2 = (m_1+m_2)^2, \; p_0 - m_2 < 0$$

Then $\mathcal{R}_r$ is the intersection of the region situated below h_s^+ and of the region situated above h_u^-.

Let us now consider any complex hyperbola $h^{(c)}[\zeta]$ with equation $k^2 \doteq k_0^2 - k_r^2 = \zeta$. On such a complex curve either one of the variables k_0 or s or u is a good parameter for holomorphic functions which are even in k_r, like $H_{p_2}(k_0, k_r)$. If ζ is real, any complex point $k = p + iq$ of $h^{(c)}[\zeta]$ is such that p^2 and q^2 have opposite signs (since $p \cdot q = 0$). Therefore, the sign of q^2 is always opposite to the sign of $\zeta (= p^2 - q^2)$: if ζ is negative, all the complex points of $h^{(c)}[\zeta]$ thus belong to $T_r^+ \cup T_r^-$; the union of all these points with the real points of $h^{(c)}[\zeta]$ in $\mathcal{R}_r$ is therefore a subset of D_r, which is represented in the complex plane of s by the cut-plane Δ_ζ. The function $\hat{H}_{p_2}(\zeta, s)$ is therefore analytic (and univalent) in Δ_ζ for each $\zeta < 0$. Moreover, the existence of moderate bounds of type [12] on H_{p_2} in D (resulting from the temperateness assumption) then implies the validity of dispersion relations (with subtractions) for $\hat{H}_{p_2}(\zeta, s)$ in Δ_ζ.

The problem of analytic completion to the complex mass-shell hyperbola $h^{(c)}[m_1^2]$: what is provided by the Jost–Lehmann–Dyson domain A basic fact in complex geometry in n variables, with $n \geq 2$, is the existence of a distinguished class of domains, called holomorphy domains: for each domain $\mathcal{U}$ in this class, there exists at least one function which is holomorphic in $\mathcal{U}$ and cannot be analytically continued at any point of the boundary of $\mathcal{U}$. In one dimension, every domain is a holomorphy domain. In dimension larger than one, a general domain $\mathcal{U}$ is not a holomorphy domain, but it admits a holomorphy envelope $\hat{\mathcal{U}}$, which is a holomorphy domain containing $\mathcal{U}$, such that every function holomorphic in $\mathcal{U}$ admits an analytic continuation in $\hat{\mathcal{U}}$.

It turns out that the domain D_r considered above in the last subsection) is not a holomorphy domain; its holomorphy envelope $\hat{D}_r$ (obtained geometrically by Bros, Messiah, and Stora in 1961) coincides with a domain introduced by Jost–Lehmann (1957) and Dyson (1958) by methods of wave equations. This domain can be characterized as the union of D_r with all the complex points of all the hyperbolae with equations $(k_0 - a)^2 - (k_r - b)^2 = c^2$ (for all a, b, c real, including the complex straight lines for which $c = 0$) whose both branches have a nonempty intersection with the real region $\mathcal{R}_r$.

In particular, one easily sees that all the hyperbolae $h^{(c)}[\zeta]$ with $0 \leq \zeta < m_1^2$ belong to the previous class. It follows that for any ζ in this positive interval, the function $\hat{H}_{p_2}(\zeta, s)$ can still be analytically continued as a holomorphic function of s in the cut-plane Δ_ζ and thereby satisfies the corresponding dispersion relations.

The physical mass shell hyperbola $h^{(c)}[m_1^2]$ thus appears as a limiting case of the previous family (for ζ tending to m_1^2 from below). The analyticity of $\hat{H}_{p_2}(m_1^2, s)$ in $\Delta_{m_1^2}$ can then be justified provided one knows that this function is analytic at at least one point of $\Delta_{m_1^2}$: but this additional information results from a more thorough exploitation of the analyticity properties resulting from the QFT postulates. This will be now briefly outlined below.

Further information coming from the four-point function in complex momentum space It is possible to obtain further analyticity properties of $\hat{H}_{p_2}(\zeta, s) \doteq H_{p_2}(k)$ by considering the latter as the restriction to the submanifold $k_1 = -k_3 = k$; $k_2 = -k_4 = p_2$ of a master analytic function $H_4(k_1, k_2, k_3, k_4)$, called the four-point function of the field Φ in complex energy–momentum space (*see* Scattering in Relativistic Quantum Field Theory: The Analytic Program). This function is holomorphic in a well-defined primitive domain D_4 of the linear submanifold $k_1 + k_2 + k_3 + k_4 = 0$. It is then possible to compute some local parts situated near the reals of the holomorphy envelope of D_4, which implies, as a by-product, that the function $\hat{H}_{p_2}(\zeta, s)$ can be analytically continued in a set Σ of the form

$$\Sigma = \{(\zeta, s);\ \zeta \in \delta,\ s \in \mathcal{V}_{s_1}(\zeta)\} \cup \{(\zeta, s);\ \zeta \in \delta,\ u = 2\zeta + 2m_2^2 - s \in \mathcal{V}_{u_1}(\zeta)\} \quad [30]$$

with the following specifications:

1. δ is a domain in the ζ-plane, which is a complex neighborhood of a real interval of the form $-a < \zeta < M_1^2$; here M_1 denotes a spectral mass threshold in the theory such that $M_1 > m_1$;
2. for each ζ, $\mathcal{V}_{s_1}(\zeta)$ (resp. $\mathcal{V}_{u_1}(\zeta)$) is a cut-neighborhood in the s-plane of the real half-line $s > s_1$ (resp. of the half-line $u = 2\zeta + 2m_2^2 -$ s real $> u_1$); s_1 and u_1 denote appropriate real numbers independent of ζ.

The final analytic completion: crossing domains on $h^{(c)}[m_1^2]$. Dispersion relations for π_0–π_0 meson scattering and "quasi-dispersion-relations" for proton–proton scattering We now wish to describe briefly the final step of analytic completion, which displays the existence of a "quasi-cut-plane domain" in s for the function $\hat{H}_{p_2}(m_1^2, s)$, even in the more general case when the s-cut and u-cut are associated with different scattering channels, whose respective mass thresholds $s = M_{12}^2$ and $u = M_{12}'^2$ are unequal.

This general situation may occur as soon as one charged particle Π_1 of the s-channel is replaced by the corresponding antiparticle $\overline{\Pi}_1$ in the u-channel, in contrast with the case of neutral particles (like the π_0 meson) which coincide with their own antiparticles. Here it is important to note that the two real branches $h^+[m_1^2]$ and $h^-[m_1^2]$ of the mass shell hyperbola $h^{(c)}[m_1^2]$ correspond, respectively, to the physical region of the "direct scattering channel" of the reaction $\Pi_1 + \Pi_2 \to \Pi_1 + \Pi_2$ with squared total energy s, and to the physical region of the "crossed scattering channel" of the reaction $\overline{\Pi}_1 + \Pi_2 \to \overline{\Pi}_1 + \Pi_2$ with squared total energy u. A typical and important example is the case of proton–proton scattering in the s-channel, where M_{12} equals twice the mass $m(=m_1=m_2)$ of the proton, while the corresponding u-channel refers to the proton–antiproton scattering, whose threshold M'_{12} equals twice the mass μ of the π meson.

In that general case, the analysis of the subsection "'Off-shell dispersion relations' as a first step" still applies, so that the function $\hat{H}_{p_2}(\zeta, s)$ is always analytic in a set of the form

$$S_0 = \{(\zeta, s);\ -a < \zeta < 0,\ s \in \Delta_\zeta\} \qquad [31]$$

Then, the additional information described above in the last subsection allows one to use the following crucial property of analytic completion, which we call

Crossing lemma *If a function $G(\zeta, s)$ is holomorphic in a domain which contains the union of the sets Σ and S_0 (see eqns* [30] *and* [31]*), then it admits an analytic continuation in a set of the following form:*

$$\{(\zeta, s);\ \zeta \in \delta,\ s \in \Delta_\zeta;$$
$$|s - \zeta - m_2^2| = |u - \zeta - m_2^2| > R(\zeta)\}$$

By applying this property to the function $\hat{H}_{p_2}(\zeta, s)$ and restricting ζ to the mass-shell value m_1^2 which belongs to δ, one obtains the analyticity of the scattering function $\hat{F}_0(s) \doteq \hat{H}_{p_2}(m_1^2, s)$ in a crossing domain of the complex mass shell hyperbola $h^{(c)}[m_1^2]$: the crossing between the two physical regions $h^+[m_1^2]$ $(s \geq M_{12}^2)$ and $h^-[m_1^2]$ $(u \geq M_{12}'^2)$ is ensured by a complex domain of $h^{(c)}[m_1^2]$ whose image in the s-plane is the "cut-neighborhood of infinity" $\{s; s \in \Delta_{m_1^2}, |s - m_1^2 - m_2^2| = |u - m_1^2 - m_2^2| > R(m_1^2)\}$. Note that the relevant boundary values of $\hat{F}_0$ for obtaining the scattering amplitudes of the two collision processes with respective physical regions $h^+[m_1^2]$ and $h^-[m_1^2]$ have to be taken from the respective sides $\operatorname{Im} s > 0$ and $\operatorname{Im} u = -\operatorname{Im} s > 0$ of the corresponding s- and u-cuts.

It is only for the neutral case, where $M_{12} = M'_{12} = m_1 + m_2$, that a more favorable scenario occurs, as explained earlier: in this case, the interval $\{\zeta \in]-a, 0[\}$ of the set S_0 is replaced by $\{\zeta \in]-a, m_1^2[\}$, so that the whole cut-plane domain $\Delta_{m_1^2}$ is obtained in the result of the previous crossing lemma. The scattering amplitudes of π_0–π_0 meson scattering and of π meson–proton scattering enjoy this property and, therefore, satisfy genuine dispersion relations in which the scattering function is even (see the second basic example described at the beginning of this article). In the general case of crossing domains obtained above, corresponding Cauchy integral relations have been written and used under the name of "quasi-dispersion-relations."

Complementary results Some comments can now be added concerning the passage from the purely geometrical results (i.e., analyticity domains) described above to the writing of precise (quasi-) dispersion relations with two subtractions:

Polynomial bounds and dispersion relations with N subtractions The previous methods of analytic completion also allow one to control the bounds at infinity in the relevant complex domains. As it has been noticed after eqn [24], the Fourier–Laplace transforms of the retarded and advanced kernels, and thereby the holomorphic functions $H^{\pm}_{p_2}(k)$ discussed at the start of this section are bounded at most by a power of a suitable norm of k in their respective tubes $\mathcal{T}^{\pm}$. Correspondingly, the holomorphic function $H_{p_2}(k)$ (resp. $H_{p_2}(k_0, k_r)$) admits the same type of bound in its primitive analyticity domain D (resp. D_r). These bounds are a consequence of the tempered distribution character of the structure functions of the fields which is built-in in the Wightman field-theoretical framework. Then it can be checked that in the holomorphy envelope $\hat{D}_r$ of D_r, and thereby in the cut-plane (or crossing) domains obtained in the intersection of $\hat{D}_r$ and of the complex mass shell $h^{(c)}[m_1^2]$, the same type of power bound is still valid: $\hat{F}_0(s)$ is therefore bounded by some power $|s|^{N-1}$ of $|s|$ and thus satisfies a (quasi-)dispersion relation with N subtractions. The same type of argument holds for all the similar cut-domains (or crossing domains) in s obtained for $\hat{F}_t(s)$ for all negative value of t.

It is also worthwhile to mention that a similar remarkable (since not at all predictable) result was also obtained in the Haag, Kastler, and Araki framework of algebraic QFT (Epstein, Glaser, Martin, 1969; *see* Scattering in Relativistic Quantum Field Theory: The Analytic Program for further comments).

In this connection, one can also mention a more recent result. In the Buchholz–Fredenhagen axiomatic approach of charged fields (1982), in which

locality is replaced by the more general notion of "stringlike locality" (*see* Algebraic Approach to Quantum Field Theory, Axiomatic Quantum Field Theory, and Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools), a proof of forward dispersion relations has again been obtained (Bros, Epstein, 1994).

The extension of the analyticity domains by positivity and the derivation of bounds by unitarity (Martin 1966; see the book by Martin (1969)). The following ingredients have been used:

1. Positivity conditions on the absorptive part of $F(s,t)$, which are expressed by the infinite set of inequalities $(\mathrm{d}/\mathrm{d}t)^n \mathrm{Im} F(s,t)_{|t=0} \geq 0$ (for all integers n),
2. The existence of a two-dimensional complex neighborhood of some point $(s=s_0, t=0)$ in the analyticity domain resulting from QFT.

The following results have then been obtained:

(a) It is justified to differentiate the forward (subtracted) dispersion relations with respect to t at any order.
(b) $\hat{F}(s,t)$ can be analytically continued in a fixed circle $|t| < t_{\max}$ for all values of s. The latter implies the extension of dispersion relations in s to positive (and complex) values of t.
(c) In a last step, the use of unitarity conditions for the "partial waves" $f_\ell(s)$ of $F(s,t)$ (*see* Scattering in Relativistic Quantum Field Theory: The Analytic Program) allows one to obtain Froissart-type bounds on the scattering amplitudes and thereby to justify the writing of (quasi-)dispersion relations with at most two subtractions for all the admissible values of t.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Perturbation Theory and its Techniques; Scattering in Relativistic Quantum Field Theory: The Analytic Program; Scattering, Asymptotic Completeness and Bound States; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools.

Further Reading

Haag R (1992) *Local Quantum Physics*.Berlin: Springer.
Iagolnitzer D (1992) *Scattering in Quantum Field Theories: The Axiomatic and Constructive Approaches*. Princeton Series in Physics. Princeton: Princeton University Press.
Jost R (1965) *The General Theory of Quantized Fields*. AMS, Providence, RI: American Mathematical Society.
Klein L (ed.) (1961) *Dispersion Relations and the Abstract Approach to Field Theory*. New York: Gordon and Breach.
Martin A (1969) *Scattering Theory: Unitarity, Analyticity and Crossing*, Lecture Notes in Physics. Berlin: Springer.
Nussenzweig HM (1972) *Causality and Dispersion Relations*. New York: Academic Press.
Streater RF and Wightman AS (1964, 1980) *PCT, Spin and Statistics, and all that*. Princeton: Princeton University Press.
Vernov YuS (1996) *Dispersion Relations in the Historical Aspect*, IHEP Publications, Protvino Conf.: Fundamental Problems of High Energy Physics and Field Theory.

Euclidean Field Theory

F Guerra, Università di Roma "La Sapienza," Rome, Italy

© 2006 Elsevier Ltd. All rights reserved.

Introduction

In this article, we consider Euclidean field theory as a formulation of quantum field theory which lives in some Euclidean space, and is expressed in probabilistic terms. Methods arising from Euclidean field theory have been introduced in a very successful way in the study of concrete models of constructive quantum field theory.

Euclidean field theory was initiated by Schwinger (1958) and Nakano (1959), who proposed to study the vacuum expectation values of field products analytically continued into the Euclidean region (Schwinger functions), where the first three (spatial) coordinates of a world point are real and the last one (time) is purely imaginary (Schwinger points). The possibility of introducing Schwinger functions, and their invariance under the Euclidean group are immediate consequences of the now classic formulation of quantum field theory in terms of vacuum expectation values given by Wightman (Streater and Wightman 1964). The convenience of dealing with the Euclidean group, with its positive-definite scalar product, instead of the Lorentz group, is evident, and has been exploited by several authors, in different contexts.

The next step was made by Symanzik (1966), who realized that Schwinger functions for boson fields have a remarkable positivity property, allowing to introduce Euclidean fields on their own sake. Symanzik also pointed out an analogy between Euclidean field theory and classical statistical mechanics, at least for some interactions (Symanzik 1969).

This analogy was successfully extended, with a different interpretation, to all boson interactions by Guerra *et al.* (1975), with the purpose of using rigorous results of modern statistical mechanics for

Notes:

the study of constructive quantum field theory, within the program advocated by Wightman (1967), and further pursued by Glimm and Jaffe (see Glimm and Jaffe (1981) for an overall presentation).

The most dramatic advance of Euclidean theory was due to Nelson (1973a, b). He was able to isolate a crucial property of Euclidean fields (the Markov property) and gave a set of conditions for these fields, which allow us to derive all properties of relativistic quantum fields satisfying Wightman axioms. The Nelson theory is very deep and rich in new ideas. Even after so many years since the basic papers were published, we lack a complete understanding of the radical departure from the conventional theory afforded by Nelson's ideas, especially about their possible further developments.

By using the Nelson scheme, in particular a very peculiar symmetry property, it was very easy to prove (Guerra 1972) the convergence of the ground-state energy density, and the van Hove phenomenon in the infinite-volume limit for two-dimensional boson theories. A subsequent analysis (Guerra *et al.* 1972) gave other properties of the infinite-volume limit of the theory, and allowed a remarkable simplification in the proof of a very important regularity property for fields, previously established by Glimm and Jaffe.

Since then, all work on constructive quantum field theory has exploited in different ways ideas coming from Euclidean field theory. Moreover, a very important reconstruction theorem has been established by Osterwalder and Schrader (1973), allowing a reconstruction of relativistic quantum fields from the Euclidean Schwinger functions, and avoiding the previously mentioned Nelson reconstruction theorem, which is technically more difficult to handle.

This article is intended to be an introduction to the general structure of Euclidean quantum field theory, and to some of the applications to constructive quantum field theory. Our purpose is to show that, 50 years after its introduction, the Euclidean theory is still interesting, both from the point of view of technical applications and physical interpretation.

The article is organized as follows. In the next section, by considering simple systems made of a single spinless relativistic particle, we introduce the relevant structures in both Euclidean and Minkowski worlds. In particular, a kind of (pre)Markov property is introduced already at the one-particle level.

Next we present a description of the procedure of second quantization on the one-particle structure. The free Markov field is introduced, and its crucial Markov property explained. Following Nelson, we use probabilistic concepts and methods, whose relevance for constructive quantum field theory became immediately more and more apparent. The very structure of classical statistical mechanics for Euclidean fields is firmly based on these probabilistic methods. This is followed by an introduction of interaction, and we show the connection between the Markov theory and the Hamiltonian theory, for two-dimensional space-cutoff interacting scalar fields. In particular, we present the Feynman–Kac–Nelson formula that gives an explicit expression of the semigroup generated by the space-cutoff Hamiltonian in Φ_{2}^{4} space. We also deal with some applications to constructive quantum field theory. This is followed by a short discussion about the physical interpretation of the theory. In particular, we discuss the Osterwalder–Schrader reconstruction theorem on Euclidean Schwinger functions, and the Nelson reconstruction theorem on Euclidean fields. For the sake of completeness, we sketch the main ideas of a proposal, advanced in Guerra and Ruggiero (1973), according to which the Euclidean field theory can be interpreted as a stochastic field theory in the physical Minkowski spacetime.

Our treatment will be as simple as possible, by relying on the basic structural properties, and by describing methods of presumably very long lasting power. The emphasis given to probabilistic methods, and to the statistical mechanics analogy, is a result of the historical development. Our opinion is that not all possibilities of Euclidean field theory have been fully exploited yet, both from technical and physical points of view.

One-Particle Systems

A system made of only one relativistic scalar particle, of mass $m > 0$, has a quantum state space represented by the positive-frequency solutions of the Klein–Gordon equation. In momentum space, with points $p_\mu, \mu = 0, 1, 2, 3$, let us introduce the upper mass hyperboloid, characterized by the constraints $p^2 \equiv p_0^2 - \sum_{i=1}^{3} p_i^2 = m^2, p_0 \geq m$, and the relativistic invariant measure on it, formally given by $d\mu(p) = \theta(p_0)\delta(p^2 - m^2)\, dp$, where θ is the step function $\theta(x) = 1$ if $x \geq 0$, and $\theta(x) = 0$ otherwise, and dp is the four-dimensional Lebesgue measure. The Hilbert space of quantum states F is given by the square-integrable functions on the mass hyperboloid equipped with the invariant measure $d\mu(p)$. Since in some reference frame the mass hyperboloid is uniquely characterized by the space values of the momentum $\boldsymbol{p}$, with the energy given by $p_0 \equiv \omega(p) = \sqrt{\boldsymbol{p}^2 + m^2}$, the Hilbert space F of the states is, in fact, made of those complex-valued tempered distributions f in the configuration space R^3 whose Fourier transforms, $\tilde{f}(\boldsymbol{p})$, are square-integrable functions in momentum space with respect to the image of the relativistic invariant measure $d\boldsymbol{p}/2\omega(\boldsymbol{p})$, where

$d\boldsymbol{p}$ is the Lebesgue measure in momentum space. The scalar product on F is defined by

$$\langle f, g\rangle_F = (2\pi)^3 \int \tilde{f}^*(\boldsymbol{p})\tilde{g}(\boldsymbol{p}) \frac{d\boldsymbol{p}}{2\omega(\boldsymbol{p})}$$

where we have normalized the Fourier transform in such a way that

$$f(\boldsymbol{x}) = \int \exp(i\boldsymbol{p}.\boldsymbol{x})\tilde{f}(\boldsymbol{p})\, d\boldsymbol{p}$$

$$\tilde{f}(\boldsymbol{p}) = (2\pi)^{-3}\int \exp(-i\boldsymbol{p}.\boldsymbol{x})\tilde{f}(\boldsymbol{x})\, d\boldsymbol{x}$$

$$\int \exp(i\boldsymbol{p}.\boldsymbol{x})\, d\boldsymbol{p} = (2\pi)^3\delta(\boldsymbol{x})$$

The scalar product on F can also be expressed in the form

$$\langle f, g\rangle_F = \int\int f(\boldsymbol{x}')^* W(\boldsymbol{x}' - \boldsymbol{x})g(\boldsymbol{x})\, d\boldsymbol{x}'\, d\boldsymbol{x}$$

where we have introduced the two-point Wightman function at fixed time, defined by

$$W(\boldsymbol{x}' - \boldsymbol{x}) = (2\pi)^{-3}\int \exp(i\boldsymbol{p}.(\boldsymbol{x}' - \boldsymbol{x})) \frac{d\boldsymbol{p}}{2\omega(\boldsymbol{p})}$$

A unitary irreducible representation of the Poincaré group can be defined on F in the obvious way. In particular, the generators of space translations are given by multiplication by the components of $\boldsymbol{p}$ in momentum space, and the generator of time translations (the energy of the particle) is given by $\omega(\boldsymbol{p})$.

For the scalar product of time-evolved wave functions, we can write

$$\langle \exp(-it')f, \exp(-it)g\rangle_F = \int\int f(\boldsymbol{x}')^* W(t' - t, \boldsymbol{x}' - \boldsymbol{x})g(\boldsymbol{x})\, d\boldsymbol{x}' d\boldsymbol{x}$$

where we have introduced the two-point Wightman function, defined by

$$W(t' - t, \boldsymbol{x}' - \boldsymbol{x}) = (2\pi)^{-3}\int \exp(-i(t - t'))\exp(i\boldsymbol{p}.(\boldsymbol{x}' - \boldsymbol{x})) \frac{d\boldsymbol{p}}{2\omega(\boldsymbol{p})}$$

To the physical single-particle system living in Minkowski spacetime, we associate a kind of mathematical image, living in Euclidean space, from which all properties of the physical system can be easily derived. We start from the two-point Schwinger function

$$S(x) = \frac{1}{(2\pi)^4}\int \frac{\exp(ip \cdot x)}{\sqrt{p^2 + m^2}}\, dp$$

which is the analytic continuation of the previously given two-point Wightman function into the Schwinger points. Here $x, p \in R^4$, and $p \cdot x = \sum_{i=1}^4 x_i p_i$. Here dp and dx are the Lebesgue measures in the R^4 momentum and configuration spaces, respectively. The function $S(x)$ is positive and analytic for $x \neq 0$, decreases as $\exp(-m\|x\|)$ as $x \to \infty$, and satisfies the equation

$$(-\Delta + m^2)S(x) = \delta(x)$$

where $\Delta = \sum_{i=1}^4 \partial^2/\partial x_i^2$ is the Laplacian in four dimensions.

The mathematical image we are looking for is described by the Hilbert space N of those tempered distributions in four-dimensional configuration space R^4 whose Fourier transforms are square integrable with respect to the measure $dp/\sqrt{p^2 + m^2}$. The scalar product on N is defined by

$$\langle f, g\rangle_N = (2\pi)^4 \int \tilde{f}^*(p)\tilde{g}(p) \frac{dp}{\sqrt{p^2 + m^2}}$$

Four-dimensional Fourier transforms are normalized as follows:

$$f(x) = \int \exp(ip.x)\tilde{f}(p)\, dp$$

$$\tilde{f}(p) = (2\pi)^{-4}\int \exp(-ip.x)\tilde{f}(x)\, dx$$

$$\int \exp(ip.x)\, dp = (2\pi)^4\delta(x)$$

We also write

$$\langle f, g\rangle_N = \int\int f^*(x)S(x - y)g(y)\, dx\, dy = \left\langle f, (-\Delta + m^2)^{-1}g\right\rangle$$

where $\langle , \rangle$ is the ordinary Lebesgue product defined on Fourier transforms and, in momentum space, $(-\Delta + m^2)^{-1}$ amounts to a multiplication by $(p^2 + m^2)^{-1}$. The Schwinger function $S(x - y)$ is formally the kernel of the operator $(-\Delta + m^2)^{-1}$. The Hilbert space N is the carrier space of a unitary (nonirreducible) representation of the four-dimensional Euclidean group $E(4)$. In fact, let (a, R) be an element of $E(4)$

$$(a, R) : R^4 \to R^4$$
$$x \to Rx + a$$

where $a \in R^4$, and R is an orthogonal matrix, $RR^T = R^T R = 1_4$. Then the transformation $u(a, R)$ defined by

$$u(a, R) : N \to N$$
$$f(x) \to (u(a, R)f)(x) = f(R^{-1}(x - a))$$

provides the representation. In particular, we consider the reflection r_0 with respect to the hyperplane

$x_4 = 0$, and the translations $u(t)$ in the x_4-direction. Then we have $r_0 u(t) r_0 = u(-t)$, and analogously for other hyperplanes.

Now we introduce a local structure on N by considering, for any closed region A of R^4, the subspace N_A of N made by distributions in N with support on A. We call e_A the orthogonal projection on N_A. It is obvious that if $A \in B$ then $N_A \in N_B$ and $e_A e_B = e_B e_A = e_A$. A kind of (pre)Markov property for one-particle systems is introduced as follows. Consider a closed three-dimensional piecewise smooth manifold σ, which divides R^4 in two closed regions A and B, having σ in common. Therefore, $\sigma \in A, \sigma \in B, A \cap B = \sigma, A \cup B = R^4$. Let N_A, N_B, N_σ, and e_A, e_B, e_σ be the associated subspaces and projections, respectively. Then $N_\sigma \subset N_A, N_\sigma \subset N_B$, and $e_\sigma e_A = e_A e_\sigma = e_\sigma, e_\sigma e_B = e_B e_\sigma = e_\sigma$. It is very simple to prove the following:

Theorem 1 *Let e_A, e_B, e_σ be defined as above, then $e_A e_B = e_B e_A = e_\sigma$.*

Clearly, it is enough to show that for any $f \in N$ we have $e_A e_B f \in N_\sigma$. In that case, $e_\sigma e_A e_B f = e_A e_B f$, from which the theorem easily follows. Since $e_A e_B f$ has support on A, we must show that for any C_0^∞ function g with support on A_σ we have $\langle g, e_A e_B f \rangle = 0$. Then $e_A e_B f$ has support on σ, and the proof is complete. Now we have

$$\begin{aligned}\langle g, e_A e_B f \rangle &= \langle (-\Delta + m^2) g, e_A e_B f \rangle_N \\ &= \langle e_A (-\Delta + m^2) g, e_B f \rangle_N \\ &= \langle (-\Delta + m^2) g, e_B f \rangle_N \\ &= \langle g, e_B f \rangle = 0\end{aligned}$$

where we have used the definition of $\langle\,\rangle_N$ in terms of $\langle\,\rangle$, the fact that $e_A(-\Delta + m^2)g = (-\Delta + m^2)g$, since $(-\Delta + m^2)g$ has support on A_σ, and the fact that $e_B f$ has support on B. This ends the proof of the (pre)Markov property for one-particle systems.

A very important role in the theory is played by subspaces of N associated to hyperplanes in R^4. To fix ideas, consider the hyperplane $x_4 = 0$ and the associated subspace N_0. A tempered distribution in N with support on $x_4 = 0$ has necessarily the form $(f \otimes \delta_0)(x) \equiv f(x)\delta(x_4)$, with $f \in F$. By using the basic magic formula, for $x \geq 0$ and $M > 0$,

$$\int_{-\infty}^{+\infty} \frac{\exp(ipx)}{p^2 + M^2}\,dp = \frac{\pi}{M}\exp(-Mx)$$

it is immediate to verify that $\|f \otimes \delta_0\|_N = \|f\|_F$. Therefore, we have an isomorphic and isometric identification of the two Hilbert spaces F and N_0. Obviously, similar considerations hold for any hyperplane. In particular, we consider the hyperplanes $x_4 = t$ and the associated subspaces N_t. Let us introduce injection operators j_t defined by

$$\begin{aligned} j_t &: F \to N \\ &f \to f \otimes \delta_t \end{aligned}$$

where f is a generic element of F, with values $f(x)$, and $(f \otimes \delta_t)(x) = f(x)\delta(x_4 - t)$. It is immediate to verify the following properties for j_t and its adjoint j_t^*: the range of j_t is N_t; moreover, j_t is an isometry, so that $j_t^* j_t = 1_F, j_t j_t^* = e_t$, where 1_F is the identity on F, and e_t is the projection on N_t. Moreover, $e_t j_t = j_t$ and $j_t^* = j_t^* e_t$.

If we introduce translations $u(t)$ along the x_4-direction and the reflection r_0 with respect to $x_4 = 0$, then we also have the covariance property $u(t) j_s = j_{t+s}$, and the reflexivity property $r_0 j_0 = j_0$, $j_0^* r_0 = j_0^*$. The reflexivity property is very important. It tells us that r_0 leaves N_0 pointwise invariant, and it is an immediate consequence of the fact that $\delta(x_4) = \delta(-x_4)$.

Therefore, if we start from N we can obtain F, by taking the projection j_π with respect to some hyperplane π, in particular $x_4 = 0$. It is also obvious that we can induce on F a representation of $E(3)$ by taking those elements of $E(4)$ that leave π invariant.

Let us now see how we can define the Hamiltonian on F starting from the properties of N. Since we are considering the simple case of the one-particle system, we could just perform the following construction explicitly by hand, through a simple application of the basic magic formula given earlier. But we prefer to follow a route that emphasizes Markov property and can be immediately generalized to more complicated cases.

Let us introduce the operator $p(t)$ on F defined by the dilation $p(t) = j_0^* j_t = j_0^* u(t) j_0, t \geq 0$. Then we prove the following:

Theorem 2 *The operator $p(t)$ is bounded and self-adjoint. The family $\{p(t)\}$, for $t \geq 0$, is a norm-continuous semigroup.*

Proof Boundedness and continuity are obvious. Self-adjointness is a consequence of reflexivity. In fact,

$$p^*(t) = j_0^* u(-t) j_0 = j_0^* r_0 u(t) r_0 j_0 = j_0^* u(t) j_0 = p(t)$$

The semigroup property is a consequence of the Markov property. In fact, let us introduce N_+, N_0, N_- as subspaces of N made by distributions with support in the regions $x_4 \geq 0, x_4 = 0, x_4 \leq 0$, respectively, and call e_+, e_0, e_- the respective projections. By Markov property, we have $e_0 = e_- e_+$. Now write, for $s, t \geq 0$,

$$p(t)p(s) = j_0^* u(t) j_0 j_0^* u(s) j_0 = j_0^* u(t) e_0 u(s) j_0$$

If e_0 could be cancelled, then the semigroup property would follow from the group property of the translations $u(t)u(s)=u(t+s)$ (a miracle of the dilations!). For this, consider the matrix element

$$\langle f, p(t)p(s)g\rangle_F = \langle u(-t)j_0 f, e_0 u(s)j_0 g\rangle_N$$

recall $e_0 = e_- e_+$, and use $u(s)j_0 g \in N_+$ and $u(-t)j_0 f \in N_-$.

Let us call h the generator of $p(t)$, so that $p(t)=\exp(-th)$, for $t \geq 0$. By definition, h is the Hamiltonian of the physical system. A simple explicit calculation shows that h is just the energy ω introduced earlier. Starting from the representation of the Euclidean group $E(3)$ already given and from the Hamiltonian, we immediately get a representation of the full Poincaré group on F. Therefore, all physical properties of the one-particle system have been reconstructed from its Euclidean image on the Hilbert space N.

As a last remark of this section, let us note that we can consider the real Hilbert spaces N_r and F_r, made of real elements (in configuration space) in N and F. The operators $u(a,t), u(t), r_0, j_\pi, j_\pi^*, e_A$ are all reality preserving, that is, they map real spaces into real spaces.

This completes our discussion about the one-particle system. For more details we refer to Guerra *et al.* (1975) and Simon (1974). We have introduced the Euclidean image, discussed its main properties, and shown how we can derive all properties of the physical system from its Euclidean image. In the next sections, we will show how this kind of construction carries through the second-quantized case and the interacting case.

Second Quantization and Free Fields

We begin this section with a short review about the procedure of second quantization based on probabilistic methods, by following mainly Nelson (1973b); see also Guerra *et al.* (1975) and Simon (1974). Probabilistic methods are particularly useful in the framework of the Euclidean theory.

Let $\mathcal{H}$ be a real Hilbert space with symmetric scalar product $\langle , \rangle$. Let $\phi(u)$ be the elements of a family of centered Gaussian random variables indexed by $u \in \mathcal{H}$, uniquely defined by the expectation values $E(\phi(u))=0, E(\phi(u)\phi(v))=\langle u, v\rangle$. Since ϕ is Gaussian, we also have

$$E(\exp(\lambda\phi(u))) = \exp\left(\tfrac{1}{2}\lambda^2\langle u,u\rangle\right)$$

and

$$E(\phi(u_1)\phi(u_2)\cdots\phi(u_n)) = [u_1 u_2 \cdots u_n]$$

Here $[\ldots]$ is the Hafnian of elements $[u_i u_j]=\langle u_i, u_j\rangle$, defined to be zero for odd n, and for even n given by the recursive formula

$$[u_1 u_2 \cdots u_n] = \sum_{i=2}^{n} [u_1 u_i][u_1 u_2 \cdots u_n]'$$

where in $[\ldots]'$ the terms u_1 and u_i are suppressed. Hafnians, from the Latin name of Copenhagen, the first seat of the theoretical group of CERN, were introduced in quantum field theory by Caianiello (1973), as a useful tool when dealing with Bose statistics.

Let (Q, Σ, μ) be the underlying probability space where ϕ are defined as random variables. Here Q is a compact space, Σ a σ-algebra of subsets of Q, and μ a regular, countable additive probability measure on Σ, normalized to $\mu(Q)=\int_Q d\mu = 1$.

The fields $\phi(u)$ are represented by measurable functions on Q. The probability space is uniquely defined, but for trivial isomorphisms, if we assume that Σ is the smallest σ-algebra with respect to which all fields $\phi(u)$, with $u \in \mathcal{H}$, are measurable. Since $\phi(u)$ are Gaussian, they are represented by $L^p(Q, \Sigma, \mu)$ functions, for any p with $1 \leq p < \infty$, and the expectations will be given by

$$E(\phi(u_1)\phi(u_2)\cdots\phi(u_n)) = \int_Q \phi(u_1)\phi(u_2)\cdots\phi(u_n)\, d\mu$$

where, by a mild abuse of notation, $\phi(u_i)$ on the right-hand side denote the Q space functions which represent the random variables $\phi(u_i)$. We call the complex Hilbert space $\mathcal{F}=\Gamma(\mathcal{H})=L^2(Q,\Sigma,\mu)$ the Фок space constructed on $\mathcal{H}$, and the function $\Omega_0 \equiv 1$ on Q the Фок vacuum.

In order to introduce the concept of second quantization of operators, we must introduce subspaces of $\mathcal{F}$ with a "fixed number of particles." Call $\mathcal{F}_{(0)}=\{\lambda\Omega_0\}$, where λ is any complex number. Define $\mathcal{F}_{(\leq n)}$ as the subspace of $\mathcal{F}$ generated by complex linear combinations of monomials of the type $\phi(u_1)\cdots\phi(u_j)$, with $u_i \in \mathcal{H}$, and $j \leq n$. Then $\mathcal{F}_{(\leq n-1)}$ is a subspace of $\mathcal{F}_{(\leq n)}$. We define $\mathcal{F}_{(n)}$, the n-particle subspace, as the orthogonal complement of $\mathcal{F}_{(\leq n-1)}$ in $\mathcal{F}_{(\leq n)}$, so that

$$\mathcal{F}_{(\leq n)} = \mathcal{F}_{(n)} \oplus \mathcal{F}_{(\leq n-1)}$$

By construction, the $\mathcal{F}_{(n)}$ are orthogonal, and it is not difficult to verify that

$$\mathcal{F} = \bigoplus_{n=0}^{\infty} \mathcal{F}_{(n)}$$

Let us now introduce the Wick normal products by the definition

$$:\phi(u_1)\phi(u_2)\cdots\phi(u_n):=E_{(n)}\phi(u_1)\phi(u_2)\cdots\phi(u_n)$$

where $E_{(n)}$ is the projection on $\mathcal{F}_{(n)}$. It is not difficult to prove the usual Wick theorem (see, e.g., Guerra *et al.* (1975), and its inversion given by Caianiello (1973).

It is interesting to remark that, in the framework of the second quantization performed with probabilistic methods, it is not necessary to introduce creation and destruction operators as in the usual treatment. However, the two procedures are completely equivalent, as shown, for example, in Simon (1974).

Given an operator A from the real Hilbert space $\mathcal{H}_1$ to the real Hilbert space $\mathcal{H}_2$, we define its second-quantized operator $\Gamma(A)$ through the following definitions:

$$\Gamma(A)\Omega_{01}=\Omega_{02}$$
$$\Gamma(A)\,:\phi_1(u_1)\phi_1(u_2)\ldots\phi_1(u_n):$$
$$=:\phi_2(Au_1)\phi_2(Au_2)\cdots\phi_2(Au_n):$$

where we have introduced the probability spaces Q_1 and Q_2, their vacua Ω_{01} and Ω_{02}, and the random variables ϕ_1 and ϕ_2, associated to $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. The following remarkable theorem by Nelson (1973b) gives a full characterization of $\Gamma(A)$, very useful in the applications.

Theorem 3 *Let A be a contraction from the real Hilbert space $\mathcal{H}_1$ to the real Hilbert space $\mathcal{H}_2$. Then $\Gamma(A)$ is an operator from $L^1_{(1)}$ to $L^1_{(2)}$ which is positivity preserving, $\Gamma(A)u\geq 0$ if $u\geq 0$, and such that $E(\Gamma(A)u)=E(u)$. Moreover, $\Gamma(A)$ is a contraction from $L^p_{(1)}$ to $L^p_{(2)}$ for any p, $1\leq p<\infty$. Finally, $\Gamma(A)$ is also a contraction from $L^p_{(1)}$ to $L^q_{(2)}$, with $q\geq p$, if $\|A\|^2\leq(p-1)/(q-1)$.*

We have indicated with $L^p_{(1)}, L^p_{(2)}$ the L^p spaces associated to $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. This is the celebrated best hypercontractive estimate given by Nelson. For the proof, we refer to the original paper of Nelson (1973b); see also Simon (1974).

This completes our short review on the theory of second quantization based on probabilistic methods.

The usual time-zero quantum field $\bar{\phi}(u), u\in F_r$, in the Φοκ representation, can be obtained through second quantization starting from F_r. We call $(\bar{Q},\bar{\Sigma},\bar{\mu})$ the underlying probability space, and $\mathcal{F}=\Gamma(F_r)=L^2(\bar{Q},\bar{\Sigma},\bar{\mu})$ the Hilbert Φοκ space of the free physical particles.

Now we introduce the free Markov field $\phi(f), f\in N_r$, by taking N_r as the starting point. We call (Q,Σ,μ) the associated probability space. We introduce the Hilbert space $\mathcal{N}=\Gamma(N_r)=L^2(Q,\Sigma,\mu)$, and the operators $U(a,R)=\Gamma(u(a,R))$, $R_0=\Gamma(r_0)$, $U(t)=\Gamma(u(t))$, $E_A=\Gamma(e_A)$, and so on, for which the previous Nelson theorem holds (take $\mathcal{H}_1=\mathcal{H}_2=N_r$).

Since in general $\Gamma(AB)=\Gamma(A)\Gamma(B)$, we have immediately the following expression of the Markov property $E_\sigma=E_AE_B$, where the closed regions A,B,σ of the Euclidean space have the same properties as explained earlier in the proof of the (pre)Markov property for one-particle systems.

It is obvious that E_A can also be understood as conditional expectation with respect to the sub-σ-algebra Σ_A generated by the field $\phi(f)$ with $f\in N_r$ and the support of f on A.

The relation, previously pointed out, between N_t subspaces and F are also valid for their real parts N_{rt} and F_r. Therefore, they carry out through the second quantization procedure. We introduce $J_t=\Gamma(j_t)$ and $J_t^*=\Gamma(j_t^*)$; then the following properties hold. J_t is an isometric injection of $L^p(\bar{Q},\bar{\Sigma},\bar{\mu})$ into $L^p(Q,\Sigma,\mu)$; the range of J_t as an operator $L^2\to L^2$ is obviously $\mathcal{N}_t=\Gamma(N_{rt})$; moreover, $J_tJ_t^*=E_t$. The free Hamiltonian H_0 is given for $t\geq 0$ by

$$J_0^*J_t=\exp(-tH_0)=\Gamma(\exp(-t\omega))$$

Moreover, we have the covariance property $U(t)J_0=J_t$, and the reflexivity $R_0J_0=J_0, J_0^*R_0=J_0^*$.

These relations allow a very simple expression for the matrix elements of the Hamiltonian semigroup in terms of Markov quantities. In fact, for $u,v\in\mathcal{F}$ we have

$$\langle u,\exp(-tH_0)v\rangle=\int_Q (J_tu)^*J_0v\,d\mu$$

In the next section, we will generalize this representation to the interacting case.

Finally, let us derive the hypercontractive property of the free Hamiltonian semigroup.

Since $\|\exp(-t\omega)\|\leq\exp(-tm)$, where m is the mass of the particle, we have immediately, by a simple application of Nelson theorem,

$$\|\exp(-tH_0)\|_{p,q}\leq 1$$

provided $q-1\leq(p-1)\exp(2tm)$, where $\|\ldots\|_{p,q}$ denotes the norm of an operator from L^p to L^q spaces.

Interacting Fields

The discussion of the previous sections was limited to free fields both in Minkowski and Euclidean spaces. Now we must introduce interaction in order to get nontrivial theories.

First, as a general motivation, we will proceed quite formally and then we will resort to precise statements.

Let us recall that in standard quantum field theory, for scalar self-coupled fields, the time-ordered products of quantum fields in Minkowski spacetime can be expressed formally through the formula

$$\frac{\langle T(\phi(x_1)\cdots\phi(x_n)\exp(\mathrm{i}\int\mathcal{L}\,\mathrm{d}x))\rangle}{\langle T\exp(\mathrm{i}\int\mathcal{L}\,\mathrm{d}x)\rangle}$$

where T denotes time ordering, ϕ are free fields in Minkowski spacetime, $\mathcal{L}$ is the interaction Lagrangian, and $\langle\ldots\rangle$ are vacuum averages. As is well known, this expression can be put, for example, at the basis of perturbative expansions, giving rise to terms expressed through Feynman graphs. The appropriately chosen normalization provides automatic cancelation of the vacuum to vacuum graphs.

Now we can introduce a formal analytic continuation to the Schwinger points, as previously done for the one-particle system, and obtain the following expression for the analytic continuation of the field time-ordered products, now called Schwinger functions,

$$S(x_1,\ldots,x_n)=\frac{\langle\phi(x_1)\cdots\phi(x_n)\exp U\rangle}{\langle\exp U\rangle}$$

Here $x_1,\ldots,x_n$ denote points in Euclidean space, ϕ are the Euclidean fields introduced earlier. The chronological time ordering disappears, because the fields ϕ are commutative, and there is no distinguished "time" direction in Euclidean space. Here the symbol $\langle\ldots\rangle$ denotes the expectation values represented by $\int\ldots\mathrm{d}\mu$, as explained earlier, and U is the Euclidean "action" of the system formally given by the integral on Euclidean space

$$U=-\int P(\phi(x))\,\mathrm{d}x$$

if the field self-interaction is produced by the polynomial P.

Therefore, these formal considerations suggest that the passage from the free Euclidean theory to the fully interacting one is obtained through a change of the free probability measure $\mathrm{d}\mu$ to the interacting measure

$$\exp U\,\mathrm{d}\mu/\int_Q\exp U\,\mathrm{d}\mu$$

The analogy with classical statistical mechanics is evident. The expression exp U acts as *Boltzmannfaktor*, and $Z=\int_Q\exp U\,\mathrm{d}\mu$ is the partition function.

Our task will be to make these statements precise from a mathematical point of view. We will be obliged to introduce cutoffs, and then be involved in their careful removal.

For the sake of convenience, we make the substantial simplification of considering only two-dimensional theories (one space, one time dimension in the Minkowski region) for which the well-known ultraviolet problem of quantum field theory gives no trouble. There is no difficulty in translating the contents of the previous sections to the two-dimensional case.

Let P be a real polynomial, bounded below and normalized to $P(0)=0$. We introduce approximations h to the Dirac δ function at the origin of the two-dimensional Euclidean space R^2, with $h\in N_r$. Let h_x be the translate of h by x, with $x\in R^2$. The introduction of h, equivalent to some ultraviolet cutoff, is necessary, because local fields, of the formal type $\phi(x)$, have no rigorous meaning, and some smearing is necessary.

For some compact region Λ in R^2, acting as space cutoff (infrared cutoff), introduce the Q space function

$$U_\Lambda^{(h)}=-\int_\Lambda{:}P(\phi(h_x)){:}\,\mathrm{d}x$$

where $\mathrm{d}x$ is the Lebesgue measure in R^2. It is immediate to verify that $U_\Lambda^{(h)}$ is well defined, bounded below and belongs to $L^p(Q,\Sigma,\mu)$, for any p, $1\le p<\infty$. This is the infrared and ultraviolet cutoff action. Notice the presence of the Wick normal products in its definition. They provide a kind of automatic introduction of counterterms, in the framework of renormalization theory.

The following theorem allows us to remove the ultraviolet cutoff.

Theorem 4 *Let $h\to\delta$, in the sense that the Fourier transforms $\tilde{h}$ are uniformly bounded and converge pointwise in momentum space to the Fourier transform of the δ-function given by $(2\pi)^{-2}$. Then $U_\Lambda^{(h)}$ is L^p-convergent for any $p, 1\le p<\infty$, as $h\to\delta$. Call U_Λ the L^p-limit, then U_Λ, $\exp U_\Lambda\in L^p(Q,\Sigma_\Lambda,\mu)$, for $1\le p<\infty$.*

The proof uses standard methods of probability theory, and originates from pioneering work of Nelson in (1966). It can be found for example in Guerra *et al.* (1975), and Simon (1974).

Since U_Λ is defined with normal products, and the interaction polynomial P is normalized to $P(0)=0$, an elementary application of Jensen inequality gives

$$\int_Q\exp U_\Lambda\,\mathrm{d}\mu\ge\exp.\int_Q U_\Lambda\,\mathrm{d}\mu=1$$

Therefore, we can rigorously define the new space cutoff measure in Q space:

$$d\mu_\Lambda = \exp U_\Lambda \, d\mu / \int_Q \exp U_\Lambda \, d\mu$$

The space-cutoff interacting Euclidean theory is defined by the same fields on Q space, but with a change in the measure and, therefore, in the expectation values. The correlations for the interacting fields $\bar{\phi}$ are the cutoff Schwinger functions

$$\begin{aligned} S_\Lambda(x_1, \ldots, x_n) &= \langle \bar{\phi}(x_1) \cdots \bar{\phi}(x_n) \rangle \\ &= Z_\Lambda^{-1} \langle \phi(x_1) \cdots \phi(x_n) \exp U_\Lambda \rangle \end{aligned}$$

where the partition function is

$$Z_\Lambda = \langle \exp U_\Lambda \rangle$$

We see that the analogy with statistical mechanics is complete here. Of course, the introduction of the space cutoff Λ destroys translation invariance. The full Euclidean covariant theory must be recovered by taking the infinite-volume limit $\Lambda \to R^2$ on field correlations. For the removal of the space cutoff, all methods of statistical mechanics are available. In particular, correlation inequalities of ferromagnetic type can be easily exploited, as shown, for example, in Guerra *et al.* (1975) and Simon (1974).

We would like to conclude this section by giving the connection between the space-cutoff Euclidean theory and the space-cutoff Hamiltonian theory in the physical ΦOK space.

For $\ell \geq 0, t \geq 0$, consider the rectangle in R^2,

$$\Lambda(\ell, t) = \left\{ (x_1, x_2) \colon -\frac{\ell}{2} \leq x_1 \leq \frac{\ell}{2}, 0 \leq x_2 \leq t \right\}$$

and define the operator in the physical ΦOK space

$$P_\ell(t) = J_0^* \exp U_\Lambda(\ell, t) J_t$$

where J_0 and J_t are injections relative to the lines $x_2 = 0$ and $x_2 = t$, respectively. Then the following theorem, largely due to Nelson, holds.

Theorem 5 *The operator $P_\ell(t)$ is bounded and self-adjoint. The family $\{P_\ell(t)\}$, for ℓ fixed and $t \geq 0$, is a strongly continuous semigroup. Let H_ℓ be its lower bounded self-adjoint generator, so that $P_\ell(t) = \exp(-tH_\ell)$. On the physical ΦOK space, there is a core $\mathcal{D}$ for H_ℓ such that on $\mathcal{D}$ the equality $H_\ell = H_0 + V_\ell$ holds, where H_0 is the free Hamiltonian introduced earlier and V_ℓ is the volume-cutoff interaction given by*

$$V_\ell = \lim \int_{-\ell/2}^{\ell/2} : P(\bar{\phi}(h_{x_1})) : \, dx_1$$

where h_{x_1} are the translates of approximations to the δ-function at the origin on the x_1-space, and the limit is taken in L^p, in analogy to what has been explained for the two-dimensional case in the definition of U_Λ.

While we refer to Guerra *et al.* (1975) and Simon (1974) for a full proof, we mention here that boundedness is related to hypercontractivity of the free Hamiltonian, self-adjointness is a consequence of reflexivity, and the semigroup property follows from Markov property. This theorem is remarkable, because it expresses the cutoff interacting Hamiltonian semigroup in an explicit form in the Euclidean theory through probabilistic expectations. In fact, we have

$$\langle u, \exp(-tH_\ell) v \rangle = \int_Q (J_t u)^* J_0 v \exp U_\Lambda(\ell, t) \, d\mu$$

We could call this expression as the Feynman–Kac–Nelson formula, in fact it is nothing but a path integral expressed in stochastic terms, and adapted to the Hamiltonian semigroup.

By comparison with the analogous formula given for the free Hamiltonian semigroup, we see that the introduction of the interaction inserts the *Boltzmannfaktor* under the integral.

As an immediate consequence of the Feynman–Kac–Nelson formula, together with Euclidean covariance, we have the following astonishing Nelson symmetry:

$$\langle \Omega_0, \exp(-tH_\ell) \Omega_0 \rangle = \langle \Omega_0, \exp(-\ell H_t) \Omega_0 \rangle$$

which was at the basis of Guerra (1972) and Guerra *et al.* (1972), and played some role in showing the effectiveness of Euclidean methods in constructive quantum field theory.

It is easy to establish, through simple probabilistic reasoning, that H_ℓ has a unique ground state Ω_ℓ of lowest energy E_ℓ. For a convenient choice of normalization and phase factor, one has $\|\Omega_\ell\|_2 = 1$, and $\Omega_\ell > 0$ almost everywhere on Q space (for bosonic systems, ground states have no nodes in configuration space!). Moreover, $\Omega_\ell \in L^p$, for any $1 \leq p < \infty$. If $\ell > 0$ and the interaction is not trivial, then $\Omega_\ell \neq \Omega_0, E_\ell < 0$, and $\|\Omega_\ell\|_1 < 1$. Obviously, $\|\exp(-tH_\ell)\|_{2,2} = \exp(-tE_\ell)$.

The general structure of Euclidean field theory, as explained in this section, has been at the basis of all applications in constructive quantum field theory. These applications include the proof of the existence of the infinite-volume limit, with the establishment of all Wightman axioms, for two- and three-dimensional theories. Moreover, the existence of phase transitions and symmetry breaking has been firmly established.

Extensions have also been given to theories involving Fermions, and to gauge field theory. Due to the scope of this review, limited to a description of the general structure of Euclidean field theory, we cannot give a detailed treatment of these applications. Therefore, we refer to recent general reviews on constructive quantum field theory for a complete description of all results (see, e.g., Jaffe (2000)). For recent applications of Euclidean field theory to quantum fields on curved spacetime manifolds we refer, for example, to Schlingemann (1999).

The Physical Interpretation of Euclidean Field Theory

Euclidean field theory has been considered by most researchers as a very useful tool for the study of quantum field theory. In particular, it is quite easy, for example, to obtain the fully interacting Schwinger functions in the infinite-volume limit in two-dimensional spacetime. At this point, there arises the problem of connecting these Schwinger functions with observable physical quantities in Minkowski spacetime. A very deep result of Osterwalder and Schrader (1973) gives a very natural interpretation of the resulting limiting theory. In fact, the Euclidean theory, as has been shown earlier, arises from an analytic continuation from the physical Minkowski spacetime to the Schwinger points, through a kind of analytic continuation in time (also called Wick rotation, because Wick exploited this trick in the study of the Bethe–Salpeter equation). Therefore, having obtained the Schwinger functions for the full covariant theory, after all cutoff removal, it is very natural to try to reproduce the inverse analytic continuation in order to recover the Wightman functions in Minkowski spacetime. Therefore, Osterwalder and Schrader have been able to identify a set of conditions, quite easy to verify, wich allow us to recover Wightman functions from Schwinger functions. A key role in this reconstruction theorem is played by the so-called reflection positivity for Schwinger functions, a property quite easy to verify. In this way, a fully satisfactory solution for the physical interpretation of Euclidean field theory is achieved.

From a historical point of view, an alternate route is possible. In fact, at the beginning of the exploitation of Euclidean methods in constructive quantum field theory, Nelson was able to isolate a set of axioms for the Euclidean *fields* (Nelson 1973a), allowing the reconstruction of the physical theory. Of course, Nelson axioms are more difficult to verify, since they also involve properties of the Euclidean fields and not only of the Schwinger functions. However, it is still very interesting to investigate whether the Euclidean fields play only an auxiliary role in the construction of the physical content of relativistic theories, or if they have a more fundamental meaning.

From a physical point of view, the following considerations could also lead to further developments along this line. By its very structure, the Euclidean theory contains the fixed-time quantum correlations in the vacuum. In elementary quantum mechanics, it is possible to derive all physical content of the theory from the simple knowledge of the ground state wave function, including scattering data. Therefore, at least in principle, it should be possible to derive all physical content of the theory directly from the Euclidean theory, without any analytic continuation.

We conclude this short section on the physical interpretation of the Euclidean theory with a mention of a quite surprising result (Guerra and Ruggiero 1973) obtained by submitting classical field theory to the procedure of stochastic quantization in the sense of Nelson (1985). The procedure of stochastic quantization associates a stochastic process to each quantum state. In this case, in a fixed reference frame, the procedure of stochastic quantization, applied to interacting fields, produces, for the ground state, a process in the physical spacetime that has the same correlations as Euclidean field theory. This opens the way to a possible interpretation of Euclidean field theory directly in Minkowski spacetime. However, a consistent development along this line requires a new formulation of representations of the Poincaré group in the form of measure-preserving transformations in the probability space where the Euclidean fields are defined. This difficult task has not been accomplished as yet.

Acknowledgment

Research connected with this work was supported in part by MIUR (Italian Minister of Instruction, University and Research), and by INFN (Italian National Institute for Nuclear Physics).

See also: Axiomatic Quantum Field Theory; Constructive Quantum Field Theory; Feynman Path Integrals; Functional Integration in Quantum Physics; High T_c Superconductor Theory; Malliavin Calculus; Quantum Chromodynamics; Quantum Field Theory: A Brief Introduction; Quantum Fields with Indefinite Metric: Non-Trivial Models; Relativistic Wave Equations including Higher Spin Fields; Renormalization: General Theory; Two-dimensional Models.

Further Reading

Caianiello ER (1973) *Combinatorics and Renormalization in Quantum Field Theory*. Reading, MA: Benjamin.

Glimm J and Jaffe A (1981) *Quantum Physics, A Functional Integral Point of View*. Berlin: Springer.

Guerra F (1972) Uniqueness of the vacuum energy density and van Hove phenomenon in the infinite volume limit for two-dimensional self-coupled Bose fields. *Physical Review Letters* 28: 1213.

Guerra F, Rosen L, and Simon B (1972) Nelson's symmetry and the infinite volume behavior of the vacuum in $P(\phi)_2$. *Communication in Mathematical Physics* 27: 10.

Guerra F, Rosen L, and Simon B (1975) The $P(\phi)_2$ Euclidean quantum field theory as classical statistical mechanics. *Annuals of Mathematics* 101: 111.

Guerra F and Ruggiero P (1973) New interpretation of the Euclidean Markov field in the framework of physical Minkowski space-time. *Physical Review Letters* 31: 1022.

Jaffe A (2000) *Constructive quantum field theory*, available on http://www.arthurjaffe.com.

Nakano T (1959) Quantum field theory in terms of Euclidean parameters. *Progress in Theoretical Physics* 21: 241.

Nelson E (1966) A quartic interaction in two dimensions. In: Goodman R and Segal I (eds.) *Conference on the Mathematical Theory of Elementary Particles*. Cambridge, MA: MIT Press.

Nelson E (1973a) Construction of quantum fields from Markoff fields. *Journal of Functional Analysis* 12: 97.

Nelson E (1973b) The free Markoff field. *Journal of Functional Analysis* 12: 211.

Nelson E (1985) *Quantum Fluctuations*. Princeton, NJ: Princeton University Press.

Osterwalder K and Schrader R (1973) Axioms for Euclidean Green's functions. *Communication in Mathematical Physics* 31: 83.

Schlingemann D (1999) *Euclidean field theory on a sphere*, available on http://arXiv.org.

Schwinger J (1958) On the Euclidean structure of relativistic field theory. *Proceedings of the National Academy Sciences* 44: 956.

Simon B (1974) *The $P(\phi)_2$ Euclidean (Quantum) Field Theory*. Princeton, NJ: Princeton University Press.

Streater R and Wightman AS (1964) *PCT, Spin and Statistics and All That*. New York: Benjamin.

Symanzik K (1966) Euclidean quantum field theory, I, Equations for a scalar model. *Journal of Mathematical Physics* 7: 510.

Symanzik K (1969) Euclidean quantum field theory. In: Jost R (ed.) *Local Quantum Theory*. New York: Academic Press.

Wightman AS (1967) An introduction to some aspects of the relativistic dynamics of quantized fields. In: Levy M (ed.) *1964 Cargèse Summer School Lectures*. New York: Gordon and Breach.

Gerbes in Quantum Field Theory

J Mickelsson, KTH Physics, Stockholm, Sweden

© 2006 Elsevier Ltd. All rights reserved.

Definitions and an Example

A gerbe can be viewed as a next step in a ladder of geometric and topological objects on a manifold which starts from ordinary complex-valued functions and in the second step of sections of complex line bundles.

It is useful to recall the construction of complex line bundles and their connections. Let M be a smooth manifold and $\{U_\alpha\}$ an of open cover of M which trivializes a line bundle L over M. Topologically, up to equivalence, the line bundle is completely determined by its Chern class, which is a cohomology class $[c] \in H^2(M, \mathbb{Z})$. On each open set U_α we may write $2\pi i c = \mathrm{d}A_\alpha$, where A_α is a 1-form. On the overlaps $U_{\alpha\beta} = U_\alpha \cap U_\beta$ we can write

$$A_\alpha - A_\beta = f_{\alpha\beta}^{-1} \mathrm{d} f_{\alpha\beta} \qquad [1]$$

at least when $U_{\alpha\beta}$ is contractible, where $f_{\alpha\beta}$ is a circle-valued complex function on the overlap. The data $\{c, A_\alpha, f_{\alpha\beta}\}$ define what is known as a (representative of a) Deligne cohomology class on the open cover $\{U_\alpha\}$. The 1-forms A_α are the local potentials of the curvature form $2\pi i c$ and the $f_{\alpha\beta}$'s are the transition functions of the line bundle L. Each of these three different data defines separately the equivalence class of the line bundle but together they define the line bundle with a connection.

The essential thing here is that there is a bijection between the second integral cohomology of M and the set of equivalence classes of complex line bundles over M. It is natural to ask whether there is a geometric realization of integral third (or higher) cohomology. In fact, gerbes provide such a realization. Here, we shall restrict to a smooth differential geometric approach which by no means is the most general possible, but it is sufficient for most applications to quantum field theory. However, there are examples of gerbes over orbifolds that do not need to come from finite group action on a manifold, which are not covered by the following definition.

For the examples in this article, it is sufficient to adapt the following definition. A gerbe over a manifold M (without geometry) is simply a principal bundle $\pi: P \to M$ with fiber equal to $PU(H)$, the projective unitary group of a Hilbert space H. The Hilbert space may be either finite or infinite dimensional.

The quantum field theory applications discussed in this article are related to the chiral anomaly for

Notes:

fermions in external fields. The link comes from the fact that the chiral symmetry breaking leads in the generic case to projective representations of the symmetry groups. For this reason, when modding out by the gauge or diffeomorphism symmetries, one is led to study bundles of projective Hilbert spaces. The anomaly is reflected as a nontrivial characteristic class of the projective bundle, known in mathematics literature as the Dixmier–Douady class.

In a suitable open cover, the bundle P has a family of local trivializations with transition functions $g_{\alpha\beta}: U_{\alpha\beta} \to PU(H)$, with the usual cocycle property

$$g_{\alpha\beta}g_{\beta\gamma}g_{\gamma\alpha} = 1 \qquad [2]$$

on triple overlaps. Assuming that the overlaps are contractible, we can choose lifts $\hat{g}_{\alpha\beta}: U_{\alpha\beta} \to U(H)$, to the unitary group of the Hilbert space. However,

$$\hat{g}_{\alpha\beta}\hat{g}_{\beta\gamma}\hat{g}_{\gamma\alpha} = f_{\alpha\beta\gamma} \qquad [3]$$

where the f's are circle-valued functions on triple overlaps. They satisfy automatically the cocycle property

$$f_{\alpha\beta\gamma}f^{-1}_{\alpha\beta\delta}f_{\alpha\gamma\delta}f^{-1}_{\beta\gamma\delta} = 1 \qquad [4]$$

on quadruple overlaps. There is an important difference between the finite- and infinite-dimensional cases. In the finite-dimensional case, the circle bundle $U(H) \to U(H)/S^1 = PU(H)$ reduces to a bundle with fiber $\mathbb{Z}/N\mathbb{Z} = \mathbb{Z}_N$, where $N = \dim H$. This follows from $U(N)/S^1 = SU(N)/Z_N$ and the fact that $SU(N)$ is a subgroup of $U(N)$. For this reason one can choose the lifts $\hat{g}_{\alpha\beta}$ such that the functions $f_{\alpha\beta\gamma}$ take values in the finite subgroup $\mathbb{Z}_N \subset S^1$.

The functions $f_{\alpha\beta\gamma}$ define an element $a = \{a_{\alpha\beta\gamma\delta}\}$ in the Čech cohomology $H^3(\mathcal{U}, \mathbb{Z})$ by a choice of logarithms,

$$2\pi i a_{\alpha\beta\gamma\delta} = \log f_{\alpha\beta\gamma} - \log f_{\alpha\beta\delta} + \log f_{\alpha\gamma\delta} - \log f_{\beta\gamma\delta} \qquad [5]$$

In the finite-dimensional case, the Čech cocycle is necessarily torsion, $Na = 0$, but not so if H is infinite dimensional. In the finite-dimensional case (by passing to a good cover and using the Čech – de Rham equivalence over real or complex numbers), the class is third de Rham cohomology constructed from the transition functions is necessarily zero. Thus, in general one has to work with Čech cohomology to preserve torsion information. One can prove:

Theorem *The construction above is a one-to-one map between the set of equivalence classes of $PU(H)$ bundles over M and elements of $H^3(M, \mathbb{Z})$.*

The characteristic class in $H^3(M, \mathbb{Z})$ of a $PU(H)$ bundle is called the *Dixmier–Douady class*.

First example

Let M be an oriented Riemannian manifold and FM its bundle of oriented orthonormal frames. The structure group of FM is the rotation group $SO(n)$ with $n = \dim M$. The spin bundle (when it exists) is a double covering $\mathrm{Spin}(M)$ of FM, with structure group $\mathrm{Spin}(n)$, a double cover of $SO(n)$. Even when the spin bundle does not exist there is always the bundle $\mathrm{Cl}(M)$ of Clifford algebras over M. The fiber at $x \in M$ is the Clifford algebra defined by the metric g_x, that is, it is the complex 2^n-dimensional algebra generated by the tangent vectors $v \in T_x(m)$ with the defining relations

$$\gamma(u)\gamma(v) + \gamma(v)\gamma(u) = 2g_x(u, v)$$

The Clifford algebra has a faithful representation in $N = 2^{[n/2]}$ dimensions ($[x]$ is the integral part of x) such that

$$\gamma(a \cdot u) = S(a)\gamma(u)S(a)^{-1}$$

where S is an unitary representation of $\mathrm{Spin}(n)$ in $\mathbb{C}^N$. Since $\mathrm{Spin}(n)$ is a double cover of $SO(n)$, the representation S may be viewed as a projective representation of $SO(n)$. Thus again, if the overlaps $U_{\alpha\beta}$ are contractible, we may choose a lift of the frame bundle transition functions $g_{\alpha\beta}$ to unitaries $\hat{g}_{\alpha\beta}$ in $H = \mathbb{C}^N$. In this case, the functions $f_{\alpha\beta\gamma}$ reduce to $\mathbb{Z}_2$-valued functions, and the obstruction to the lifting problem, which is the same as the obstruction to the existence of spin structure, is an element of $H^2(M, \mathbb{Z}_2)$, known as the second Stiefel–Whitney class w_2. The image of w_2 with respect to the Bockstein map (in this case, given by the formula [5]) gives a 2-torsion element in $H^3(M, \mathbb{Z})$, the Dixmier–Douady class.

Another way to think of a gerbe is the following (we shall see that this arises in a natural way in quantum field theory). There is a canonical complex line bundle L over $PU(H)$, the associated line bundle to the circle bundle $S^1 \to U(H) \to PU(H)$. Pulling back L by the local transition functions $g_{\alpha\beta} \to PU(H)$, we obtain a family of line bundles $L_{\alpha\beta}$ over the open sets $U_{\alpha\beta}$. By the cocycle property [2] we have natural isomorphisms

$$L_{\alpha\beta} \otimes L_{\beta\gamma} = L_{\alpha\gamma} \qquad [6]$$

We can take this as a definition of a gerbe over M: a collection of line bundles over intersections of open sets in an open cover of M, satisfying the cocycle condition [6]. By [6] we have a trivialization

$$L_{\alpha\beta} \otimes L_{\beta\gamma} \otimes L_{\gamma\alpha} = f_{\alpha\beta\gamma} \cdot 1 \qquad [7]$$

where the f's are circle-valued functions on the triple overlaps. By the theorem above, we conclude

that indeed the data in [6] define (an equivalence class of) a principal $PU(H)$ bundle.

If $L_{\alpha\beta}$ and $L'_{\alpha\beta}$ are two systems of local line bundles over the same cover, then the gerbes are equivalent if there is a system of line bundles L_α over open sets U_α such that

$$L'_{\alpha\beta} = L_{\alpha\beta} \otimes L^*_\alpha \otimes L_\beta \qquad [8]$$

on each $U_{\alpha\beta}$.

A gerbe may come equipped with geometry, encoded in a Deligne cohomology class with respect to a given open covering of M. The Deligne class is given by functions $f_{\alpha\beta\gamma}$, 1-forms $A_{\alpha\beta}$, 2-forms F_α, and a global 3-form (the Dixmier–Douady class of the gerbe) Ω, subject to the conditions

$$\begin{aligned} \mathrm{d}F_\alpha &= 2\pi\mathrm{i}\Omega \\ F_\alpha - F_\beta &= \mathrm{d}A_{\alpha\beta} \\ A_{\alpha\beta} - A_{\alpha\gamma} + A_{\beta\gamma} &= f^{-1}_{\alpha\beta\gamma}\mathrm{d}f_{\alpha\beta\gamma} \end{aligned} \qquad [9]$$

Gerbes from Canonical Quantization

Let D_x be a family of self-adjoint Fredholm operators in a complex Hilbert space H parametrized by $x \in M$. This situation arises in quantum field theory, for example, when M is some space of external fields, coupled to Dirac operator D on a compact manifold. The space M might consist of gauge potentials (modulo gauge transformations) or M might be the moduli space of Riemann metrics. In these examples, the essential spectrum of D_x is both positive and negative and the family D_x defines an element of $K^1(M)$. In fact, one of the definitions of $K^1(M)$ is that its elements are homotopy classes of maps from M to the space $\mathcal{F}_*$ of self-adjoint Fredholm operators with both positive and negative essential spectrum. In physics applications, one deals most often with unbounded Hamiltonians, and the operator norm topology must be replaced by something else; popular choices are the Riesz topology defined by the map $F \mapsto F/(|F|+1)$ to bounded operators or the gap topology defined by graph metric.

The space $\mathcal{F}_*$ is homotopy equivalent to the group $G = U_1(H)$ of unitary operators g in H such that $g-1$ is a trace-class operator. This space is a classifying space for principal U_{res} bundles, where U_{res} is the group of unitary operators g in a polarized complex Hilbert space $\mathcal{H} = \mathcal{H}_+ \oplus \mathcal{H}_-$ such that the off-diagonal blocks of g are Hilbert–Schmidt operators. This is related to Bott periodicity. There is a natural principal bundle P over $G = U_1(H)$ with fiber equal to the group ΩG of based loops in G. The total space P consists of smooth paths $f(t)$ in G starting from the neutral element such that $f^{-1}\mathrm{d}f$ is smooth and periodic. The projection $P \to G$ is the evaluation at the end point $f(1)$. The fiber is clearly ΩG. By Bott periodicity, the homotopy groups of ΩG are shifted from those of G by one dimension, that is,

$$\pi_n \Omega G = \pi_{n+1} G$$

The latter are zero in even dimensions and equal to $\mathbb{Z}$ in odd dimensions. On the other hand, it is known that the even homotopy groups of $U_{\mathrm{res}}(\mathcal{H})$ are equal to $\mathbb{Z}$ and the odd ones vanish. In fact, with a little more effort, one can show that the embedding of ΩG to $U_{\mathrm{res}}(\mathcal{H})$ is a homotopy equivalence, when $\mathcal{H} = L^2(S^1, H)$, the polarization being the splitting to non-negative and negative Fourier modes and the action of ΩG is the pointwise multiplication on H-valued functions on the circle S^1.

Since P is contractible, it is indeed the classifying bundle for U_{res} bundles. Thus, we conclude that "$K^1(M)$ = the set of homotopy classes of maps $M \to G$ = the set of equivalence classes of U_{res} bundles over M." The relevance of this fact in quantum field theory follows from the properties of representations of the algebra of canonical anti-commutation relations (*CAR*). For any complex Hilbert space H, this algebra is the algebra generated by elements $a(v)$ and $a^*(v)$, with $v \in H$, subject to the relations

$$a^*(u)a(v) + a(v)a^*(u) = 2 < v, u >$$

where the Hilbert space inner product on the right-hand side is antilinear in the first argument, and all other anticommutators vanish. In addition, $a^*(u)$ is linear and $a(v)$ antilinear in its argument.

An irreducible Dirac representation of the CAR algebra is given by a polarization $H = H_+ \oplus H_-$. The representation is characterized by the existence of a vacuum vector ψ in the fermionic Fock space $\mathcal{F}$ such that

$$a^*(u)\psi = 0 = a(v)\psi \quad \text{for } u \in H_-,\ v \in H_+ \qquad [10]$$

A theorem of D Shale and W F Stinespring says that two Dirac representations defined by a pair of polarizations H_+, H'_+ are equivalent if and only if there is $g \in U_{\mathrm{res}}(H_+ \oplus H_-)$ such that $H'_+ = g \cdot H_+$. In addition, in order that a unitary transformation g is implementable in the Fock space, that is, there is a unitary operator $\hat{g}$ in $\mathcal{F}$ such that

$$\hat{g}a^*(v)\hat{g}^{-1} = a^*(gv), \quad \forall v \in H \qquad [11]$$

and similarly for the $a(v)$'s, one must have $g \in U_{\mathrm{res}}$ with respect to the polarization defining the vacuum vector. This condition is both necessary and sufficient.

The polarization of the one-particle Hilbert space comes normally from a spectral projection onto the positive-energy subspace of a Hamilton operator. In the background field problems one studies families of Hamilton operators D_x and then one would like to construct a family of fermionic Fock spaces parametrized by $x \in M$. If none of the Hamilton operators has zero modes, this is unproblematic. However, the presence of zero modes makes it impossible to define the positive-energy subspace $H_+(x)$ as a continuous function of x. One way out of this is to weaken the condition for the polarization: each $x \in M$ defines a Grassmann manifold $\mathrm{Gr}_{\mathrm{res}}(x)$ consisting of all subspaces $W \subset H$ such that the projections onto W and $H_+(x)$ differ by Hilbert–Schmidt operators. The definition of $\mathrm{Gr}_{\mathrm{res}}(x)$ is stable with respect to finite-rank perturbations of $D_x/|D_x|$. For example, when D_x is a Dirac operator on a compact manifold then $(D_x - \lambda)/|D_x - \lambda|$ defines the same Grassmannian for all real numbers λ because in each finite interval there are only a finite number of eigenvalues (with multiplicities) of D_x. From this follows that the Grassmannians form a locally trivial fiber bundle Gr over families of Dirac operators.

If the bundle Gr has a global section $x \mapsto W_x$ then we can define a bundle of Fock space representations for the CAR algebra over the parameter space M. However, there are important situations when no global sections exist. It is easier to explain the potential obstruction in terms of a principal U_{res} bundle P such that Gr is an associated bundle to P.

The fiber of P at $x \in M$ is the set of all unitaries g in H such that $g \cdot H_+ \in \mathrm{Gr}_x$ where $H = H_+ \oplus H_-$ is a fixed reference polarization. Then we have

$$\mathrm{Gr} = P \times_{U_{\mathrm{res}}} \mathrm{Gr}_{\mathrm{res}},$$

where the right action of $U_{\mathrm{res}} = U_{\mathrm{res}}(H_+ \oplus H_-)$ in the fibers of P is the right multiplication on unitary operators and the left action on $\mathrm{Gr}_{\mathrm{res}}$ comes from the observation that $\mathrm{Gr}_{\mathrm{res}} = U_{\mathrm{res}}/(U_+ \times U_-)$, where $U_\pm$ are the diagonal block matrices in U_{res}. By a result of N Kuiper, the subgroup $U_- \times U_-$ is contractible and so Gr has a global section if and only if P is trivial.

Thus, when P is trivial we can define the family of Dirac representations of the CAR algebra parametrized by M such that in each of the Fock spaces we have a Dirac vacuum which, in a precise sense, is close to the vacuum defined by the energy polarization. However, the triviality of P is not a necessary condition. Actually, what is needed is that P has a prolongation to a bundle $\hat{P}$ with fiber $\hat{U}_{\mathrm{res}}$. The group $\hat{U}_{\mathrm{res}}$ is a central extension of U_{res} by the group S^1.

The Lie algebra $\hat{\boldsymbol{u}}_{\mathrm{res}}$ is as a vector space the direct sum $\boldsymbol{u}_{\mathrm{res}} \oplus \mathrm{i}\mathbb{R}$, with commutators

$$[X + \lambda, Y + \mu] = [X, Y] + c(X, Y) \qquad [12]$$

where c is the Lie algebra cocycle

$$c(X, Y) = \tfrac{1}{4} \mathrm{tr}\, \epsilon [\epsilon, X][\epsilon, Y] \qquad [13]$$

Here ϵ is the grading operator with eigenvalues ± 1 on $H_\pm$. The trace exists since the off diagonal blocks of X, Y are Hilbert–Schmidt.

The group $\hat{U}_{\mathrm{res}}$ is a circle bundle over U_{res}. The Chern class of the associated complex line bundle is the generator of $H^2(U_{\mathrm{res}}, \mathbb{Z})$ and is given explicitly at the identity element as the antisymmetric bilinear form $c/2\pi\mathrm{i}$ and at other points on the group manifold through left-translation of $c/2\pi\mathrm{i}$. If P is trivial, then it has an obvious prolongation to the trivial bundle $M \times \hat{U}_{\mathrm{res}}$. In any case, if the prolongation exists we can define the bundle of Fock spaces carrying CAR representations as the associated bundle

$$\mathcal{F} = \hat{P} \times_{\hat{U}_{\mathrm{res}}} \mathcal{F}_0$$

where is $\mathcal{F}_0$ is the fixed Fock space defined by the same polarization $H = H_+ \oplus H_-$ used to define U_{res}. By the Shale–Stinespring theorem, any $g \in U_{\mathrm{res}}$ has an implementation $\hat{g}$ in $\mathcal{F}_0$, but $\hat{g}$ is only defined up to phase, thus the central S^1 extension.

The action of the CAR algebra in the fibers is given as follows. For $x \in M$ choose any $\hat{g} \in \hat{P}_x$. Define

$$a^*(v) \cdot (\hat{g}, \psi) = (\hat{g}, a^*(g^{-1}v)\psi)$$

where $\psi \in \mathcal{F}_0$ and $v \in H$; similarly for the operators $a(v)$. It is easy to check that this definition passes to the equivalence classes in $\mathcal{F}$. Note that the representations in different fibers are in general inequivalent because the tranformation g is not implementable in the Fock space $\mathcal{F}_0$.

The potential obstruction to the existence of the prolongation of P is again a 3-cohomology class on the base. Choose a good cover of M. On the intersections $U_{\alpha\beta}$ of the open cover the transition functions $g_{\alpha\beta}$ of P can be prolonged to functions $\hat{g}_{\alpha\beta} : U_{\alpha\beta} \to \hat{U}_{\mathrm{res}}$. We have

$$\hat{g}_{\alpha\beta}\hat{g}_{\beta\gamma}\hat{g}_{\gamma\alpha} = f_{\alpha\beta\gamma} \cdot 1 \qquad [14]$$

for functions $f_{\alpha\beta\gamma} : U_{\alpha\beta\gamma} \to S^1$, which by construction satisfy the cocycle property [4]. Since the cocycle is defined on a good cover, it defines an integral Čech cohomology class $\omega \in H^3(M, \mathbb{Z})$.

Let us return to the universal U_{res} bundle P over $G = U_1(H)$. In this case the prolongation obstruction can be computed relatively easily. It turns out that

the 3-cohomology class is represented by the de Rham class which is the generator of $H^3(G,\mathbb{Z})$. Explicitly,

$$\omega = \frac{1}{24\pi^2}\mathrm{tr}\,(g^{-1}\mathrm{d}g)^3 \qquad [15]$$

Any principal U_{res} bundle over M comes from a pullback of P with respect to a map $f: M \to G$, so the Dixmier–Douady class in the general case is the pullback $f^*\omega$.

The line bundle construction of the gerbe over the parameter space M for Dirac operators is given by the observation that the spectral subspaces $E_{\lambda\lambda'}(x)$ of D_x, corresponding to the open interval $]\lambda,\lambda'[$ in the real line, form finite-rank vector bundles over open sets $U_{\lambda\lambda'} = U_\lambda \cap U_{\lambda'}$. Here U_λ is the set of points $x \in M$ such that λ does not belong to the spectrum of D_x. Then we can define, as top exterior power,

$$L_{\lambda\lambda'} = \bigwedge^{\mathrm{top}}(E_{\lambda\lambda'})$$

as the complex vector bundle over $U_{\lambda\lambda'}$. It follows immediately from the definition that the cocycle property [6] is satisfied.

Example 1 (Fermions on an interval). Let K be a compact group and ρ its unitary representation in a finite-dimensional vector space V. Let H be the Hilbert space of square-integrable V-valued functions on the interval $[0,2\pi]$ of the real axis. For each $g \in K$ let $\mathrm{Dom}_g \subset H$ be the dense subspace of smooth functions ψ with the boundary condition $\psi(2\pi) = \rho(g)\psi(0)$. Denote by D_g the operator $-\mathrm{i}\,\mathrm{d}/\mathrm{d}x$ on this domain. The spectrum of D_g is a function of the eigenvalues λ_k of $\rho(g)$, consisting of real numbers $n + \log(\lambda_k)/2\pi\mathrm{i}$ with $n \in \mathbb{Z}$. For this reason the splitting of the one-particle space H to positive and negative modes of the operator D_g is in general not continuous as function of the parameter g. This leads to the problems described above. However, the principal U_{res} bundle can be explicitly constructed. It is the pullback of the universal bundle P with respect to the map $f: K \to G$ defined by the embedding $\rho(K) \subset G$ as $N \times N$ block matrices, $N = \dim V$. Thus, the Dixmier–Douady class in this example is

$$\omega = \frac{1}{24\pi^2}\mathrm{tr}\,(\rho(g)^{-1}\,\mathrm{d}\rho(g))^3 \qquad [16]$$

Example 2 (Fermions on a circle). Let $H = L^2(S^1, V)$ and $D_A = -\mathrm{i}(\mathrm{d}/\mathrm{d}x + A)$ where A is a smooth vector potential on the circle taking values in the Lie algebra $\boldsymbol{k}$ of K. In this case, the domain is fixed, consisting of smooth V-valued functions on the circle. The $\boldsymbol{k}$-valued function A is represented as a multiplication operator through the representation ρ of K. The parameter space $\mathcal{A}$ of smooth vector potentials is flat; thus, there cannot be any obstruction to the prolongation problem. However, in quantum field theory, one wants to pass to the moduli space $\mathcal{A}/\mathcal{G}$ of gauge potentials. Here $\mathcal{G}$ is the group of smooth based gauge transformations, that is, $\mathcal{G} = \Omega K$. Now the moduli space is the group of holonomies around the circle, $\mathcal{A}/\mathcal{G} = K$. Thus, we are in a similar situation as in Example 1. In fact, these examples are really two different realizations of the same family of self-adjoint Fredholm operators. The operator D_A with $k = \mathrm{holonomy}(A)$ has exactly the same spectrum as D_k in Example 1. For this reason, the Dixmier–Douady class on K is the same as before.

The case of Dirac operators on the circle is simple because all the energy polarizations for different vector potentials are elements in a single Hilbert–Schmidt Grassmannian $\mathrm{Gr}(H_+ \oplus H_-)$, where we can take as the reference polarization the splitting to positive and negative Fourier modes. Using this polarization, the bundle of fermionic Fock spaces over $\mathcal{A}$ can be trivialized as $\mathcal{F} = \mathcal{A} \times \mathcal{F}_0$. However, the action of the gauge group $\mathcal{G}$ on $\mathcal{F}$ acquires a central extension $\hat{\mathcal{G}} \subset \widehat{LK}$, where LK is the free loop group of K. The Lie algebra cocycle determining the central extension is

$$c(X,Y) = \frac{1}{2\pi\mathrm{i}}\int_{S^1}\mathrm{tr}_\rho X\,\mathrm{d}Y \qquad [17]$$

where tr_ρ is the trace in the representation ρ of K. Because of the central extension, the quotient $\mathcal{F}/\hat{\mathcal{G}}$ defines only a projective vector bundle over $\mathcal{A}/\mathcal{G}$, the Dixmier–Douady class being given by [16].

In the Example 1 (and Example 2) above, the complex line bundles can be constructed quite explicitly. Let us study the case $K = \mathrm{SU}(n)$. Define $U_\lambda \subset K$ as the set of matrices g such that λ is not an eigenvalue of g. Select n different points λ_j on the unit circle such that their product is not equal to 1. We assume that the points are ordered counterclockwise on the circle. Then the sets $U_j = U_{\lambda_j}$ form an open cover of $\mathrm{SU}(n)$. On each U_j we can choose a continuous branch of the logarithmic function $\log: U_j \to \boldsymbol{su}(n)$. The spectrum of the Dirac operator D_g with the holonomy g consists of the infinite set of numbers $\mathbb{Z} + \mathrm{Spec}(-\mathrm{i}\log(g))$. In particular, the numbers $\mathbb{Z} - \mathrm{i}\log\lambda_j$ do not belong to the spectrum of D_g. Choosing $\mu_k = -\mathrm{i}\log\lambda_k$ as an increasing sequence in the interval $[0,2\pi]$, we can as well define $U_j = \{x \in M | \mu_j \notin \mathrm{Spec}(D_x)\}$. In any case, the

top exterior power of the spectral subspace $E_{\mu_j,\mu_k}(x)$ is given by zero Fourier modes consisting of the spectral subspace of the holonomy g in the segment $[\lambda_j, \lambda_k]$ of the unit circle.

Index Theory and Gerbes

Gauge and gravitational anomalies in quantum field theory can be computed by Atiyah–Singer index theory. The basic setup is as follows. On a compact even-dimensional spin manifold S (without boundary) the Dirac operators coupled to vector potentials and metrics form a family of Fredholm operators. The parameter space is the set $\mathcal{A}$ of smooth vector potentials (gauge connections) in a vector bundle over S and the set of smooth Riemann metrics on S. The family of Dirac operators is covariant with respect to gauge transformations and diffeomorphisms of S; thus, we may view the Dirac operators parametrized by the moduli space $\mathcal{A}/\mathcal{G}$ of gauge connections and the moduli space $\mathcal{M}/\mathrm{Diff}_0(S)$ of Riemann metrics. Again, in order that the moduli spaces are smooth manifolds, one has to restrict to the based gauge transformations, that is, those which are equal to the neutral element in a fixed base point in each connected component of S. Similarly, the Jacobian of a diffeomorphism is required to be equal to the identity matrix at the base points. Passing to the quotient modulo gauge transformations and diffeomorphims, we obtain a vector bundle over the space

$$S \times \mathcal{A}/\mathcal{G} \times \mathcal{M}/\mathrm{Diff}_0(S) \qquad [18]$$

Actually, we could as well consider a generalization in which the base space is a fibering over the moduli space with model fiber equal to S, but for simplicity we stick to [18].

According to the Atiyah–Singer index formula for families, the K-theory class of the family of Dirac operators acting on the smooth sections of the tensor product of the spin bundle and the vector bundle V over [18] is given through the differential forms

$$\hat{A}(R) \wedge \mathrm{ch}(V)$$

where $\hat{A}(R)$ is the A-roof genus, a function of the Riemann curvature tensor R associated with the Riemann metric,

$$\hat{A}(R) = \det{}^{1/2}\left(\frac{R/4\pi\mathrm{i}}{\sinh(R/4\pi\mathrm{i})}\right)$$

and $\mathrm{ch}(V)$ is the Chern character

$$\mathrm{ch}(V) = \mathrm{tr}\, e^{F/2\pi\mathrm{i}}$$

where F is the curvature tensor of a gauge connection. Here both R and F are forms on the infinite-dimensional base space [18]. After integrating over the fiber S,

$$\mathrm{Ind} = \int_S \hat{A}(R) \wedge \mathrm{ch}(V) \qquad [19]$$

we obtain a family of differential forms ϕ_{2k}, one in each even dimension, on the moduli space.

The (cohomology classes of) forms ϕ_{2k} contain important topological information for the quantized Yang–Mills theory and for quantum gravity. The form ϕ_2 describes potential chiral anomalies. The chiral anomaly is a manifestation of gauge or reparametrization symmetry breaking. If the class $[\phi_2]$ is nonzero, the quantum effective action cannot be viewed as a function on the moduli space. Instead, it becomes a section of a complex line bundle DET over the moduli space.

Since the Dirac operators are Fredholm (on compact manifolds), at a given point in the moduli space we can define the complex line

$$\mathrm{DET}_x = \bigwedge{}^{\mathrm{top}}(\ker D_x^+) \otimes \bigwedge{}^{\mathrm{top}}(\mathrm{coker}\, D_x^+) \qquad [20]$$

for the chiral Dirac operators D_x^+. In the even-dimensional case, the spin bundle is $\mathbb{Z}_2$ graded such that the grading operator Γ anticommutes with D_x. Then $D_x^+ = P_- D_x P_+$, where $P_\pm = (1/2)(1 \pm \Gamma)$ are the chiral projections. $\wedge^{\mathrm{top}}$ means the operation on finite-dimensional vector spaces W taking the exterior power of W to dim W.

When the dimensions of the kernel and cokernel of D_x are constant, eqn [20] defines a smooth complex line bundle over the moduli space. In the case of varying dimensions, a little extra work is needed to define the smooth structure.

The form ϕ_2 is the Chern class of DET. So if DET is nontrivial, gauge covariant quantization of the family of Dirac operators is not possible.

One can also give a geometric and topological meaning to the chiral symmetry breaking in Hamiltonian quantization, and this leads us back to gerbes on the moduli space. Here we have to use an odd version of the index formula [19]. Assuming that the physical spacetime is even dimensional, at a fixed time the space is an odd-dimensional manifold S. We still assume that S is compact. In this case, the integration in [19] is over odd-dimensional fibers and, therefore, the formula produces a sequence of odd forms on the moduli space.

The first of the odd forms ϕ_1 gives the spectral flow of a one-parameter family of operators $D_{x(s)}$. Its integral along the path $x(t)$, after a correction by the difference of the eta invariant at the end points

of the path, in the moduli space, gives twice the difference of positive eigenvalues crossing over to the negative side of the spectrum minus the flow of eigenvalues in the opposite direction. The second term ϕ_3 is the Dixmier–Douady class of the projective bundle of Fock spaces over the moduli space. In Examples 1 and 2, the index theory calculation gives exactly the form [16] on K.

Example Consider Dirac operators on the three-dimensional sphere S^3 coupled to vector potentials. Any vector bundle on S^3 is trivial, so let $V = S^3 \times \mathbb{C}^N$. Take SU($N$) as the gauge group and let $\mathcal{A}$ be the space of 1-forms on S^3 taking values in the Lie algebra $\boldsymbol{su}(N)$ of SU(N). Fix a point x_s on S^3, the "south pole," and let $\mathcal{G}$ be the group of gauge transformations based at x_s. That is, $\mathcal{G}$ consists of smooth functions $g: S^3 \to \mathrm{SU}(N)$ with $g(x_s) = 1$. In this case $\mathcal{A}/\mathcal{G}$ can be identified as $\mathrm{Map}(S^2, \mathrm{SU}(N))$ times a contractible space. This is because any point x on the equator of S^3 determines a unique semicircle from the south pole to the north pole through x. The parallel transport along this path with respect to a vector potential $A \in \mathcal{A}$ defines an element $g'_A(x) \in \mathrm{SU}(N)$, using the fixed trivialization of V. Set $g_A(x) = g'_A(x) g'_A(x_0)^{-1}$, where x_0 is a fixed point on the equator. The element $g_A(x)$ then depends only on the gauge equivalence class $[A] \in \mathcal{A}/\mathcal{G}$. It is not difficult to show that the map $A \mapsto g_A$ is a homotopy equivalence from the moduli space of gauge potentials to the group $\mathcal{G}_2 = \mathrm{Map}_{x_0}(S^2, \mathrm{SU}(N))$, based at x_0. When $N > 2$, the cohomology $H^5(\mathrm{SU}(N), \mathbb{Z}) = \mathbb{Z}$ transgresses to the cohomology $H^3(\mathcal{G}_2, \mathbb{Z}) = \mathbb{Z}$. In particular, the generator

$$\omega_5 = \left(\frac{\mathrm{i}}{2\pi}\right)^3 \frac{2}{5!} \mathrm{tr}(g^{-1}\,\mathrm{d}g)^5$$

of $H^5(\mathrm{SU}(N), \mathbb{Z})$ gives the generator of $H^3(\mathcal{G}_2, \mathbb{Z})$ by contraction and integration,

$$\Omega = \int_{S^2} \omega_5$$

Gauge Group Extensions

The new feature for gerbes associated with Dirac operators in higher than one dimension is that the gauge group, acting on the bundle of Fock spaces parametrized by vector potentials, is represented through an abelian extension. On the Lie algebra level this means that the Lie algebra extension is not given by a scalar cocycle c as in the one-dimensional case but by a cocycle taking values in an abelian Lie algebra. In the case of Dirac operators coupled to vector potentials, the abelian Lie algebra consists of a certain class of complex functions on $\mathcal{A}$. The extension is then defined by the commutators

$$[(X,\alpha),(Y,\beta)] = ([X,Y], \mathcal{L}_X\beta - \mathcal{L}_Y\alpha + c(X,Y)) \quad [21]$$

where α, β are functions on $\mathcal{A}$ and $\mathcal{L}_X\beta$ denotes the Lie derivative of β in the direction of the infinitesimal gauge transformation X. The 2-cocycle property of c is expressed as

$$c([X,Y],Z) + \mathcal{L}_X c(Y,Z) + \text{cyclic permutations of } X, Y, Z = 0$$

In the case of Dirac operators on a 3-manifold S the form c is the Mickelsson–Faddeev cocycle

$$c(X,Y) = \frac{i}{12\pi^2} \int_S \mathrm{tr}_\rho A \wedge (\mathrm{d}X \wedge \mathrm{d}Y - \mathrm{d}Y \wedge \mathrm{d}X) \quad [22]$$

The corresponding gauge group extension is an extension of $\mathrm{Map}(S, G)$ by the normal subgroup $\mathrm{Map}(\mathcal{A}, S^1)$. As a topological space, the extension is the product

$$\mathrm{Map}(\mathcal{A}, S^1) \times_{S^1} P$$

where P is a principal S^1 bundle over $\mathrm{Map}(S, G)$.

The Chern class c_1 of the bundle P is again computed by transgression from ω_5; this time

$$c_1 = \int_S \omega_5$$

In fact, we can think of the cocycle c as a 2-form on the space of flat vector potentials $A = g^{-1}\mathrm{d}g$ with $g \in \mathrm{Map}(S^3, G)$. Then one can show that the cohomology classes $[c]$ and $[c_1]$ are equal.

As we have seen, the central extension of a loop group is the key to understanding the quantum field theory gerbe. Here is a brief description of it starting from the 3-form [16] on a compact Lie group G. First define a central extension $\mathrm{Map}(D, G) \times S^1$ of the group of smooth maps from the unit disk D to G, with pointwise multiplication. The group multiplication is given as

$$(g,\lambda)\cdot(g',\lambda') = (gg', \lambda\lambda' \cdot \mathrm{e}^{2\pi\mathrm{i}\gamma(g,g')})$$

where

$$\gamma(g,g') = \frac{1}{8\pi^2} \int_D \mathrm{tr}_\rho\, g^{-1}\mathrm{d}g \wedge \mathrm{d}g' g'^{-1} \quad [23]$$

where the trace is computed in a fixed unitary representation ρ of G. This group contains as a normal subgroup the group N consisting of pairs $(g, \mathrm{e}^{2\pi\mathrm{i}C(g)})$ with

$$C(g) = \frac{1}{24\pi^2} \int_B \mathrm{tr}_\rho\,(g^{-1}\mathrm{d}g)^3 \quad [24]$$

Here $g(x)=1$ on the boundary circle $S^1=\partial D$, and thus can be viewed as a function $S^2\to G$. The three-dimensional unit ball B has S^2 as a boundary and g is extended in an arbitrary way from the boundary to the ball B. The extension is possible since $\pi_2(G)=0$ for any finite-dimensional Lie group. The value of $C(g)$ depends on the extension only modulo an integer and therefore $e^{2\pi iC(g)}$ is well defined.

The central extension is then defined as

$$\widehat{LG}=(\mathrm{Map}(D,G)\times S^1)/N$$

One can show easily that the Lie algebra of $\widehat{LG}$ is indeed given through the cocycle [17]. When $G=\mathrm{SU}(n)$ in the defining representation, this central extension is the basic extension: The cohomology class is the generator of $H^2(LG,\mathbb{Z})$. In general, to obtain the basic extension one has to correct [23] and [24] by a normalization factor.

This construction generalizes to the higher loop groups $\mathrm{Map}(S,G)$ for compact odd-dimensional manifolds S. For example, in the case of a 3-manifold, one starts from an extension of $\mathrm{Map}(D,G)$, where D is a 4-manifold with boundary S. The extension is defined by a 2-cocycle γ, but now for given g,g' the cocycle γ is a real-valued function of a point $g_0\in\mathrm{Map}(S,G)$, which is a certain differential polynomial in the Maurer–Cartan 1-forms $g_0^{-1}\mathrm{d}g_0, g^{-1}\mathrm{d}g, g^{-1}\mathrm{d}g$. The normal subgroup N is defined in a similar way; now $C(g)$ is the integral of the 5-form ω_5 over a 5-manifold B with boundary ∂B identified as $D/\sim$, the equivalence shrinking the boundary of D to one point. This gives the extension only over the connected component of identity in $\mathrm{Map}(S,G)$, but it can be generalized to the whole group. For example, when $S=S^3$ and G is simple, the connected components are labeled by elements of the third homotopy group $\pi_3G=\mathbb{Z}$.

In some cases, the de Rham cohomology class of the extension vanishes but the extension still contains interesting torsion information. In quantum field theory this comes from Hamiltonian formulation of global anomalies. A typical example of this phenomenon is the Witten SU(2) anomaly in four spacetime dimensions. In the Hamiltonian formulation, we take S^3 as the physical space, the gauge group $G=\mathrm{SU}(2)$. In this case, the second cohomology of $\mathrm{Map}(S^3,G)$ becomes pure torsion, related to the fact that the 5-form ω_5 on SU(2) vanishes for dimensional reasons. Here the homotopy group $\pi_4(G)=\mathbb{Z}_2$ leads the nontrivial fundamental group $\mathbb{Z}_2$ in each connected component of $\mathrm{Map}(S^3,G)$. Using this fact, one can show that there is a nontrivial $\mathbb{Z}_2$ extension of the group $\mathrm{Map}(S^3,G)$.

See also: Anomalies; Bosons and Fermions in External Fields; Characteristic Classes; Dirac Operator and Dirac Field; Index Theorems; *K*-Theory.

Further Reading

Araki H (1987) Bogoliubov automorphisms and Fock representations of canonical anticommutation relations. *Contemporary Mathematics*, vol. 62. Providence: American Mathematical Society.

Berline N, Getzler E, and Vergne M (1992) *Heat Kernels and Dirac Operators*, Die Grundlehren der mathematischen Wissenschaften, vol. 298. Berlin: Springer.

Brylinski J-L *Loop Spaces, Characteristic Classes, and Geometric Quantization*, Progress in Mathematics, vol. 107. Boston: Birkhäuser.

Carey AL, Mickelsson J, and Murray M (2000) Bundle gerbes applied to quantum field theory. *Reviews in Mathematical Physics* 12: 65–90.

Gawedzki K and Reis N (2002) WZW branes and gerbes. *Reviews in Mathematical Physics* 14: 1281–1334.

Giraud J (1971) *Cohomologie Non Abelienne*, (French) Die Grundlehren der Mathematischen Wissenschaften, vol. 179. Berlin: Springer.

Hitchin N (2001) Lectures on special Lagrangian submanifolds. In: *Winter School on Mirror Symmetry, Vector Bundles and Lagrangian Submanifolds (Cambridge, MA, 1999)*, AMS/IP Stud. Adv. Math., vol. 23, pp. 151–182. Providence, RI: American Mathematical Society.

Mickelsson J (1989) *Current Algebras and Groups*. London: Plenum.

Treiman SB, Jackiw R, Zumino B, and Witten E (1985) *Current Algebra and Anomalies*, Princeton Series in Physics. Princeton, NJ: Princeton University Press.

Holonomic Quantum Fields

J Palmer, University of Arizona, Tucson, AZ, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The term, holonomic field, was coined by Sato, Miwa, and Jimbo (SMJ) in 1978 and the subject was investigated by them in a series of five long papers and many shorter notes in the period from 1978–81 (Sato *et al.* 1979a, 1979b, 1979c, 1980, Tracy and Widom 1994). The term refers to a special class of two-dimensional interacting quantum field theories whose n point correlations can be expressed in terms of the solution to a holonomic system of differential equations. A holonomic system is an overdetermined system of differential equations with only a finite-dimensional family of solutions. There is a sense in which these interacting systems with infinitely many degrees of freedom have a finite-dimensional substrate (at the level of n point functions for fixed n). After developing their theory, SMJ realized that such quantum fields made an earlier appearance in work of Thirring and Federbush. The models considered by Thirring and Federbush are self-interacting fermionic systems whose nonlinear classical field equations have solutions that are an explicit nonlinear transformation of solutions to the free field equations. This inspired the idea of trying to study these models by "quantizing" the nonlinear transformation. Expressions were obtained for the correlations and S-matrix but the connection with deformation theory was not understood until the SMJ work.

In what follows we will sketch the SMJ theory and discuss some of its offshoots. There is one circumstance that it might help the reader to be aware of even though it will be mostly glossed over. Quantum fields in one space and one time dimension have correlations which transform under the symmetries of spacetime with metric signature (1,1). Since the work of Osterwalder and Schrader, it is customary to pass back and forth between this Minkowski regime and the Schwinger functions obtained by analytically continuing the n point functions to pure imaginary values for the time variable where they possess the rotational symmetries associated with a positive-definite metric. The Ising model, which we take up next, is naturally considered in the Euclidean domain where the correlations have an interpretation in statistical mechanics as the expected value of a product of random variables. Ultimately, the SMJ deformation analysis is done in the Euclidean domain.

The Two-Dimensional Ising Model

The SMJ theory was inspired by, and provides an attractive setting for, an earlier result of Wu, McCoy, Tracy, and Baruch (WMTB), concerning the spin–spin scaling functions of the two-dimensional Ising model (Wu *et al.* 1976). Since the Ising model is the example with the most direct significance for physics, we will take some time to explain the WMTB result and to sketch the way in which it fits into the SMJ theory.

The Ising model is a statistical model of magnestism on a lattice that incorporates ferromagnetic interactions of nearest-neighbor spins. In the 1920s, Ising solved the model for the one-dimensional lattice and showed that there was no phase transition in the infinite volume limit. Interest in the two-dimensional model intensified dramatically following Onsager's calculation of the specific heat in the infinite volume limit (see Palmer and Tracy (1981) and references within). His formula for the specific heat was the first instance of a thermodynamic quantity in a nearest-neighbor model which exhibits the sort of discontinuity in temperature dependence expected at a phase transition. For many years, the Ising model served as a testbed for the now accepted notion that the infinite volume limit of Gibbsian statistical mechanics provides a suitable setting for the study of phase transitions.

A configuration for the Ising model on a finite subset, Λ, of the integer lattice, $\mathbf{Z}^2$, is a map C: $\Lambda \to \{+1, -1\}$, which assigns to each site on the lattice either an up spin $(+1)$ or a down spin (-1). The energy function of the Ising model, $E_\Lambda(\sigma)$, is defined by

$$E_\Lambda(\sigma) = -J \sum_{\langle i,j\rangle \in \Lambda} \sigma(i)\sigma(j)$$

for $J > 0$ and a spin configuration σ is a sum over pairs of nearest-neighbor sites i, j in Λ (boundary terms require special consideration). This energy function tends to favor spin configurations, σ, in which the nearest-neighbor spins are aligned in the sense that the Boltzmann weight, $e^{-E_\Lambda(\sigma)/kT}$, is larger for such configurations. In the Gibbs ensemble, which is expected to describe systems in equilibrium at temperature T, the configuration σ occurs with a probability proportional to the Boltzmann weight. The factor k which appears is a conversion factor between thermal and kinetic energy called the Boltzmann constant. It is clear from the formula for the Boltzmann weight that small temperatures (near 0) tend to accentuate the difference in

statistical weights assigned to configurations with different energies, and large temperatures tend to wash out the difference in statistical weights associated to configurations with different energies.

Remarkably, there is a sharp critical temperature $0 < T_c < \infty$ so that for $T < T_c$ the propensity for order built into the energy triumphs in the infinite volume limit $\Lambda \uparrow \mathbf{Z}^2$, and for $T > T_c$ the randomness or disorder associated with high temperatures governs the infinite volume behavior. More specifically, if $T < T_c$ and the infinite volume limit is taken with plus spins assigned to the boundary of Λ, the system exhibits a residual magnetism (there is a positive expected value, $\langle\sigma\rangle$, for the spin per site). This infinite volume plus state is the quintessential example of symmetry breaking – the spin flip symmetry possessed by the bulk energy is broken below T_c in the thermodynamic limit. For $T > T_c$, the spin per site is 0 no matter what boundary conditions are imposed on the infinite volume limit.

Pure equilibrium states both above and below T_c exhibit clustering in the thermodynamic limit (uniqueness for the ground state in field theory). This is the tendency of spin variables $\sigma(a)$ and $\sigma(b)$ at sites $a, b \in \mathbf{Z}^2$ to become statistically independent as the distance $|a-b|$ tends to ∞. In such a pure state the two-point function, which is the expected value of the product of spin variables, $\langle\sigma(a)\sigma(b)\rangle$, will tend to the square $\langle\sigma\rangle^2$ both below ($\langle\sigma\rangle \neq 0$) and above ($\langle\sigma\rangle = 0$) the critical temperature T_c as $|a-b| \to \infty$. To leading order, this clustering takes place at an exponential rate, $e^{-|a-b|/\xi(T)}$, for a function $\xi(T)$ called the correlation length. The correlation length $\xi(T) \to \infty$ as $T \to T_c$. The scaling limit (from below T_c) of the spin–spin correlation is the leading-order correction to the clustering behavior of the correlations when these correlations are examined at the scale of the correlation length. It is the limit

$$\langle\sigma(a)\sigma(b)\rangle = \lim_{T \uparrow T_c} \frac{\langle\sigma(\xi(T)a)\sigma(\xi(T)b)\rangle_T}{\langle\sigma\rangle_T^2}$$

where the correlations on the right-hand side are thermodynamic correlations on the lattice at temperature T. Since $\langle\sigma\rangle_T$ tends to 0 as $T \to T_c$, the normalization by $\langle\sigma\rangle_T^{-1}$ on the right produces an "infinite wave function renormalization" in the limit.

Equivalently, one may think of this continuum limit being achieved by letting the lattice spacing shrink to 0 as T approaches T_c so that the correlation length stays fixed on the new scale. The scaling limit from above T_c turns out to be different from the scaling limit from below T_c and since $\langle\sigma\rangle_T = 0$ for $T > T_c$, it is defined by a different wave function renormalization. The resulting asymptotics are expected to capture quite a lot about what is interesting in the behavior of the correlations near the phase transition. In the late 1970s, the scaling behavior in this model was also a prototype for the emerging connection between renormalization group ideas in quantum field theory and statistical mechanics.

Wu *et al.* (1976) showed that the two-point scaling function, $\langle\sigma(0)\sigma(x)\rangle$, is a function of $r = |x|$ and can be written as

$$\langle\sigma(0)\sigma(x)\rangle = \begin{cases} \cosh(\psi/2) \\ \quad \times \exp\frac{1}{4}\int_r^\infty \left(\sinh^2\psi - \left(\frac{d\psi}{dt}\right)^2\right) s\,ds \\ \quad \times (T \uparrow T_c) \\ \sinh(\psi/2) \\ \quad \times \exp\frac{1}{4}\int_r^\infty \left(\sinh^2\psi - \left(\frac{d\psi}{dt}\right)^2\right) s\,ds \\ \quad \times (T \downarrow T_c) \end{cases} \qquad [1]$$

where $\psi = \psi(r)$ satisfies the differential equation,

$$\frac{d}{dr}\left(r\frac{d\psi}{dr}\right) = \frac{r}{2}\sinh(2\psi)$$

The substitution $\eta = e^{-\psi}$ transforms this differential equation into a Painlevé equation of the third kind. This was used by McCoy, Tracy, and Wu (see Palmer and Tracy (1981) and references within) to study the short-distance behavior, $r \to 0$, of the scaling functions – behavior which is far from manifest in the infinite series expansions obtained for the scaling functions.

Deformation Theory

Sato, Miwa, and Jimbo showed that there was a class of quantum field theories that included the scaling limits of the Ising model which have the property that the n-point correlations are "tau functions" for monodromy-preserving deformations of the Dirac equation in two dimensions. The two-dimensional (Euclidean) Dirac operator is

$$\mathcal{D} = \begin{bmatrix} m & -2\partial \\ -2\bar{\partial} & m \end{bmatrix}$$

with

$$\partial = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right), \qquad \bar{\partial} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)$$

the usual complex derivatives acting on smooth functions on R^2. The monodromy-preserving deformations mentioned above are families of (multivalued) solutions $w(x)$ to

$$\mathcal{D}w = 0 \qquad [2]$$

which are branched at points $a_j \in R^2 (j = 1, 2, \ldots, n)$ and change by a factor $e^{2\pi i \ell_j}$ as x makes a small circuit about a_j. SMJ (1979b) show that the $L^2(R^2)$ (square-integrable) solutions $w(x)$ of the Dirac equation with this prescribed branching behavior comprise an n-dimensional subspace of $L^2(R^2)$. The constants $e^{2\pi i \ell_j}$ are called the monodromy of the solutions and the term "deformation" in the description refers to the fact that the monodromy constants do not change as the branch points a_j are varied. SMJ show that it is possible to choose a basis $w_j(x, a) = w_j(x, a_1, \ldots, a_n)$ $(j = 1, \ldots, n)$ so that the vector

$$W(x, a) := [w_1(x, a), w_2(x, a), \ldots, w_n(x, a)]$$

becomes a section for a flat (Dirac compatible) connection in the (x, a) variables. That is,

$$d_{x,a} W = \Omega(x, a) W$$

where $d_{x,a}$ is the exterior derivative in the x and a variables and Ω is a matrix-valued 1-form that satisfies the zero curvature condition,

$$d_{x,a} \Omega = \Omega \wedge \Omega$$

They also introduced the notion of a tau function, $\tau(a)$, for such deformations. The logarithmic derivative $d_a \log \tau(a) = \omega$, where ω is a 1-form on $R^2 \backslash \{a_1, a_2, \ldots, a_n\}$ expressed in terms of the matrix elements of Ω. The 1-form ω introduced by SMJ is shown to be closed when Ω satisfies the zero curvature condition above. The scaling limit of the Ising model is related to the situation for monodromy multipliers -1 and when the scaling limits of correlations are identified as suitable τ-functions in this case, the WMTB result emerges when the nonlinear zero curvature condition is identified with a Painlevé equation.

The connection between the deformation theory and quantum field theory is developed in the computationally intensive paper SMJ (1980). Extensive use is made of local operator product expansions, analytic continuation, and formal series expansions that are infinite-dimensional analogs of Wick-type theorems for finite-dimensional spin representations (developed by SMJ (1978)). One can get a feeling for the source of the connection by recalling that in one of the "exact solutions" of the two-dimensional Ising model the spin operators, $\sigma(a)$, are identified as elements of an infinite spin representation of the orthogonal group and are characterized by their linear action on Fermionic variables (Palmer and Tracy 1981). In the physics literature, the $\sigma(a)$ are referred to as Bogoliubov transformations. In the scaling limit the associated representation space is the home to a free Fermi field $\psi(x)$, an operator-valued solution to the Dirac equation. Of course, $\psi(x)$ has components $\psi_j(x)$ but for simplicity we will suppress such details in the mostly schematic discussion that follows. For coincident second coordinates $x_2 = a_2$ the Fermi field $\psi(x)$ and $\sigma(a)$ satisfy the commutation relation

$$\sigma(a)\psi(x) = -\mathrm{sgn}(x_1 - a_1)\psi(x)\sigma(a) \qquad [3]$$

which is a surviving remnant of the linear action of $\sigma(a)$ on lattice fermions. In the transfer matrix formalism, which is natural for statistical mechanics, translation in the "space" variable x_1 is unitary, but translation in the "time" variable, x_2, is governed by the transfer matrix, the generator of a contractive semigroup. Because of this, the quantities that are well behaved in this formalism are "time-ordered vacuum expectations"; these involve only "positive" powers of the transfer matrix. Let $\mathcal{T}$ denote the "time"-ordering operator; a sequence of operators depending on coordinates in R^2 is reordered following $\mathcal{T}$ so that the second coordinates appear in increasing order from left to right. Sign changes are incorporated whenever it is necessary to exchange Fermi type operators like $\psi^*(x)$ and $\psi(y)$ to put them in the correct order. In the Euclidean setting (pure imaginary time) it is well known that

$$G(x, y) = \langle \mathcal{T} \psi^*(x) \psi(y) \rangle$$

is a Green function for the Dirac operator $\mathcal{D}$ (the distribution kernel for $\mathcal{D}^{-1}$).

This observation and [3] suggests that the hybrid vacuum expectation

$$G(x, y; a) = \frac{\langle \mathcal{T} \psi^*(x) \psi(y) \sigma(a_1) \cdots \sigma(a_n) \rangle}{\langle \mathcal{T} \sigma(a_1) \cdots \sigma(a_n) \rangle}$$

should be the Green function for a Dirac operator with a domain containing "functions" branched at the points a_j having "monodromy" -1 there. It is possible to recast the SMJ analysis so that a Dirac operator, $\mathcal{D}(a)$, on a suitable vector bundle with base $R^2 \backslash \{a_1, \ldots, a_n\}$ becomes the central player (see Palmer *et al.* (1994) and references therein). The data for the vector bundle includes the factors $e^{2\pi i \ell_j}$ incorporated in transition functions for the bundle. The τ-function becomes an infinite determinant (or Pfaffian in the Ising case)

$$\tau(a) = \det \mathcal{D}(a) \qquad [4]$$

in the Segal–Wilson sense (see Palmer *et al.* (1994) and references therein). The Green function $G(x,y;a)$ has a finite-rank derivative,

$$d_a G(x,y;a) = \sum_j r_j(x,a)s_j(y,a)\,da_j + u_j(x,a)v_j(y,a)\,d\bar{a}_j \qquad [5]$$

which is the key result in this version of the SMJ analysis (this observation appears in SMJ (1980) but does not have a central role there). The "wave functions" r, s, u, and v are closely related to the L^2 wave functions w_j described above. Equation [5] is both the source of the deformation equations for r, s, u, and v which arise from $d_a^2 G = 0$ coupled with the rotational and translational symmetries of the Green function, and also of the expression for $d_a \log \tau(a)$ in terms of data associated with the deformation theory. A "transfer matrix" calculation of the determinant allows one to make the connection with the scaling limits of lattice fields including the Ising model (see Palmer *et al.* (1994) and references therein).

The short-distance behavior of the two-point function for the Ising model scaling functions has been rigorously calculated by Tracy and later by Tracy and Widom (see Harnad and Its (2002) and references therein). A less detailed analysis of the short-distance behavior of the n point functions that uses the deformation analysis of the correlations in a crucial way can be found in Palmer (2000).

The Riemann–Hilbert Problem

In SMJ (1979b), a "massless" version of holonomic fields is developed. This concerns monodromy-preserving deformations of the Cauchy–Riemann operator $\bar{\partial}$. The techniques used to study this lead back to the Riemann–Hilbert problem – the problem of determining a linear differential equation in the complex plane with rational coefficients and prescribed monodromy at the poles of the coefficients. More specifically, suppose one is given n distinct points $\{a_1,\ldots,a_n\}$ in $\boldsymbol{P}^1$, the Riemann sphere, and a base point a_0 distinct from the $a_j, j \neq 0$. Let γ_j denote a simple closed curve based at a_0 which winds counterclockwise once around a_j but has winding number 0 for the other points $a_k, k \neq j$. Choose n invertible $p \times p$ matrices M_j which satisfy the single condition $M_1 M_2 \cdots M_n = I$. Then, the homotopy classes of the curves γ_j are the generators for the fundamental group of the punctured sphere $\boldsymbol{P}^1 \backslash \{a_1,\ldots,a_n\}$ with base point a_0 and the map which sends $\gamma_j \to M_j$ determines a representation of the fundamental group. One version of the Riemann–Hilbert problem is to find $p \times p$ complex matrices A_j for $j = 1,\ldots,n$ so that the linear differential equation

$$\frac{dY}{dz} = \sum_j \frac{A_j}{z - a_j} Y \qquad [6]$$

has monodromy representation given by $\gamma_j \to M_j$. This means that the fundamental solution $Y(z)$ defined in a neighborhood of $z = a_0$ and normalized so the $Y(a_0) = I$ (the identity) will become the fundamental solution $Y(z)M_j^{-1}$ after analytic continuation around the curve γ_j. This form of the problem does not always have a solution but when it does, it is interesting to consider deformations $a \to A_j(a)$ that preserve the monodromy multipliers M_j. Such monodromy-preserving deformations were first considered by Schlesinger in 1912 (see Palmer and Tracy (1981) and references therein) and he discovered that the coefficients A_j must satisfy nonlinear differential equations that, for $a_0 = \infty$, can be written as

$$d_a A_j = \sum_{k \neq j} \frac{A_k - A_j}{a_k - a_j} d(a_k - a_j)$$

SMJ introduced τ-functions associated with these deformations and they gave these τ-functions a quantum field theory interpretation as n point functions. Eventually this theory was extended to include the Birkhoff generalization of the Riemann–Hilbert problem, a generalization which incorporates the additional information needed to fix local holomorphic equivalence at higher-order poles (formal asymptotics and Stokes' multipliers) (Jimbo and Miwa 1981, Sato *et al.* 1978). Roughly speaking, the problem is to reconstruct a global connection with specified singularities from its local holomorphic equivalence data and its global monodromy representation. Thinking of the differential equation [6] as a holomorphic connection proved very helpful in a geometric reworking of the SMJ analysis given by Malgrange (1983a, 1983b) who showed that the zeros of the τ-function occurred at points where a suitably defined Riemann–Hilbert problem fails to have a solution (see also Palmer (1999) references within). The mathematical significance of massless holonomic quantum fields as (quantized) singular elements of a gauge group is apparent from the SMJ work and later work of Miwa but the possibility of interesting physics in these models does not seem to have been much investigated at this time. These quantum fields are also conformal fields; however, a comprehensive integration into the highly developed formalism of conformal field theory on compact Riemann surfaces has not currently been developed

(an analog of [5] should survive on compact Riemann surfaces but the deformation analysis of the correlations is likely limited to symmetric spaces).

Further Developments

This work on massless holonomic fields and the connection with the Riemann–Hilbert problem is doubtless the aspect of holonomic fields with the most "spin offs" in the mathematics and physics literature. These include an analysis of the delta-function gas done by Jimbo, Miwa, Mori, and Sato in 1981, random matrix models first looked at by Jimbo, Miwa, Mori, and Sato and later systematically investigated by Tracy and Widom (1994), the deformations of line bundles on Riemann surfaces that led to KdV in the work of Segal and Wilson (1985), which emerged from work of Sato, Miwa, Jimbo and collaborators, the analysis of Painlevé equations starting with work of McCoy, Tracy and Wu (see Palmer and Tracy (1981) and references within) and more systematically developed by Its and Novokshenov (1986), and the revival of interest in monodromy-preserving deformations (Harnad and Its 2002).

Holonomic fields are related to free fields in a well-understood way and it is natural to study them in situations where free fields make sense. In particular, they are an interesting testbed for the nonperturbative investigation of the influence of geometry (or curvature) on quantum fields. In Palmer *et al.* (1994), the deformation analysis of τ-functions for holonomic fields is carried out for the Poincaré disk. The two-point functions are shown to be expressible in terms of solutions to the family of Painlevé VI equations. A quantum field theory interpretation of these τ-functions is given by Doyon and there are natural analogs of the scaling limit of the Ising model on the Poincaré disk as well. The role of "spacetime" symmetries in the deformation theory suggests that such analysis will be limited to symmetric spaces. In addition to the plane and the Poincaré disk, the cylinder, the sphere, and the torus round out the possibilities in two dimensions. Lisovyy has recently worked out the analysis for the cylinder, which is important for the study of thermodynamic correlations. It should be possible to recast the analysis of the continuum Ising model on the torus (Zuber and Itzykson 1977) in deformation theoretic terms. It does not appear that the holonomic fields associated with the Dirac operator for the constant curvature metric on the 2-sphere have been studied yet.

See also: Deformation Theory; Integrable Systems: Overview; Isomonodromic Deformations; Riemann–Hilbert Problem; Two-Dimensional Ising Model.

Further Reading

Harnad J and Its A (2002) *CRM Workshop: Isomonodromic Deformations and Applications in Physics*, vol. 31 American Mathematical Society.

Its AR and Novokshenov VYu (1986) *The Isomonodromic Deformation Method in the Theory of Painlevé Equations*, Lecture Notes in Mathematics, vol. 1191. Berlin: Springer.

Jimbo M and Miwa T (1981) Monodromy preserving deformations of linear ordinary differential equations with rational coefficients II. *Physica D* 2: 407–448.

Jimbo M, Miwa T, and Ueno K (1981) Monodromy preserving deformations of linear ordinary differential equations with rational coefficients I. *Physica D* 2: 306–352.

Malgrange B (1983a) *Sur les déformations isomonodromiques I, Singularités régulières*, Progress in Mathematics vol. 37, pp. 401–426. Boston: Birkhäuser.

Malgrange B (1983b) *Sur les déformations isomonodromiques II, Singularités irrégulières*, Progress in Mathematics vol. 37, pp. 427–438. Boston: Birkhäuser.

Palmer J (1999) Zeros of the Jimbo, Miwa, Ueno tau function. *Journal of Mathematical Physics* 40: 6638–6681.

Palmer J (2000) Short distance asymptotics of Ising correlations. *Journal of Mathematical Physics* 7: 1–50.

Palmer J, Beatty M, and Tracy C (1994) Tau functions for the Dirac operator on the Poincaré disk. *Communications in Mathematical Physics* 165: 97–173.

Palmer J and Tracy CA (1981) Two dimensional Ising correlations: Convergence of the scaling limit. *Advances in Applied Mathematics* 2: 329–388.

Sato M, Miwa T, and Jimbo M (1978) Holonomic quantum fields I. *Publications of the Research Institute for Mathematical Sciences, Kyoto University* 14: 223–267.

Sato M, Miwa T, and Jimbo M (1979a) Holonomic quantum fields II. *Publications of the Research Institute for Mathematical Sciences, Kyoto University* 15: 201–278.

Sato M, Miwa T, and Jimbo M (1979b) Holonomic quantum fields III. *Publications of the Research Institute for Mathematical Sciences, Kyoto University* 15: 577–629.

Sato M, Miwa T, and Jimbo M (1979c) Holonomic quantum fields IV. *Publications of the Research Institute for Mathematical Sciences, Kyoto University* 15: 871–972.

Sato M, Miwa T, and Jimbo M (1980) Holonomic quantum fields V. *Publications of the Research Institute for Mathematical Sciences, Kyoto University* 16: 531–584.

Tracy CA and Widom H (1994) Fredholm determinants, differential equations and matrix models. *Communications in Mathematical Physics* 163: 33–72.

Wu TT, McCoy B, Tracy CA, and Barouch E (1976) Spin–spin correlation functions for the two dimensional Ising model: exact theory in the scaling region. *Physical Review B* 13: 316–374.

Wilson G and Segal G (1985) Loop groups and equations of KdV type. *Publications of Mathématiques IHES* 61: 5–65.

Zuber JB and Itzykson C (1977) Quantum field theory and the two dimensional Ising model. *Physical Review D* 15: 2875–2884.

Hopf Algebra Structure of Renormalizable Quantum Field Theory

D Kreimer, IHES, Bures-sur-Yvette, France

© 2006 Elsevier Ltd. All rights reserved.

Overview

Renormalization theory is a venerable subject put to daily use in many branches of physics. Here, we focus on its applications in quantum field theory, where a standard perturbative approach is provided through an expansion in Feynman diagrams. Whilst the combinatorics of the Bogoliubov recursion, solved by suitable forest formulas, has been known for a long time, the subject regained interest on the conceptual side with the discovery of an underlying Hopf algebra structure behind these recursions.

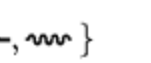
Perturbative expansions in quantum field theory are organized in terms of one-particle irreducible (1PI) Feynman graphs. The goal is to calculate the corresponding 1PI Green functions order by order in the coupling constants of the theory, by applying Feynman rules to these 1PI graphs of a

Notes:

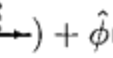
renormalizable theory under consideration. This allows one to disentangle the problem into an algebraic part and an analytic part.

For the algebraic part, one studies Feynman graphs as combinatorial objects which lead to the Lie and Hopf algebras discussed below. Feynman rules then assign analytic expressions to these graphs, with the analytic structure of finite renormalized quantum field theory largely dictated by the underlying algebra.

The objects of interest in quantum field theory are the 1PI Green functions. They are parametrized by the quantum numbers – masses, momenta, spin, and such – of the particles participating in the scattering process under consideration. We call a set of such quantum numbers an external leg structure $\underline{r}$. For example, the three terms in the Lagrangian of massless quantum electrodynamics correspond to

$$\underline{r} \in \{ \text{} \} \qquad [1]$$

Note that the Lagrangian L of massless quantum electrodynamics is obtained accordingly as

$$\begin{aligned} L &= \hat{\phi}(\text{})^{-1} + \hat{\phi}(\text{}) + \hat{\phi}(\text{})^{-1} \\ &= \bar{\psi}\partial\psi + \bar{\psi}A\psi + \tfrac{1}{4}F^2 \end{aligned} \qquad [2]$$

where $\hat{\phi}$ are coordinate space Feynman rules.

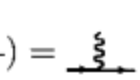
The renormalized 1PI Green function in momentum space, $G^{\underline{r}}_{\mathrm{R}}(\{g\};\{p\},\{m\};\mu)$, is obtained as the image under renormalized Feynman rules ϕ_{R} applied to a series of graphs:

$$\Gamma^{\underline{r}} = 1 + \sum_{k=1}^{\infty} g^k c^{\underline{r}}_k \equiv 1 + \sum_{\mathrm{res}(\Gamma)=\underline{r}} g^{|\Gamma|} \frac{\Gamma}{\mathrm{Sym}(\Gamma)} \qquad [3]$$

Here $\underline{r}$ is a given such external leg structure, while $c^{\underline{r}}_k$ is the finite sum of 1PI graphs having k loops,

$$c^{\underline{r}}_k = \sum_{\substack{\mathrm{res}(\Gamma)=\underline{r} \\ |\Gamma|=k}} \frac{\Gamma}{\mathrm{Sym}(\Gamma)} \qquad [4]$$

and $0 < g < 1$ is a coupling constant. The generalization to the case of several couplings $\{g\}$ and masses $\{m\}$ is straightforward. In the above, the sum is over all 1PI graphs with the same given external leg structure. We have denoted the map which assigns $\underline{r}$ to a given graph a residue, for example,

$$\mathrm{res}(\text{}) = \text{} \qquad [5]$$

The unrenormalized but regularized Feynman rules ϕ assign to a graph a function

$$\begin{aligned} &\phi(\Gamma)(\{g\};\{p\},\{m\};\mu,z) \\ &= \int \prod_{v\in\Gamma^{[0]}} \delta^{(4)}\left(\sum_{f \text{ incident } v} k_f \right) \prod_{e\in\Gamma^{[1]}_{\mathrm{int}}} \mathrm{Prop}(k_e) \frac{\mathrm{d}^4 k_e}{4\pi^2} \end{aligned} \qquad [6]$$

and formally the unrenormalized Green function

$$\begin{aligned} &G^{\underline{r}}_{\mathrm{u}}(\{g\};\{p\},\{m\};\mu,z) \\ &\quad = \phi(\Gamma^{\underline{r}})(\{g\};\{p\},\{m\};\mu,z) \end{aligned} \qquad [7]$$

which is a function of a suitably chosen regulator z. Note that in [6] the four-dimensional Dirac-δ distribution guarantees momentum conservation at each vertex and restricts the number of four-dimensional integrations to the number of independent cycles in the graph. It is assumed that the reader is familiar with the readily established fact that these integrals suffer from UV singularities, which render the integration over the momenta in internal cycles ill-defined. We also remind the reader that the problem persists in coordinate space, where one confronts the continuation of products of distributions to regions of coinciding support. We restrict ourselves here to a discussion of the situation in momentum space and refer the reader to the literature for the situation in coordinate space.

Ignoring problems of convergence in the sum over all graphs, the problem of renormalization is to make sense of these functions term by term: We have to determine invertible series $Z^{\underline{r}}(\{g\},z)$ in the couplings g such that the modified Lagrangian

$$\tilde{L} = \sum_{\underline{r}} Z^{\underline{r}}(\{g\},z)\, \hat{\phi}(\underline{r}) \qquad [8]$$

produces a perturbation series in graphs that allows for the removal of the regulator z.

This amounts to a transition from unrenormalized to renormalized Feynman rules $\phi \to \phi_{\mathrm{R}}$. Let us first describe how this transition is achieved using the Lie and Hopf algebra structure of the perturbative expansion, which is described in detail below:

- Decide on the free fields and local interactions of the theory, appropriately specifying quantum numbers (spin, mass, flavor, color, and such) of fields, restricting interactions so as to obtain a renormalizable theory.
- Consider the set of all 1PI graphs with edges corresponding to free-field propagators. Define vertices for local interactions. This allows one to construct a pre-Lie algebra of graph insertions. Antisymmetrize this pre-Lie product to get a Lie algebra $\mathcal{L}$ of graph insertions and define the Hopf algebra $\mathcal{H}$ which is dual to the enveloping algebra $\mathcal{U}(\mathcal{L})$ of this Lie algebra.
- Realize that the coproduct and antipode of this Hopf algebra give rise to the forest formula, which generates local counter-terms upon introducing a Rota–Baxter map, a renormalization scheme in physicists' parlance.

- Use the Hochschild cohomology of this Hopf algebra to show that one can absorb singularities in local counter-terms.
- Determine the corepresentations of this Hopf algebra to identify the sub-Hopf algebras corresponding to time-ordered products in physical fields. This is most easily achieved by rewriting the Dyson–Schwinger equations using Hochschild 1-cocycles.

The last point exhibits close connections, in particular, between the structure of gauge theories and the corepresentation theory of their perturbative Hopf algebras which we discuss below in brief.

This program can be carried out in coordinate space as well as momentum space renormalization. It has given a firm mathematical background to the process of renormalization, justifying the practice of quantum field theory. The notion of locality has achieved a precise formulation in terms of the Hochschild cohomology of the perturbation expansion. In momentum space, this approach emphasizes the connections to number theory, which emerge when one investigates the role of the Hopf algebra primitives, which in turn furnish the Hochschild 1-cocycles underlying locality.

The next sections describe the above setup in some detail.

Lie and Hopf Algebras of Graphs

All algebras are supposed to be over some field $\mathbb{K}$ of characteristic zero, associative and unital, and similarly for coalgebras. The unit (and, by abuse of notation, also the unit map) will be denoted by I, the counit map by $\bar{e}$. All algebra homomorphisms are supposed to be unital. A bialgebra $(A=\bigoplus_{i=0}^{\infty}A_i, m, \mathrm{I}, \Delta, \bar{e})$ is called graded connected if $A_iA_j\subset A_{i+j}$ and $\Delta(A_i)\subset\bigoplus_{j+k=i}A_j\otimes A_k$, and if $\Delta(\mathrm{I})=\mathrm{I}\otimes\mathrm{I}$ and $A_0=k\mathrm{I}, \bar{e}(\mathrm{I})=1\in\mathbb{K}$ and $\bar{e}=0$ on $\bigoplus_{i=1}^{\infty}A_i$. We call $\ker\bar{e}$ the augmentation ideal of A and denote by P the projection $A\to\ker\bar{e}$ onto the augmentation ideal, $P=\mathrm{id}-\mathrm{I}\bar{e}$. Furthermore, we use Sweedler's notation, $\Delta(h)=\sum h'\otimes h''$, for the coproduct. We define

$$\mathrm{Aug}^{(k)}=\left(\underbrace{P\otimes\cdots\otimes P}_{k\,\text{times}}\right)\Delta^{k-1},\qquad A\to\{\ker\bar{e}\}^{\otimes k} \tag{9}$$

as a map into the k-fold tensor product of the augmentation ideal. We let $A^{(k)}=\ker\mathrm{Aug}^{(k+1)}/\ker\mathrm{Aug}^{(k)}, \forall\, k\geq 1$. All bialgebras considered here are bigraded in the sense that

$$A=\bigoplus_{i=0}^{\infty}A_{(i)}=\bigoplus_{k=0}^{\infty}A^{(k)} \tag{10}$$

where $A_{(k)}\subset\oplus_{j=1}^{k}A^{(j)}$ for all $k\geq 1$. $A_{(0)}\simeq A^{(0)}\simeq\mathbb{K}$.

The first construction we have to study is the pre-Lie algebra structure of 1PI graphs.

The Pre-Lie Structure

For each Feynman graph we have vertices as well as internal and external edges. External edges are edges that have an open end not connected to a vertex. They indicate the particles participating in the scattering amplitude under consideration and each such edge carries the quantum numbers of the corresponding free field. The internal edges and vertices form a graph in their own right. For an internal edge, both ends of the edge are connected to a vertex.

We consider 1PI Feynman graphs. A graph Γ is 1PI if and only if all graphs, obtained by removal of any one of its internal edges, are still connected. Such 1PI graphs are naturally graded by their number of independent loops, the rank of their first homology group $H_{[1]}(\Gamma,\mathbb{Z})$. We write $|\Gamma|$ for this degree of a graph Γ. Note that $|\mathrm{res}(\Gamma)|=0$, where we let $\mathrm{res}(\Gamma)$ be the graph obtained when all edges in $\Gamma_{\mathrm{int}}^{[1]}$ shrink to a point, as before. Note that the graph obtained in this manner consists of a single vertex, to which the edges $\Gamma_{\mathrm{ext}}^{[1]}$ are attached.

For a 1PI graph Γ, $\Gamma^{[0]}$ denotes its set of vertices and $\Gamma^{[1]}=\Gamma_{\mathrm{int}}^{[1]}\cup\Gamma_{\mathrm{ext}}^{[1]}$ its set of internal and external edges. In addition, let $\omega_{\underline{r}}$ be the number of spacetime derivatives appearing in the corresponding monomial in the Lagrangian.

Having specified free quantum fields and local interaction terms between them, one immediately obtains the set of 1PI graphs. One can then consider for a given external leg structure $\underline{r}$ the set of graphs with that external leg structure. For a renormalizable theory, we can define a superficial degree of divergence,

$$\omega=\sum_{\underline{r}\in\Gamma_{\mathrm{int}}^{[1]}\cup\Gamma^{[0]}}\omega_{\underline{r}}-4|H_{[1]}(\Gamma,\mathbb{Z})| \tag{11}$$

for each such external leg structure: $\omega(\Gamma)=\omega(\Gamma')$ if $\mathrm{res}(\Gamma)=\mathrm{res}(\Gamma')$; all graphs with the same external leg structure have the same superficial degree of divergence, and only for a finite number of distinct external leg structures $\underline{r}$ will this degree indeed signify a divergence.

This leaves a finite number of external leg structures to be considered to which we restrict ourselves from now. Our first observation is that there is a natural pre-Lie algebra structure on 1PI graphs.

To this end, we define a bilinear operation

$$\Gamma_1 * \Gamma_2 = \sum_{\Gamma} n(\Gamma_1, \Gamma_2; \Gamma)\Gamma \qquad [12]$$

where the sum is over all 1PI graphs Γ. Here, $n(\Gamma_1, \Gamma_2; \Gamma)$ is a section coefficient which counts the number of ways in which a subgraph Γ_2 can be reduced to a point in Γ such that Γ_1 is obtained. The above sum is evidently finite as long as Γ_1 and Γ_2 are finite graphs, and the graphs which contribute necessarily fulfill $|\Gamma| = |\Gamma_1| + |\Gamma_2|$ and $\mathrm{res}(\Gamma) = \mathrm{res}(\Gamma_1)$.

One then has the following theorem.

Theorem 1 *The operation $*$ is pre-Lie*:

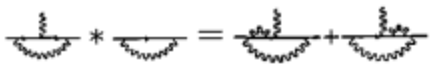
$$[\Gamma_1 * \Gamma_2] * \Gamma_3 - \Gamma_1 * [\Gamma_2 * \Gamma_3] = [\Gamma_1 * \Gamma_3] * \Gamma_2 - \Gamma_1 * [\Gamma_3 * \Gamma_2] \qquad [13]$$

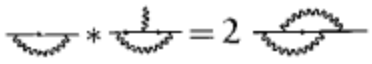
which is evident when one rewrites the $*$-product in suitable gluing operations.

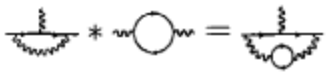
To understand this theorem, note that the equation claims that the lack of associativity in the bilinear operation $*$ is invariant under permutation of the elements indexed 2, 3. This suffices to show that the antisymmetrization of this map fulfills a Jacobi identity. Hence, we get a Lie algebra $\mathcal{L}$ by antisymmetrizing this operation:

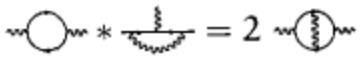
$$[\Gamma_1, \Gamma_2] = \Gamma_1 * \Gamma_2 - \Gamma_2 * \Gamma_1 \qquad [14]$$

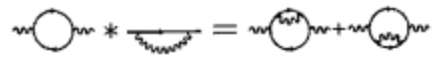
This Lie algebra is graded and of finite dimension in each degree. Let us look at a couple of examples for pre-Lie products:


[15]


[16]


[17]


[18]


[19]


[20]

Together with $\mathcal{L}$ one is led to consider the dual of its universal enveloping algebra $\mathcal{U}(\mathcal{L})$ using the theorem of Milnor and Moore. For this we use the above grading by the loop number.

This universal enveloping algebra $\mathcal{U}(\mathcal{L})$ is built from the tensor algebra

$$T = \bigoplus_k T^k, \quad T^k = \underbrace{\mathcal{L} \otimes \cdots \otimes \mathcal{L}}_{k \text{ times}} \qquad [21]$$

by dividing out the ideal generated by the relations

$$a \otimes b - b \otimes a = [a, b] \in \mathcal{L} \qquad [22]$$

Note that in $\mathcal{U}(\mathcal{L})$ we have a natural concatenation product m_*. Furthermore, $\mathcal{U}(\mathcal{L})$ carries a natural Hopf algebra structure with this product. For that, the Lie algebra $\mathcal{L}$ furnishes the primitive elements:

$$\Delta_*(a) = a \otimes 1 + 1 \otimes a, \quad \forall\, a \in \mathcal{L} \qquad [23]$$

It is, by construction, a connected finitely graded Hopf algebra which is co-commutative but not commutative. We can then consider its graded dual, which will be a Hopf algebra $\mathcal{H}(m, \mathbb{I}, \Delta, \bar{e})$ that is commutative but not cocommutative. One finds it upon using a Kronecker pairing

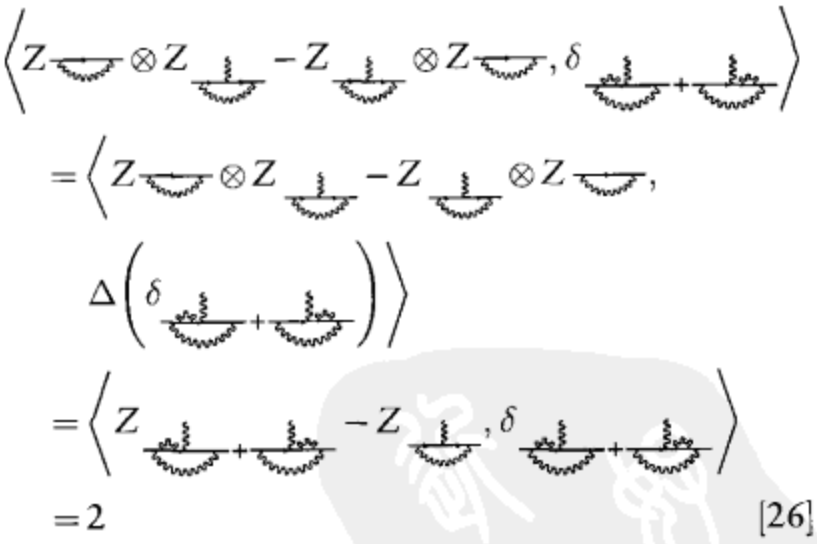
$$< Z_\Gamma, \delta_{\Gamma'} > = \begin{cases} 1, & \Gamma = \Gamma' \\ 0, & \text{else} \end{cases} \qquad [24]$$

The space of primitives of $\mathcal{U}(\mathcal{L})$ is in one-to-one correspondence with the set $\mathrm{Indec}(\mathcal{H})$ of indecomposables of $\mathcal{H}$, which is the linear span of its generators. One finds the following theorem.

Theorem 2

$$< Z_{\Gamma_1} \otimes Z_{\Gamma_2} - Z_{\Gamma_2} \otimes Z_{\Gamma_1},\ \delta_\Gamma > = < Z_{[\Gamma_2, \Gamma_1]}, \delta_\Gamma > \qquad [25]$$

For example, one finds


[26]

$\mathcal{H}$ is a graded commutative Hopf algebra which suffices to describe renormalization theory, as we see in the next section. We have formulated it for the superficially divergent 1PI graphs of the theory with the understanding that the residues of these graphs are in one-to-one correspondence with the terms in the Lagrangian of a given theory. Often, several terms in a Lagrangian correspond to graphs with the same number and type of external legs, but correspond to different form-factor projections of the graph. In such cases, the above approach can be easily adopted considering suitably colored or

labeled graphs. A similar remark applies if one desires to incorporate renormalization of superficially convergent Green functions, which requires nothing more than the consideration of an easily obtained semidirect product of the Lie algebra of superficially divergent graphs with the abelian Lie algebra of superficially convergent graphs.

The Principle of Multiplicative Subtraction

The above algebra structures are available once one has decided on the set of 1PI graphs of interest. We now use them toward the renormalization of any such chosen local quantum field theory.

From the above, 1PI graphs Γ provide the linear generators δ_Γ of the Hopf algebra $\mathcal{H} = \oplus_{i=0}^{\infty} H_i$, where $\mathcal{H}_{\text{lin}} = \text{span}(\delta_\Gamma)$ and their disjoint union provides the commutative product.

Now let Γ be a 1PI graph. We find the Hopf algebra $\mathcal{H}$ as described above to have a coproduct explicitly given as $\Delta : \mathcal{H} \to \mathcal{H} \otimes \mathcal{H}$:

$$\Delta(\Gamma) = \Gamma \otimes 1 + 1 \otimes \Gamma + \sum_{\gamma \subset \Gamma} \gamma \otimes \Gamma/\gamma \qquad [27]$$

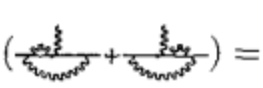
where the sum is over all unions of 1PI superficially divergent proper subgraphs, and we extend this definition to products of graphs so that we get a bialgebra.

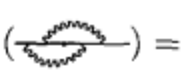
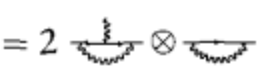
While the Lie bracket inserted graphs into each other, the coproduct disentangles them. It is this latter operation which is needed in renormalization theory: we have to render each subgraph finite before we can construct a local counter-term. That is precisely what the Hopf algebra structure maps do.

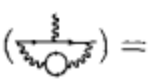
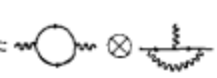
Having a coproduct, two further structure maps of $\mathcal{H}$ are immediate: the counit and the antipode. The counit $\bar{e}$ vanishes on any nontrivial Hopf algebra element, $\bar{e}(1) = 1, \bar{e}(X) = 0$. The antipode is

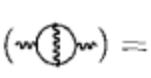
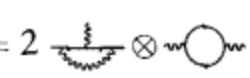
$$S(\Gamma) = -\Gamma - \sum_{\gamma \subset \Gamma} S(\gamma)\Gamma/\gamma \qquad [28]$$

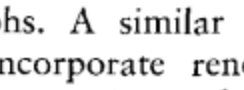
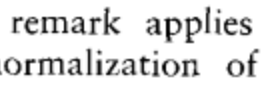
We can work out a few coproducts and antipodes as follows:

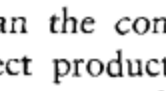
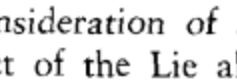
$\text{Aug}^{(2)}$() = 2  [29]

$\text{Aug}^{(2)}$() = 2  [30]

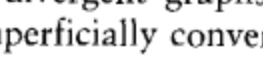
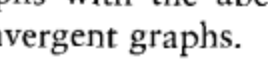
$\text{Aug}^{(2)}$() =  [31]

$\text{Aug}^{(2)}$() = 2  [32]

$\text{Aug}^{(2)}$() = 2  [33]

$\text{Aug}^{(2)}$() =  [34]

We give just one example for an antipode:

S() =  [35]

Note that for each term in the sum $\tilde{\Delta}(\Gamma) = \sum_i \Gamma'_{(i)} \otimes \Gamma''_{(i)}$, we have unique gluing data G_i such that

$$\Gamma = \Gamma''_{(i)} \leftarrow_{G_i} \Gamma'_{(i)}, \quad \forall\, i \qquad [36]$$

These gluing data describe the necessary bijections to glue the components $\Gamma'_{(i)}$ back into $\Gamma''_{(i)}$ so as to obtain Γ: using them, we can reassemble the whole from its parts. Each possible gluing can be interpreted as a composition in the insertion operad of Feynman graphs.

We have by now obtained a Hopf algebra generated by combinatorial elements, 1PI Feynman graphs. Its existence is automatic from the above choices of interactions and free fields. What remains to be done is a structural analysis of these algebras for the renormalizable theories we are confronted with in four spacetime dimensions.

The assertion underlying perturbation theory is the fact that meaningful approximations to physical observable quantities can be found by evaluating these graphs using Feynman rules.

First, as disjoint scattering processes give rise to independent amplitudes, one is led to the study of characters of the Hopf algebra, maps $\phi : \mathcal{H} \to V$ such that $\phi \circ m = m_V(\phi \otimes \phi)$.

Such maps assign to any element in the Hopf algebra an element in a suitable target space V. The study of tree-level amplitudes in lowest-order perturbation theory justifies assigning to each edge a propagator and to each elementary scattering process a vertex, which define the Feynman rules $\phi(\text{res}(\Gamma))$ and the underlying Lagrangian, on the level of residues of these very graphs. Graphs are constructed from edges and vertices which are provided precisely by the residues of those divergent graphs, hence one is led to assign to each Feynman graph an evaluation in terms of an integral over the continuous quantum numbers assigned to edges or vertices, which leads to the familiar integrals over momenta in closed loops mentioned before.

Then, with the Feynman rules providing a canonical character ϕ, we will have to make one further choice: a renormalization scheme. The need for such a choice is no surprise: after all we are eliminating short-distance singularities in the graphs,

which renders their remaining finite part ambiguous, albeit in a most interesting manner.

Hence, we choose a map $R: V \to V$, from which we obviously demand that it does not modify the UV-singular structure, and furthermore that it obeys

$$R(xy) + R(x)R(y) = R(R(x)y) + R(xR(y)) \qquad [37]$$

which guarantees the multiplicativity of renormalization and is at the heart of the Birkhoff decomposition, which emerges below: it tells us that elements in V split into two parallel subalgebras given by the image and kernel of R. Algebras for which such a map exists are known as Rota–Baxter algebras. The role Rota–Baxter algebras play for associative algebras is similar to the role Yang–Baxter algebras play for Lie algebras. The structure of these algebras allows one to connect renormalization theory to integrable systems. In addition, most of the results obtained initially for a specific renormalization scheme, such as minimal subtraction, can also be obtained, in general, upon a structural analysis of the corresponding Rota–Baxter algebras.

To see how all the above comes together in renormalization theory, we define a further character S_R^ϕ that deforms $\phi \circ S$ slightly and delivers the counter-term for Γ in the renormalization scheme R:

$$\begin{aligned} S_R^\phi(\Gamma) &= -Rm_V(S_R^\phi \otimes \phi \circ P)\Delta \\ &= -R[\phi(\Gamma)] - R\left[\sum_{\gamma \subset \Gamma} S_R^\phi(\gamma)\phi(\Gamma/\gamma)\right] \end{aligned} \qquad [38]$$

which should be compared with the undeformed

$$\begin{aligned} \phi \circ S &= m_V(S \circ \phi \otimes \phi \circ P)\Delta \\ &= -\phi(\Gamma) - \sum_{\gamma \subset \Gamma} \phi \circ S(\gamma)\phi(\Gamma/\gamma) \end{aligned} \qquad [39]$$

The fact that R is a Rota–Baxter map ensures that S_R^ϕ is an element of the character group $\mathcal{G}$ of the Hopf algebra, $S_R^\phi \in \mathrm{Spec}(\mathcal{G})$. Note that we have now determined the modified Lagrangian:

$$Z^L = S_R^\phi(\Gamma^L) \qquad [40]$$

The classical results of renormalization theory follow immediately using this group structure: we obtain the renormalization of Γ by the application of a renormalized character

$$S_R^\phi \star \phi(\Gamma) = m_V(S_R^\phi \otimes \phi)\Delta \qquad [41]$$

and Bogoliubov's $\bar{R}$ operation as

$$\begin{aligned} \bar{R}(\Gamma) &= m_V(S_R^\phi \otimes \phi)(\mathrm{id} \otimes P)\Delta(\Gamma) \\ &= \phi(\Gamma) + \sum_{\gamma \subset \Gamma} S_R^\phi(\gamma)\phi(\Gamma/\gamma) \end{aligned} \qquad [42]$$

so that

$$S_R^\phi \star \phi(\Gamma) = \bar{R}(\Gamma) + S_R^\phi(\Gamma) \qquad [43]$$

Here, $S_R^\phi \star \phi$ is an element in the group of characters of the Hopf algebra, with the group law given by the convolution

$$\phi_1 \star \phi_2 = m_V \circ (\phi_1 \otimes \phi_2) \circ \Delta \qquad [44]$$

so that the coproduct, counit, and coinverse (the antipode) give the product, unit, and inverse of this group, as befits a Hopf algebra. This Lie group has the previous Lie algebra $\mathcal{L}$ of graph insertions as its Lie algebra: $\mathcal{L}$ exponentiates to $\mathcal{G}$.

What we have achieved above is a local renormalization of quantum field theory. Let M^L be a monomial in the Lagrangian L of degree ω_L:

$$M^L = D_L\{\phi\} \qquad [45]$$

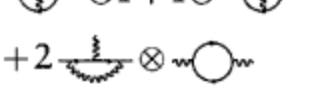
Then one can prove, using the Hochschild cohomology of $\mathcal{H}$:

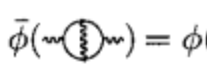
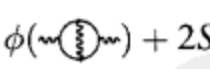
Theorem 3 (*Locality*)

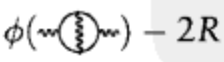
$$Z^L D_L\{\phi\} = D_L Z^L\{\phi\} \qquad [46]$$

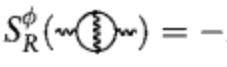
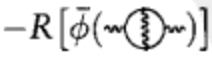
that is, renormalization commutes with infinitesimal spacetime variations of the fields.

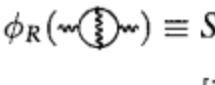
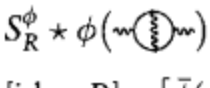
We can now work out the renormalization of a Feynman graph Γ:

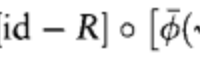
$\Delta($  $) =$  $+ 2$ [47]

$\bar{\phi}($  $) = \phi($  $) + 2S_R^\phi($ $)\phi($ $)$ [48]

$= \phi($  $) - 2R\left[\phi($ $)\right]\phi($ $)$ [49]

$S_R^\phi($  $) = -R\left[\bar{\phi}($  $)\right]$ [50]

$\phi_R($  $) \equiv S_R^\phi \star \phi($  $)$

$= [\mathrm{id} - R] \circ \left[\bar{\phi}($  $)\right]$ [51]

The formulas [47]–[51] are given in their recursive form. Zimmermann's original forest formula solving this recursion is obtained when we trace our considerations back to the fact that the coproduct can be written in nonrecursive form as a sum over forests, and similarly for the antipode.

Diffeomorphisms of Physical Parameters

In the above, we have effectively obtained a Birkhoff decomposition of the Feynman rules $\phi \in \mathrm{Spec}(\mathcal{G})$ into two characters – $\phi_+^R = S_R^\phi \star \phi \in \mathrm{Spec}(\mathcal{G})$ and $\phi_-^R = S_R^\phi \in \mathrm{Spec}(\mathcal{G})$ – for any Rota–Baxter map R. Thanks to Atkinson's theorem, this is possible for any renormalization scheme R. For the minimal subtraction scheme, it amounts to the decomposition of the Laurent series $\phi(\Gamma)(\epsilon)$, which has poles of finite order in the regulator ϵ, into a part holomorphic at the origin and a part holomorphic at complex infinity. This has a particularly nice geometric interpretation upon considering the Birkhoff decomposition of a loop around the origin, providing the clutching data for the two half-spheres defined by that very loop.

Whilst in this manner a satisfying understanding of perturbative renormalization is obtained, the character group $\mathcal{G}$ remains rather poorly understood. On the other hand, renormalization can be captured by the study of diffeomorphisms of physical parameters as, by definition, the range of allowed modification in renormalization theory is determined by the variation of the coefficients of monomials $\hat{\phi}(\underline{r})$ of the underlying Lagrangian

$$L = \sum_{\underline{r}} Z^{\underline{r}} \hat{\phi}(\underline{r}) \qquad [52]$$

Thus, one desires to obtain the whole Birkhoff decomposition at the level of diffeomorphisms of the coupling constants.

The crucial step toward that goal is to realize the role of a standard quantum field-theoretic formula of the form

$$g_{new} = g_{old} Z^g \qquad [53]$$

where

$$Z^g = \frac{Z^v}{\prod_{e \in \mathrm{res}(v)_{\mathrm{ext}}^{[1]}} \sqrt{Z^e}} \qquad [54]$$

for some vertex v, which obtains the new coupling in terms of a diffeomorphism of the old. This formula provides, indeed, a Hopf algebra homomorphism from the Hopf algebra of diffeomorphisms to the Hopf algebra of Feynman graphs, regarding Z^g (a series over counter-terms for all 1PI graphs with the external leg structure corresponding to the coupling g), in two different ways: it is, at the same time, a formal diffeomorphism in the coupling constant g_{old} and a formal series in Feynman graphs. As a consequence, there are two competing coproducts acting on Z_g. That both give the same result defines the required homomorphism, which transposes to a homomorphism from the largely unknown group of characters of $\mathcal{H}$ to the one-dimensional diffeomorphisms of this coupling.

In summary, one finds that a couple of basic facts enable one to make a transition from the abstract group of characters of a Hopf algebra of Feynman graphs (which, incidentally, equals the Lie group assigned to the Lie algebra with universal enveloping algebra the dual of this Hopf algebra) to the rather concrete group of diffeomorphisms of physical observables. These steps are given as follows:

- Recognize that Z factors are given as counter-terms over a formal series of graphs starting with 1, graded by powers of the coupling, hence invertible.
- Recognize the series Z_g as a formal diffeomorphism, with Hopf algebra coefficients.
- Establish that the two competing Hopf algebra structures of diffeomorphisms and graphs are consistent in the sense of a Hopf algebra homomorphism.
- Show that this homomorphism transposes to a Lie algebra and hence Lie group homomorphism.

The effective coupling $g_{\mathrm{eff}}(\varepsilon)$ now allows for a Birkhoff decomposition in the space of formal diffeomorphisms.

Theorem 4 *Let the unrenormalized effective coupling constant $g_{\mathrm{eff}}(\varepsilon)$ viewed as a formal power series in g be considered as a loop of formal diffeomorphisms and let $g_{\mathrm{eff}}(\varepsilon) = (g_{\mathrm{eff}_-})^{-1}(\varepsilon)\, g_{\mathrm{eff}_+}(\varepsilon)$ be its Birkhoff decomposition in the group of formal diffeomorphisms. Then the loop $g_{\mathrm{eff}_-}(\varepsilon)$ is the bare coupling constant and $g_{\mathrm{eff}_+}(0)$ is the renormalized effective coupling.*

The above results hold as they stand for any massless theory which provides a single coupling constant. If there are multiple interaction terms in the Lagrangian, one finds similar results relating the group of characters of the corresponding Hopf algebra to the group of formal diffeomorphisms in the multidimensional space of coupling constants.

The Role of Hochschild Cohomology

The Hochschild cohomology of the combinatorial Hopf algebras which we discuss here plays three major roles in quantum field theory:

1. it allows one to prove locality from the accompanying filtration by the augmentation degree coming from the kernels $\ker \mathrm{Aug}^{(k)}$;

2. it allows one to write the quantum equations of motion in terms of the Hopf algebra primitives, elements in $\mathcal{H}_{\text{lin}} \cap \{\ker \text{Aug}^{(2)}/\ker \text{Aug}^{(1)}\}$; and
3. it identifies the relevant sub-Hopf algebras formed by time-ordered products.

Before we discuss these properties, let us first introduce the relevant Hochschild cohomology.

Hochschild Cohomology of Bialgebras

Let $(A, m, \mathbb{I}, \Delta, \epsilon)$ be a bialgebra, as before. We regard linear maps $L: A \to A^{\otimes n}$ as n-cochains and define a coboundary map $b, b^2 = 0$ by

$$bL := (\text{id} \otimes L) \circ \Delta + \sum_{i=1}^{n} (-1)^i \Delta_i \circ L + (-1)^{n+1} L \otimes \mathbb{I} \qquad [55]$$

where Δ_i denotes the coproduct applied to the ith factor in $A^{\otimes n}$, which defines the Hochschild cohomology of A.

For the case $n = 1$, for $L: A \to A$, [55] reduces to

$$bL = (\text{id} \otimes L) \circ \Delta - \Delta \circ L + L \otimes \mathbb{I} \qquad [56]$$

The category of objects (A, C), which consists of a commutative bialgebra A and a Hochschild 1-cocycle C on A, has an initial object $(\mathcal{H}_{\text{rt}}, B_+)$, where $\mathcal{H}_{\text{rt}}$ is the Hopf algebra of (nonplanar) rooted trees, and the closed but nonexact 1-cocycle B_+ grafts a product of rooted trees together at a new root as described below.

The higher ($n > 1$) Hochschild cohomology of $\mathcal{H}_{\text{rt}}$ vanishes, but in what follows, the closedness of B_+ will turn out to be crucial.

The Hopf Algebra of Rooted Trees

A rooted tree is a simply connected contractible compact graph with a distinguished vertex, the root. A forest is a disjoint union of rooted trees. Isomorphisms of rooted trees or forests are isomorphisms of graphs preserving the distinguished vertex/vertices. Let t be a rooted tree with root o. The choice of o determines an orientation of the edges of t, away from the root, say. Forests are graded by the number of vertices they contain.

Let $\mathcal{H}_{\text{rt}}$ be the free commutative algebra generated by rooted trees. The commutative product in $\mathcal{H}_{\text{rt}}$ corresponds to the disjoint union of trees, such that monomials in $\mathcal{H}_{\text{rt}}$ are scalar multiples of forests. We demand that the linear operator B_+ on $\mathcal{H}_{\text{rt}}$, defined by

$$B_+(\mathbb{I}) = \bullet \qquad [57]$$

$$B_+(t_1 \dots t_n) = \qquad [58]$$



is a Hochschild 1-cocycle, which makes $\mathcal{H}_{\text{rt}}$ a Hopf algebra. The resulting coproduct can be described as follows:

$$\Delta(t) = \mathbb{I} \otimes t + t \otimes \mathbb{I} + \sum_{adm\ c} P_c(t) \otimes R_c(t) \qquad [59]$$

where the sum goes over all admissible cuts of the tree t. Such a cut of t is a nonempty set of edges of t that are to be removed. The forest which is disconnected from the root upon removal of those edges is denoted by $P_c(t)$ and the part which remains connected to the root is denoted by $R_c(t)$. A cut $c(t)$ is admissible if, for each vertex l of t, it contains at most one edge on the path from l to the root.

This Hopf algebra of nonplanar rooted trees is the universal object after which all such commutative Hopf algebras $\mathcal{H}$ providing pairs $(\mathcal{H}, B)$, for B a Hochschild 1-cocycle, are formed.

Theorem 5 *The pair $(\mathcal{H}_{\text{rt}}, B_+)$, unique up to isomorphism, is universal among all such pairs. In other words, for any pair $(\mathcal{H}, B)$ where $\mathcal{H}$ is a commutative Hopf algebra and B a closed nonexact 1-cocycle, there exists a unique Hopf algebra morphism $\mathcal{H}_{\text{rt}} \xrightarrow{\rho} \mathcal{H}$ such that $B \circ \rho = \rho \circ B_+$.*

This theorem suggests that we investigate the Hochschild cohomology of the Hopf algebras of 1PI Feynman graphs. It clarifies the structure of 1PI Green functions.

The Roles of Hochschild Cohomology

The Hochschild cohomology of the Hopf algebras of 1PI graphs sheds light on the structure of 1PI Green function in at least four different ways:

- it gives a coherent proof of locality of counterterms – the very fact that

$$[Z^L, D_{\underline{r}}] = 0 \qquad [60]$$

  means that the coefficients in the Lagrangian remain independent of momenta, and hence the Lagrangian remains a polynomial expression in fields and their derivatives;
- the quantum equation of motions takes a very succinct form, identifying the Dyson kernels with the primitives of the Hopf algebra;
- sub-Hopf algebras emerge from the study of the Hochschild cohomology, which connects the representation theory of these Hopf algebras to the structure of theories with internal symmetries; and
- these Hopf algebras are intimately connected to the structure of transcendental functions, such as

the generalized polylogarithms, which play a prominent role these days ranging from applied particle physics to recent developments in mathematics.

To determine the Hochschild 1-cocycles of some Feynman graph Hopf algebra $\mathcal{H}$, one determines first the primitives graphs γ of the Hopf algebra, which, by definition, fulfill the condition

$$\Delta(\gamma) = \gamma \otimes \mathbb{I} + \mathbb{I} \otimes \gamma \qquad [61]$$

Using the pre-Lie product above, one then determines the maps

$$B_+^\gamma : \mathcal{H} \to \mathcal{H}_{\text{lin}} \qquad [62]$$

such that

$$B_+^\gamma(h) = B_+^\gamma(h) \otimes \mathbb{I} + (\text{id} \otimes B_+^\gamma)\Delta(h) \qquad [63]$$

where $B_+^\gamma(h) = \sum_\Gamma n(\gamma, h, \Gamma)\Gamma$. The coefficients $n(\gamma, h, \Gamma)$ are closely related to the section coefficients noted earlier.

Using the definition of the Bogoliubov map $\bar{\phi}$, this immediately shows that

$$S_R^\phi(B_+^\gamma(h)) = \int D_\gamma \leftarrow_{G_i} \phi_R(h) \qquad [64]$$

which proves locality of counter-terms upon recognizing that B_+^γ increases the augmentation degree. Here, the insertion of the functions for the subgraph is achieved using the relevant gluing data of [36].

To recover the quantum equation of motions from the Hochschild cohomology, one proves that

$$\Gamma^r = 1 + \sum_\gamma \frac{g^\gamma}{\text{Sym}(\gamma)} B_+^\gamma(X_\gamma) \qquad [65]$$

where

$$X_\gamma = \prod_{e \in \gamma^{[1]}_{\text{int}}} \prod_{v \in \gamma^{[0]}} \frac{\Gamma^v}{\Gamma^e} \qquad [66]$$

has the required solution. Upon application of the Feynman rules, the maps B_+^γ turn into the integral kernels of the usual Dyson–Schwinger equations. This allows for new nonperturbative approaches which are a current theme of investigation.

Finally, we note that the 1-cocycles introduced above allow one to determine sub-Hopf algebras of the form

$$\Delta(c_n^r) = \sum P(\{c_j^s\}) \otimes c_j^r \qquad [67]$$

where the c_j^s are defined in eqn [3]. These algebras do not necessitate the considerations of single Feynman graphs any longer, but allow one to establish renormalization directly for the sum of all graphs at a given loop order. Hence, they establish a Hopf algebra structure on time-ordered products in momentum space. For theories with internal symmetries, one expects and indeed finds that the existence of these subalgebras establishes relations between graphs that are same as the Slavnov–Taylor identities between the couplings in the Lagrangian.

Outlook

Thanks to the Hopf and Lie algebra structures described above, quantum field theory has started to reveal its internal mathematical structure in recent years, which connects it to a motivic theory and arithmetic geometry. Conceptually, quantum field theory has been the most sophisticated means by which a physicist can describe the character of the physical law. We have slowly begun to understanding that, in its short-distance singularities, it encapsulates concepts of matching beauty. We can indeed expect local point-particle quantum field theory to remain a major topic of mathematical physics investigations in the foreseeable future.

See also: Bicrossproduct Hopf Algebras and Noncommutative Spacetime; Exact Renormalization Group; Hopf Algebras and q-Deformation Quantum Groups; Number Theory in Physics; Operads; Perturbation Theory and Its Techniques; Renormalization: General Theory; von Neumann Algebras: Introduction, Modular Theory, and Classification Theory.

Further Reading

Bloch S, Esnault H, and Kreimer D (2005) *On Motives associated to graph polynomials.*

Connes A and Kreimer D (1998) Hopf algebras, renormalization and noncommutative geometry. *Communications in Mathematical Physics* 199: 203 (arXiv:hep-th/9808042).

Connes A and Kreimer D (2001) Renormalization in quantum field theory and the Riemann–Hilbert problem. II: The beta-function, diffeomorphisms and the renormalization group. *Communications in Mathematical Physics* 216: 215 (arXiv:hep-th/0003188).

Connes A and Marcolli M (2004) From physics to number theory via noncommutative geometry. II: Renormalization, the Riemann–Hilbert correspondence, and motivic Galois theory, arXiv:hep-th/0411114.

Kreimer D (2003) New mathematical structures in renormalizable quantum field theories. *Annals of Physics* 303: 179 (arXiv:hep-th/0211136).

Kreimer D (2004) The residues of quantum field theory: numbers we should know, arXiv:hep-th/0404090.

Kreimer D (2005) *Anatomy of a gauge theory.*

Indefinite Metric

H Gottschalk, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

If, in a problem of quantization, state spaces with indefinite inner product are used instead of Hilbert spaces, one speaks of quantization with indefinite metric. The main domain of application is the quantization of gauge fields, like the electromagnetic vector potential $A_\mu(x)$ or Yang–Mills fields in quantum chromodynamics (QCD) and the standard model.

The conceptual problem with the indefinite metric is the occurrence of senseless negative probabilities in the formalism. Such negative probabilities, however, only arise in expectation values of fields that are not gauge invariant and hence do not correspond to observable quantities. Equivalently, the inner product of vectors generated by application of such fields to the vacuum vector with itself can be negative or null. In order to extract the observable content of an indefinite-metric quantum theory, a subsidiary condition is needed to single out the physical subspace. Restricted to this subspace, the inner product is positive semidefinite. This subsidiary condition can be seen as the implementation of a gauge, as, for example, the Lorentz gauge $\partial_\mu A^\mu(x) = 0$ in quantum electrodynamics (QED). This procedure is also known under the name Gupta–Bleuler formalism.

The use of indefinite metric in the quantization of gauge theories like QED can be avoided entirely.

Notes:

This is called quantization in a physical gauge. The problem with such gauges is that they are not Lorentz invariant and that the vector potential $A^\mu(x)$ is not a local field. An example is the Coulomb gauge defined by $A_0(x)=0$ and $\partial^i A_i(x)=0$ in QED. Furthermore, Dirac spinor fields $\psi(x)$ in such gauges do not anticommute when localized in spacelike separated regions. The Dirac fields therefore are also nonlocal quantities. Although not in contrast with special relativity, as Dirac spinors and the vector potential are not gauge invariant and hence are unobservable, this leads to severe technical problems in the formulation of interacting theories. In particular, the theory of renormalization heavily uses both locality and invariance. Therefore, the Gupta–Bleuler formalism generally is the preferred quantization procedure for a gauge theory.

That a local and invariant quantization is not possible using a (positive-metric) Hilbert space has been proved by F Strocchi in a series of articles published between 1967 and 1970. If one wants to preserve locality and/or invariance of the quantized field theory, it is thus strictly necessary to give up the positivity of the state space.

A short digression into the early history of the idea might be of interest. It dates back to 1941, where the use of indefinite metric in the quantization of relativistic equations was proposed by Paul Dirac in a lecture at the London Royal Society. The negative probabilities for the bosonic vector potential were thought to be connected with the problem of negative-energy solutions of relativistic equations as a type of surrogate of the "Dirac sea" in the quantization of fermions. Furthermore, Dirac proposed that negative-energy solutions and negative probabilities would jointly lead to the cancellation of divergences in QED. The latter idea was taken up by W Heisenberg in his lectures on the theory of elementary particles held in Munich in 1961, but the generally accepted solution to the problem of ultraviolet divergences was achieved without recourse to Dirac's original motivation. In 1950 the consistent quantization of vector potential in the Lorentz gauge was formulated by S N Gupta and K Bleuler eliminating the use of negative-energy solutions. Since then the indefinite metric has become a building block of the standard theory of quantized gauge fields.

No-Go Theorems

The strict necessity of the Gupta–Bleuler procedure for the local or covariant quantization of gauge fields has been demonstrated by F Strocchi in the form of no-go theorems for positive metric. Here we review their content for the case of the electromagnetic field. Related statements can be obtained for nonabelian gauge theories. The main problem lies in the fact that standard assumptions on the quantization of relativistic fields are in conflict with Maxwell equations that should hold as operator identities in a positive-metric theory containing no unobservable states. Let

$$F_{\nu\mu}(x) = \partial_\mu A_\nu(x) - \partial_\nu A_\mu(x) \qquad [1]$$

be the quantized electromagnetic field strength tensor. Classically, the existence of $A_\mu(x)$ is guaranteed from the first set of Maxwell equations $\epsilon^{\alpha\beta\nu\mu}\partial_\beta F_{\nu\mu}(x)=0$. Here (and henceforth) indices are raised and lowered with respect to the Minkowski metric $g_{\alpha\beta}$ and $\epsilon^{\alpha\beta\mu\nu}$ is the completely antisymmetric tensor on $\mathbb{R}^d$. Furthermore, we apply Einstein's convention on summation over repeated upper and lower indices. Standard assumptions from axiomatic quantum field theory are:

1. The field strength tensor $F_{\nu\mu}(x)$ is an operator-valued distribution acting on a (dense core of a) Hilbert space $\mathcal{H}$ with scalar product $\langle .,. \rangle$ – in the indefinite-metric case, $\langle .,. \rangle$ only needs to be an inner product.
2. $F_{\mu\nu}(x)$ transforms covariantly, that is, there is a strongly continuous unitary (with respect to $\langle .,. \rangle$) representation U of the orthochronous, proper Poincaré group on $\mathcal{H}$ such that for translation $a \in \mathbb{R}^d$ combined with a restricted Lorentz transformation Λ, one has

$$U(a,\Lambda)F_{\mu\nu}(x)U(a,\Lambda)^{-1} = (\Lambda^{-1})^\rho{}_\mu(\Lambda^{-1})^\kappa{}_\nu F_{\rho\kappa}(\Lambda x + a) \qquad [2]$$

3. There exists a unique (up to multiplication with C-numbers) translation invariant vector $\Omega \in \mathcal{H}$ (the "vacuum"), that is, $U(a,1)\Omega = \Omega \forall a \in \mathbb{R}^d$.
4. The representation of the translations fulfills the spectral condition

$$\int_{\mathbb{R}^4} \langle \Phi, U(a,1)\Psi \rangle e^{ip\cdot a}\, da = 0 \qquad [3]$$

$\forall \Psi, \Phi \in \mathcal{H}$ if p is not in the closed forward light cone $\bar{V}^+ = \{p \in \mathbb{R}^4 : p\cdot p \geq 0, p^0 \geq 0\}$. Here the dot is the Minkowski inner product.
So far the assumptions concerned only observable quantities. In the following, we also demand.
5. The vector potential $A_\mu(x)$ is realized as an operator-valued distribution on $\mathcal{H}$ and transforms covariantly under translations

$$U(a,1)A_\mu(x)U(a,1)^{-1} = A_\mu(x+a) \qquad [4]$$

The assumptions on the nature of the vector potential so far are rather weak. Strocchi's no-go theorems show that one cannot add further desirable properties as Lorentz covariance and/or locality without getting into conflict with the Maxwell equations:

Theorem 1 *Suppose that the above assumptions* (1)–(3) *and* (5) *hold. If Maxwell's equations in the absence of charges,*

$$\epsilon^{\alpha\beta\nu\mu}\partial_\beta F_{\nu\mu}(x) = 0, \quad \partial^\mu F_{\mu\nu}(x) = 0 \qquad [5]$$

are valid as δ *operator identities on* $\mathcal{H}$ *and the gauge potential transforms covariantly*

$$U(a,\Lambda)A_\mu(x)U(a,\Lambda)^{-1} = (\Lambda^{-1})^\nu{}_\mu A_\nu(\Lambda x + a) \qquad [6]$$

the two-point function of the electromagnetic field tensor vanishes identically:

$$\langle\Omega, F_{\nu\mu}(x)F_{\kappa\rho}(y)\Omega\rangle = 0 \quad \forall\, x, y \in \mathbb{R}^4 \qquad [7]$$

To gain a better understanding, where the difficulties in the quantization of the Maxwell equations arise from, here is a rough sketch of the proof: Maxwell equations and covariance imply that $f_{\mu\nu\rho}(x-y) = \langle\Omega, A_\mu(x)F_{\nu\rho}(y)\Omega\rangle$ fulfills $\partial^\alpha\partial_\alpha f_{\mu\nu\rho}(x) = 0$ and hence its Fourier transform has support in the union of the forward and backward light cone. The Fourier transform thus can be split into a positive- and a negative-frequency part, and $f_{\mu\nu\rho} = f^+_{\mu\nu\rho} + f^-_{\nu\mu\rho}$ accordingly. By the general analysis of axiomatic field theory (*see* Axiomatic Quantum Field Theory), the functions $f^\pm_{\nu\mu\rho}$ are boundary values of complex analytic functions on certain tubar domains $\mathcal{T}^\pm$ transforming covariantly under a certain representation of the complex Lorentz group. By a theorem of Araki and Hepp giving a general representation of such functions and using the antisymmetry of the field tensor, the following formula can be derived:

$$f^\pm_{\mu\nu\rho}(z) = (g_{\mu\rho}\partial_\nu - g_{\mu\nu}\partial_\rho)f^\pm(z) + \epsilon_{\mu\nu\rho\alpha}\partial^\alpha h^\pm(z)$$
$$z \in \mathcal{T}^\pm \qquad [8]$$

with $f^\pm, h^\pm$ invariant under complex Lorentz transformations. Taking boundary values in $\mathcal{T}^\pm$, one obtains $f_{\mu\nu\rho} = (g_{\mu\rho}\partial_\nu - g_{\mu\nu}\partial_\rho)f + \epsilon_{\mu\nu\rho\alpha}\partial^\alpha h$, with $f = \bar{f}^+ + \bar{f}^-$ and $h = \bar{h}^+ + \bar{h}^-$, where the bar stands for the distributional boundary value. Maxwell's equations imply $\partial^\nu f_{\mu\nu\rho} = (\partial^\nu\partial_\nu g_{\mu\rho} - \partial_\mu\partial_\rho)f = 0$ and $\epsilon_\alpha{}^{\beta\nu\rho}\partial_\beta f_{\mu\nu\rho} = (\partial^\nu\partial_\nu g_{\alpha\mu} - \partial_\alpha\partial_\mu)h = 0$. The only Lorentz-invariant solutions to these equations are constant, which implies the statement of Theorem 1.

The second no-go theorem eliminates the assumption that the vector potential $A_\mu(x)$ is covariant; however, a local gauge is assumed. The result is the same as in Theorem 1:

Theorem 2 *Suppose that the above assumptions* (1)–(5) *and Maxwell's equations hold as operator identities on* $\mathcal{H}$. *If, furthermore, the gauge is local, that is,*

$$[A_\mu(x), A_\nu(y)] = 0 \quad \textit{if } x - y \textit{ is spacelike} \qquad [9]$$

the two-point function of the field strength tensor vanishes again as in Theorem 1.

Analyzing the interplay of the covariance properties of $F_{\mu\nu}(x)$ with the locality of $A_\mu(x)$, Strocchi was able to show that the function $f_{\mu\nu\rho}(x-y)$ must have the same covariance properties as in Theorem 1, which implies the assertion of Theorem 2.

The first two no-go theorems deal with the free electromagnetic field that is not coupled to charge-carrying fields. This is, of course, already a real obstruction also for an interacting theory, since, by the LSZ formalism, one expects the asymptotic incoming and outgoing fields $A^{\text{in/out}}_\mu(x), F^{\text{in/out}}_{\mu\nu}(x)$ to be free. In fact, it has been proved by D Buchholz that, in the positive-metric case, such asymptotic fields can always be constructed. If one assumes a local and covariant gauge and positivity, the vanishing of the two-point function would also imply that the field $F_{\mu\nu}(x) = 0$ identically by the Reeh–Schlieder theorem.

The next no-go theorem shows that the problems connected to the quantization of the Maxwell equations are not connected only to the free electromagnetic fields. Let us assume that the second set of Maxwell equations is given by

$$\partial^\mu F_{\mu\nu}(x) = j_\nu(x) \qquad [10]$$

where j_ν is the leptonic current, that is, $j_\nu(x) = e:\psi^\dagger(x)\gamma_\nu\psi(x):$ in the case of QED, where ψ is the quantized Dirac field associated with electrons and positrons. Here $:\cdot:$ stands for Wick ordering and γ_ν are the Dirac matrices, $\psi^\dagger = \psi^*\gamma^0$. The conservation of the current $\partial^\nu j_\nu(x) = 0$ implies that the current charge

$$Q_C = \lim_{R\to\infty}\int_{\mathbb{R}^3}\int_{\mathbb{R}} \alpha(x^0)\chi(x/R)j_0(x^0, x)\,dx^0\,dx \qquad [11]$$

is a constant of motion, where α and χ are compactly supported infinitely differentiable functions with $\int_{\mathbb{R}}\alpha(x^0) = 1$ and $\chi(x) = 1$ for $|x| < 1$. Now, an alternative definition of charge, called gauge charge (it generates the global U(1)-gauge transformation), is given by

$$Q_G\Omega = 0, \quad [Q_C, A_\mu(x)] = 0 \quad \text{and}$$
$$[Q_G, \psi(x)] = -e\psi(x) \qquad [12]$$

A third formulation of charge, the Maxwell charge Q_M, can also be given by replacing $j^0(x)$ in [11] by $\partial_\nu F^{\nu 0}(x)$. Obviously, if Maxwell equations hold as operator identities, $Q_C = Q_M$. On observable states, all charges Q_M, Q_C, and Q_G ought to coincide. Strocchi's third theorem shows that this cannot be achieved within a local gauge:

Theorem 3 *If the Maxwell equations* [9] *hold and the Dirac field* $\psi(x)$ *is local with respect to the electromagnetic field tensor* $F_{\mu\nu}(x)$, *that is,*

$$[F_{\mu\nu}(x), \psi(y)] = 0 \quad \textit{if } x - y \textit{ is spacelike} \qquad [13]$$

then $[Q_M, \psi(x)] = 0$, *hence* $Q_C = Q_M \neq Q_C$.

The proof is a simple consequence of the observation that $j_0(x) = \partial^\nu F_{\nu 0}(x) = \partial^i F_{i0}(x)$ is a three-divergence as $F_{00}(x) = 0$ by antisymmetry of $F_{\mu\nu}(x)$. Hence,

$$\begin{aligned}[Q_C, \psi(y)] &= \lim_{R\to\infty} \int_{\mathbb{R}^4} [j_0(x), \psi(y)]\alpha(x^0)\chi(x/R)\,dx^0\,dx \\ &= -\lim_{R\to\infty} \int_{\mathbb{R}^4} [F_{i0}(x), \psi(y)]\alpha(x^0)\partial^i\chi(x/R) \\ &\quad \times dx^0\,dx = 0 \end{aligned} \qquad [14]$$

since, for R sufficiently large, the support of $\alpha(x^0)\partial_i\chi(x/R)$ becomes spacelike separated from y.

It should be noted that the proof of none of the above theorems relies on the definiteness of the inner product. The main clue of the indefinite-metric formalism, therefore, is rather to give up Maxwell equations as operator identities. In the usual positive-metric formalism, where all states in $\mathcal{H}$ are physical states, this would not be legitimate. But in indefinite metrics, many states are unobservable – in particular, those with negative "norm" $\langle \Psi, \Psi \rangle < 0$. On such states we can neglect the Maxwell equations.

Axiomatic Framework

The formalism of axiomatic quantum field theory (*see* Axiomatic Quantum Field Theory) requires a revision in order to cover the case of gauge fields. The necessary adaptations have been elaborated by G Morchio and F Strocchi, but also earlier work of E Scheibe and J Yngvasson played a significant role in this development.

Let $\phi(x)$ be a V'-valued quantum field, where V is a finite-dimensional $\mathbb{C}$-vector space with involution $*$. The prime stands for the (topological) dual. For the case of QED, V is eight dimensional, containing four dimensions for the vector potential $A_\mu(x)$ and another four for the Dirac spinors $\psi(x), \psi^\dagger(x)$.

Such a quantum field can be reconstructed from its vacuum expectation values (Wightman functions) as follows: let $S_1 = S(\mathbb{R}^4, V)$ be the space of rapidly decreasing functions $f: \mathbb{R}^4 \to V$ endowed with the Schwarz topology. Then the Borcher's algebra $\underline{S}$ be the free, unital, involutive tensor algebra over S_1, that is, $\underline{S} = \mathbb{C}\mathbf{1} \oplus_{n\geq 0} S_1^{\otimes n}$ with the multiplication induced by the tensor product and involution $(f_1 \otimes \cdots \otimes f_n)^* = f_n^* \otimes \cdots \otimes f_1^*$. $\underline{S}$ is endowed with the direct-sum topology. One can show that any linear, normalized, continuous functional $\underline{W}: \underline{S} \to \mathbb{C}$, $W(1) = 1$, is determined by its restrictions W_n to $S_1^{\otimes n}$. By the Schwarz kernel theorem, $W_n \in \mathcal{S}'(\mathbb{R}^{4n}, V^{\otimes n})$. Conversely, any such sequence of Wightman distributions W_n determines a $\underline{W}$.

Given a Hermitian Wightman functional $\underline{W}$ such that $W(f^*) = \overline{W(f)}, \forall f \in \underline{S}, \mathcal{L}_{\underline{W}} = \{f \in \underline{S}: \underline{W}(\underline{h}^* \otimes f) = 0\ \forall \underline{h} \in \underline{S}\}$ forms a left ideal and the inner product $\underline{W}(f^* \otimes \underline{h})$ induces a nondegenerate inner product $\langle ., . \rangle$ on $\mathcal{H}_0 = \underline{S}/\mathcal{L}_{\underline{W}}$. Furthermore, Borchers' algebra $\underline{S}$ acts from the left on $\mathcal{H}_0$. The quantum field $\phi(x)$ defined as the restriction of this canonical representation to the space $S_1 \subset \underline{S}$ according to $\phi(f) =$ "$\int_{\mathbb{R}^4} \phi^a(x) f_a(x) dx$" where the index a runs over a basis of V.

If the Wightman functional $\underline{W}$ has further properties from axiomatic QFT (*see* Axiomatic Quantum Field Theory) like invariance with respect to a given representation of the Lorentz group on V, translation invariance, locality, and the spectral property, the quantum field $\phi(x)$ fulfills the related requirements in analogy with the items (1)–(5) listed in the previous section for the case of the vector potential $A_\mu(x)$. The Wightman distributions W_n as in the positive-metric case are related to the vacuum expectation values of the theory by

$$W_n^{a_1,\ldots,a_n}(x_1, \ldots, x_n) = \langle \Omega, \phi^{a_1}(x_1) \cdots \phi^{a_n}(x_n)\Omega \rangle \qquad [15]$$

where Ω is the equivalence class of $\mathbf{1}$ in $\mathcal{H}_0$.

The state-space $\mathcal{H}_0$ produced by the Gelfand–Naimark–Segal (GNS) construction for inner-product spaces might be too small to contain all states of physical interest. For example, in the QED case, it does not contain charged states (cf. Theorem 3). Depending on the physical problem, one might also be interested in constructing coherent or scattering states and translation-invariant states apart from the vacuum. Such states appear in problems related to symmetry breaking and confinement (the so-called Θ-vacua) or in some problems of conformal QFT (*see* Boundary Conformal Field

Theory) in two dimensions. It, therefore, has become the standard point of view that one needs to make a suitable closure of $\mathcal{H}_0$ such that this closure includes the states of interest (for an alternative point of view, see the last paragraph of the following section).

Typically, larger closures are favorable, as they contain more states. One therefore focuses on maximal Hilbert closures of $\mathcal{H}_0$. A Hilbert topology τ is induced by an auxiliary scalar product $(.,.)$ on $\mathcal{H}_0$. It is admissible, if it dominates the indefinite inner product $|\langle\Phi,\Psi\rangle|^2 \leq C(\Psi,\Psi)(\Phi,\Phi)\,\forall\Psi,\Phi\in\mathcal{H}_0$ for some $C>0$. This guarantees that the inner product extends to the Hilbert space closure $\mathcal{H}$ of $\mathcal{H}_0$ with respect to τ. Furthermore, there exists a self-adjoint contraction η on $\mathcal{H}$ such that $\langle\Psi,\eta\Phi\rangle = (\Psi,\eta\Phi)\,\forall\Phi,\Psi\in\mathcal{H}$. A Hilbert topology τ is maximal if there is no admissible Hilbert topology τ' that is strictly weaker than $\mathcal{H}_0$. The classification of maximal admissible Hilbert topologies in terms of the metric operator η is given by the following theorem:

Theorem 4 *A Hilbert topology τ on $\mathcal{H}_0$ generated by a scalar product $(.,.)$ is maximal if and only if the metric operator η has a continuous inverse η^{-1} on the Hilbert space closure $\mathcal{H}$ of $\mathcal{H}_0$. In that case, one can replace $(.,)$ by the scalar product $(\Psi,\Phi)_1=(\Psi,|\eta|\Phi)$ without changing the topology τ. The new metric operator η_1 then fulfills $\eta_1^2=1_{\mathcal{H}}$.*

For a proof of the first statement, see the original work of Morchio and Strocchi (1980). One can easily check that $\eta_1=\eta|\eta^{-1}|$ which implies the second assertion of the theorem. A Hilbert space $(\mathcal{H},(.,.))$ with an indefinite inner product induced by a metric operator η with $\eta^2=1_{\mathcal{H}}$ is called a Krein space. For an extensive study of Krein spaces, see the monograph by Azizov and Iokhvidov (1989).

Furthermore, one can show that given a nonmaximal admissible Hilbert space topology τ induced by some $(.,.)$, one obtains a maximal admissible Hilbert topology as follows: given the metric operator η, we define a scalar product $(\Psi,\Phi)_1=(\Psi,(1-P_0)\Phi)$ on $\mathcal{H}$ with P_0 the null space projector of η. Obviously, this scalar product is still admissible and it leads to a new metric operator η_1 and a new closure $\mathcal{H}_1$ of $\mathcal{H}_0$. Furthermore, it is easy to show that the scalar product $(\Psi,\Phi)_2=(\Psi,|\eta_1|\Phi)_1$ still induces an admissible Hilbert topology which is also maximal, as $\eta_2=\eta_1|\eta_1^{-1}|$ clearly fulfills the Krein relation $\eta_2^2=1_{\mathcal{H}_2}$.

The question of the existence of a Krein space closure of $\mathcal{H}_0$, therefore, reduces to the question of the existence of an admissible Hilbert topology on $\mathcal{H}_0$. The following condition on the Wightman functions W_n replaces the positivity axiom in the case of indefinite-metric quantum fields:

Theorem 5 *Given a Wightman functional $\underline{W}$, there exists an admissible Hilbert space topology τ on $\mathcal{H}_0=\underline{\mathcal{S}}/\mathcal{L}_{\underline{W}}$ if and only if there exists a family of Hilbert seminorms p_n on $\mathcal{S}_n$ such that $|W_{n+n}(f\otimes h)|\leq p_n(f)p_m(h),\forall n,m\in\mathbb{N}_0,f\in\mathcal{S}_n,h\in\mathcal{S}_m$.*

In some cases, covering also examples with nontrivial scattering in arbitrary dimension, the condition of Theorem 5 can be checked explicitly (*see* Non-trivial Models of Quantum Fields with Indefinite Metric).

It should be mentioned that different choices of the Hilbert seminorms p_n lead to potentially different maximal Hilbert space closures (Hoffmann 1998, Constantinescu and Gheondea 2001). In fact, often the topology is not even Poincaré invariant and hence the states that can be approximated with local states depend on a chosen inertial frame. This fact, for the case of QED, has been interpreted in terms of physical gauges.

Many results from axiomatic field theory (*see* Axiomatic Quantum Field Theory) with positive metric also hold in the case of QFT with indefinite metric, like the PCT and the Reeh–Schlieder theorem, the irreducibility of the field algebra (for massive theories) and the Bisoniano–Wichmann theorem (*see* Algebraic Approach to Quantum Field Theory). Other classical results, like the Haag–Ruelle scattering theory and the spin and statistics theorem definitively do not hold, as has been proved by counterexamples. This is, however, far from being a disadvantage, as, for example, it permits the introduction of various gauges in the scattering theory of the vector potential $A_\mu(x)$ and fermionic scalar "ghost" fields in the BRST quantization (*see* BRST Quantization) formalism.

Gupta–Bleuler Gauge Procedure

Here the Gupta–Bleuler gauge procedure is presented in a slightly generalized form following Steinmann's monograph. Classically, the equations of motion for the vector potential $A_\mu(x)$,

$$\partial^\nu\partial_\nu A_\mu(x)+\lambda\partial_\mu\partial^\nu A_\nu(x)=j_\mu(x) \qquad [16]$$

together with Lorentz gauge condition $B(x)=\partial_\mu A^\mu(x)=0$ imply the Maxwell equations [10]. Here, $\lambda\in\mathbb{R}$ plays the role of a gauge parameter. As seen above, both equations, the so-called pseudo-Maxwell equations [16] and the Lorentz gauge condition $B(x)=0$, cannot both hold as operator identities. The idea for the quantization

of the theory therefore is to give up the Lorentz gauge condition as an operator identity on the entire state space $\mathcal{H}$.

Suppose one has constructed such a theory with an indefinite inner state space $\mathcal{H}$. Already for the noninteracting theory, any invariant, spectral, local, and covariant solution requires indefinite metric, cf. the explicit formula [18] below. To complete the Gupta–Bleuler program, one needs to find a subspace of (equivalence classes of) physical states $\mathcal{H}'$ of the inner-product space $\mathcal{H}'$ such that the following conditions hold:

1. the vacuum is a physical state, that is, $\Omega \in \mathcal{H}'$;
2. observable fields like $j_\mu(x)$ and $F_{\nu\mu}(x)$ map $\mathcal{H}'$ to itself;
3. the inner product $\langle .,. \rangle$ restricted to $\mathcal{H}'$ is positive semidefinite;
4. observable fields map $\mathcal{H}''$, the set of null vectors in $\mathcal{H}'$, to itself; and
5. the Maxwell equations hold on $\mathcal{H}'$ in the sense

$$\langle \Psi, \partial^\nu F_{\nu\mu}(x)\Phi\rangle = \langle \Psi, j_\mu(x)\Phi\rangle, \quad \forall \Psi, \Phi \in \mathcal{H}' \qquad [17]$$

Then one obtains $\mathcal{H}^{\text{ph}}$ as the completion of the quotient space $\mathcal{H}'/\mathcal{H}''$. The physical Hilbert space $\mathcal{H}^{\text{ph}}$ contains the vacuum Ω (1), observable fields act on $\mathcal{H}^{\text{ph}}$ (2) and (4), it is a Hilbert space (3) and the Maxwell equations hold on it (5).

To see that such a construction is possible, consider the noninteracting case $j_\nu(x) = 0$, that is, the limit case of vanishing electrical charge $e \to 0$, first. By taking the divergence of [16], one obtains $(1-\lambda)\partial^\nu\partial_\nu\partial^\mu A_\mu(x) = 0$. Excluding the Landau gauge ($\lambda = 1$), this implies $(\partial^\nu\partial_\nu)^2 A_\mu(x) = 0$. The most general solution for the two-point vacuum expectation values that is in agreement with [16] and the requirements of locality, translation invariance, the spectral condition, uniqueness of the vacuum, and the Lorentz covariance of $A^\mu(x)$ is then

$$\begin{aligned}\langle \Omega, A_\nu(x)A_\mu(y)\Omega\rangle &= (-g_{\mu\nu} + \rho\partial_\mu\partial_\nu)D^+(x-y) \\ &\quad + \frac{\lambda}{1-\lambda}\partial_\mu\partial_\nu E^+(x-y)\end{aligned} \qquad [18]$$

where D^+ and E^+ are the inverse Fourier transforms of $\theta(p^0)\delta(p^2)$ and $\theta(p^0)\delta'(p^2)$ respectively, $p^2 = p \cdot p$, θ being the Heavyside function, δ the Dirac measure on $\mathbb{R}$ of mass one in zero and δ' its derivative. ρ and λ are gauge parameters, for example, the Feynman gauge corresponds to $\lambda = \rho = 0$. We have also omitted an overall factor corresponding to a field strength normalization (choice of numerical value of $\hbar$ – here $\hbar = 1$).

Using Wick's theorem and the GNS construction for inner-product spaces (cf. the preceding section), it is possible to realize a representation of the vector potential $A_\nu(x)$ as operator-valued distribution on some indefinite-metric state space $\mathcal{H}$ with Fock structure, for example, a Krein closure of the GNS space with Ω the GNS vacuum and $\mathcal{D} \subseteq \mathcal{H}$ the canonical domain of definition. In the case of Feynman gauge, the metric operator η can be obtained by a second quantization of the operator $f_\mu \to \sum_{\nu=1}^4 g_{\mu\nu} f_\nu$ on the one-particle space $\mathcal{S}_1$.

In particular, the field $B(x)$ acts as an operator-valued distribution on $\mathcal{H}$ and, by taking the divergence of [16], it follows that $\partial^\nu\partial_\nu B(x) = 0$. Thus, $B(x) = B^+(x) + B^-(x)$ can be decomposed into a positive ("annihilation") and a negative ("creation") frequency part $B^\pm(x)$. One obtains:

Theorem 6 *The space* $\mathcal{H}' = \{\Psi \in \mathcal{D}: B^+(x)\Psi = 0\}$ *fulfills all requirements* (1)–(5) *of the Gupta–Bleuler gauge procedure.*

Condition (1) is obvious and (2) follows from the fact that the fields $F_{\nu\mu}(x)$ and $B(x)$ commute, which can be checked on the level of two-point functions [18]. In the same spirit, one can also use [18] to check (3) and (4) by explicit calculations on the one-particle space and showing that $\mathcal{H}'$ is the Fock space over the one-particle states annihilated by $B^+(x)$. Finally, by Hermiticity of $A^\mu(x)$, $B^+(x)^* = B^-(x)$ and thus $\langle \Psi, B(x)\Phi\rangle = \langle \Psi, B^+(x)\Phi\rangle + \langle B^+(x)\Psi, \Phi\rangle = 0$. As the field $B(x)$ stands for the obstruction to Maxwell equations, this implies condition (5).

It should be noted that the physical state space $\mathcal{H}^{\text{ph}}$ does not depend on the gauge parameters λ, ρ and that it is spanned by repeated application of the field tensor $F_{\mu\nu}(x)$ to the vacuum.

By current conservation, the divergence of [16] still yields $\partial^\nu\partial_\nu B(x) = 0$ also in the interacting case where $e \neq 0$. One can then choose the same gauge condition as in Theorem 6 to define $\mathcal{H}'$. One can then try to prove that this space fulfills all the requirements of the Gupta–Bleuler procedure, for example, in the sense of perturbation theory. Using more advanced formulations as, for example, BRST quantization and Bogoliubov's local S-matrix formalism, this program has been completed up to a solution of the infrared problem (*see* Perturbative Renormalization Theory and BRST).

A different procedure, motivated by the necessity of coincidence of all charges Q_C, Q_G, and Q_M on the physical state space, has been elaborated by Steinmann. It deviates from the standard procedure in the sense that the physical space $\mathcal{H}'$ is not included in $\mathcal{H}$, but $\mathcal{H}^{\text{ph}}$ is directly obtained from the GNS procedure after taking certain limits of Wightman functions restricted to

certain gauge-invariant algebras constructed from the Borchers algebra and a limiting procedure in a gauge parameter. The Wightman functional on this gauge-invariant algebras are positive (in the sense of perturbation theory), the limiting procedure, however, implies that the so-obtained physical states are singular (i.e., have diverging inner product) to states in $\mathcal{H}$, hence the so-defined state spaces corresponding to going to a physical gauge after solving the problem of a perturbative construction of an indefinite-metric solution, are not subspaces of $\mathcal{H}$.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Approach to Topological Quantum Field Theory; Axiomatic Quantum Field Theory; Boundary Conformal Field Theory; BRST Quantization; Perturbative Renormalization Theory and BRST; Quantum Fields with Indefinite Metric: Non-Trivial Models.

Further Reading

Azizov TYa and Iokhvidov IS (1989) *Linear Operators in Spaces with an Indefinite Metric*. Chichester: Wiley-Interscience.

Bleuler K (1950) Eine neue Methode zur Behandlung der longitudinalen und skalaren Photonen. *Helvetica Physica Acta* 23: 567.

Constantinescu T and Gheondea A (2001) Representations of Hermitian kernels by means of Krein spaces II: invariant kernels. *Communications in Mathematical Physics* 216: 409–430.

Gottschalk H (2002) Complex velocity transformations and the Bisogniano–Wichmann theorem for quantum fields acting on Krein spaces. *Journal of Mathematical Physics* 43(10): 4753–4769.

Gupta SN (1950) Theory of longitudinal photons in quantum electrodynamics. *Proceedings of Physical Society A* 63: 681.

Hofmann G (1998) On GNS representations on inner product spaces: I. The structure of the representation space. *Communications in Mathematical Physics* 191: 299–323.

Morchio G and Strocchi F (1980) Infrared singularities, vacuum structure and pure phases in local quantum field theory. *Annals of the Institute Henry Poincaré (Mathematical Physics)*. 33: 251–282.

Morchio G and Strocchi F (1983) A nonperturbative approach to the infrared problem in QED: construction of charged states. *Nuclear Physics B* 211: 471–508.

Morchio G, Pierotti D, and Strocchi F (1990) Infrared vacuum structure in two dimensional local quantum field theory models: the massless scalar field, SISSA Trieste. *Journal of Mathematical Physics* 31: 147.

Steinmann O (2000) *Perturbative Quantum Electrodynamics and Axiomatic Field Theory*. Berlin: Springer.

Strocchi F and Wightman A (1974) Proof of the charge superselection rule in local relativistic quantum field theory. *Journal of Mathematical Physics* 15(12): 2198–2224.

Yngvason J (1973) On the algebra of test functions for field operators. *Communications in Mathematical Physics* 34: 315–333.

Integrability and Quantum Field Theory

T J Hollowood, University of Wales Swansea, Swansea, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The notion of integrability plays many different rôles in quantum field theory (QFT). In this article we interpret it in a narrow sense and describe some QFTs that are completely integrable, in the sense that there are as many integrals of motion as degrees of freedom. Necessarily this implies, since we are talking about field theories, that there is an infinite number of conserved quantities. The existence of such a tower of conserved quantities of increasing Lorentz spin implies, via the Coleman–Mandula theorem, that the theories are trivial in spacetime dimensions greater than 2. On the other hand, in $1 + 1$ dimensions there is a rich menagerie of such integrable quantum field theories (IQFTs). These theories are fascinating in their own right as nontrivial QFTs for which data like the S-matrix and spectrum can be determined exactly. We will describe these exact S-matrices for a series of seminal examples. In addition, we briefly describe the applications of these theories to statistical systems in two dimensions.

Classical Integrable Systems and Field Theories

For a field theory to be integrable it must have an infinite number of conserved charges. Necessarily these must be spacetime symmetries which extend the Poincaré symmetry in some way. It turns out that, due to a theorem of Coleman and Mandula, such

Notes:

extensions are very restrictive: they are only possible in $1+1$ dimensions (one dimension of space and one of time) apart from noninteracting theories. Below we describe some of the most important examples.

Affine Toda Theories

These theories describe the interactions of a set of scalar fields which we write as a vector $\boldsymbol{\phi}$. The action is

$$S = \int \mathrm{d}^2x \left(\frac{1}{2} (\partial_\mu \boldsymbol{\phi})^2 - V(\boldsymbol{\phi}) \right) \qquad [1]$$

The potential has to be very specially chosen in order that the resulting theory is integrable. The resulting theories are classified by affine Lie algebras. We shall describe only the theories related to a simply laced Lie algebra $\mathfrak{g}$ (so of ADE type). In this case, for the affine version of the theory,

$$V(\boldsymbol{\phi}) = \frac{m^2}{\beta^2} \sum_{a=0}^{r} n_a \, \mathrm{e}^{\beta \boldsymbol{\alpha}_a \cdot \boldsymbol{\phi}} \qquad [2]$$

where $\boldsymbol{\phi}$ is an r-rank $\mathfrak{g}$ vector and $\boldsymbol{\alpha}_a, a = 1, \ldots, r$, are a set of simple roots of $\mathfrak{g}$. The fact that we are considering the affine version of the theory means that we include the term involving the extended root (the lowest root) $\boldsymbol{\alpha}_0 = -\sum_{a=1}^{r} n_a \boldsymbol{\alpha}_a$, which defines the integers $n_a (n_0 = 1)$. If this term is absent then the potential does not have a minimum. Such nonaffine theories are interesting in their own right since they include the Liouville theory, but we shall not describe them here.

One way to expose the infinite set of conserved charges at the classical level is to write the equations of motion in Lax form. This has the form of the vanishing of the field strength, or zero-curvature condition, of an auxiliary gauge connection in $\mathfrak{g}$ with components (A_x, A_t):

$$\begin{aligned} A_x &= \partial_t \boldsymbol{\phi} \cdot \boldsymbol{H} + \frac{m}{2\beta} \sum_{a=0}^{r} \mathrm{e}^{\beta \boldsymbol{\alpha}_a \cdot \boldsymbol{\phi}/2} (e_a + f_a) \\ A_t &= \partial_x \boldsymbol{\phi} \cdot \boldsymbol{H} + \frac{m}{2\beta} \sum_{a=0}^{r} \mathrm{e}^{\beta \boldsymbol{\alpha}_a \cdot \boldsymbol{\phi}/2} (e_a - f_a) \end{aligned} \qquad [3]$$

Here, $\{e_i, f_i\}$ are related to generators of $\mathfrak{g}$ in a Cartan–Weyl basis, via

$$\begin{aligned} &e_a = z E_{\boldsymbol{\alpha}_a}, \quad f_a = z^{-1} E_{-\boldsymbol{\alpha}_a}, \qquad a = 1, \ldots, r \\ &e_0 = z^{-h} E_{\boldsymbol{\alpha}_0}, \qquad f_0 = z^{h} E_{-\boldsymbol{\alpha}_0} \end{aligned} \qquad [4]$$

where z is a auxiliary variable known as the spectral parameter and h is the Coxeter number of $\mathfrak{g}$. Using the following commutators of $\mathfrak{g}$,

$$\begin{aligned} [E_{\boldsymbol{\alpha}_a}, E_{\boldsymbol{\alpha}_b}] &= \delta_{ab} \boldsymbol{\alpha}_a \cdot \boldsymbol{H} \\ [\boldsymbol{H}, E_{\boldsymbol{\alpha}}] &= \boldsymbol{\alpha} E_{\boldsymbol{\alpha}} \\ [E_{\boldsymbol{\alpha}_a}, E_{-\boldsymbol{\alpha}_b}] &= 0 \end{aligned} \qquad [5]$$

it is straightforward to verify that the zero-curvature condition

$$F_{xt} = \partial_x A_t - \partial_t A_x + [A_x, A_t] = 0 \qquad [6]$$

is equivalent to the equations of motion which follow from extremizing the action [1].

The fact that there exists a flat connection which depends on an auxiliary parameter z is sufficient to ensure integrability. In brief, the idea is that the gauge connection can be "abelianized" by a gauge transformation:

$$\tilde{A}_\mu = U \partial_\mu U^{-1} + U A_\mu U^{-1} \quad \text{with } [\tilde{A}_t, \tilde{A}_x] = 0 \qquad [7]$$

Hence, $\partial_t \tilde{A}_x - \partial_x \tilde{A}_t = 0$. This can be done in two inequivalent ways, such that $\tilde{A}_\mu$ are polynomials in z and z^{-1}, respectively. The corresponding coefficients are then conserved currents whose integrals give conserved charges. It can be shown that for the Toda theories these conserved charges have Lorentz spin given by an exponent $\{s_a\}$ of $\mathfrak{g}$ modulo its Coxeter number h:

$$\begin{aligned} A_n &: \quad h = n+1, \quad \{1,2,3,\ldots,n\} \\ D_n &: \quad h = 2n-2, \quad \{1,3,5,\ldots,2n-3,n-1\} \\ E_6 &: \quad h = 12, \quad \{1,4,5,7,8,11\} \\ E_7 &: \quad h = 18, \quad \{1,5,7,9,11,13,17\} \\ E_8 &: \quad h = 30, \quad \{1,7,11,13,17,19,23,29\} \end{aligned} \qquad [8]$$

This spectrum of conserved quantities seems to be a ubiquitous feature of IQFTs. These theories can be generalized by replacing $\mathfrak{g}$, or rather its (untwisted affinization) with any affine algebra.

The Sinh/Sine-Gordon Theory

These theories are the simplest of the Toda theories described above, associated to the Lie algebra A_1. In this case there is a single field and the potential has the form

$$V(\phi) = \frac{m^2}{2\beta^2} \left(\mathrm{e}^{\beta\phi} + \mathrm{e}^{-\beta\phi} \right) \qquad [9]$$

We have rescaled the field by $1/\sqrt{2}$ relative to the normalization in [2]. This potential defines the "sinh-Gordon theory." However, we can also take $\beta \to \mathrm{i}\beta$ to give the sine-Gordon theory with an action

$$S = \int \mathrm{d}^2x \left(\frac{1}{2} (\partial_\mu \phi)^2 + \frac{m^2}{\beta^2} \cos(\beta\phi) \right) \qquad [10]$$

The sine-Gordon theory is a useful paradigm for IQFTs because it exhibits most of the features of more complicated examples. To start with, it illustrates another important property of some integrable systems; namely, the existence of solitons. In the sine-Gordon case, the minima of the potential lie at $\phi = 2n\pi/\beta$, for an integer n, so there is a topological

kink that separates a vacuum n on the left and $n+1$ on the right, as well as an antikink. The explicit solution for the kink moving with velocity v is

$$\phi(x,t) = \frac{4}{\beta}\tan^{-1}\exp(m(x\cosh\theta - t\sinh\theta - \xi)) \quad [11]$$

where ξ is a constant and, since we are working in $1+1$ dimensions, we have introduced the rapidity θ, in terms of which the velocity is

$$v = \tanh\theta, \qquad -\infty \leq \theta \leq \infty \quad [12]$$

The antikink solution is simply the negative of the above. The kinks have a mass

$$M = \frac{8m}{\beta^2} \quad [13]$$

The existence of topological solitons is not a consequence of integrability, *per se*, for example, the ϕ^4 theory in $1+1$ dimensions also has kinks; however, in the integrable setting, the solitons have special properties that survive in the quantum theory. The first property is that multisoliton solutions can be found exactly using a variety of different techniques. They are most easily written down using the tau function, which is related to the field via

$$\phi = -\frac{1}{i\beta}\log\frac{\tau}{\tau^*} \quad [14]$$

The N-soliton solution can then be written compactly as

$$\tau = \sum_{\{\mu_p\}=0,1} \exp\left(\sum_{p=1}^{N}\mu_p\Phi^{(p)} + \sum_{p,q=1}^{N}\mu_p\mu_q\Upsilon^{(pq)}\right) \quad [15]$$

The sum is over the 2^N possibilities for which $\mu_p = 0$ or 1, for each p, and we have introduced

$$\Phi^{(p)} = m(x\cosh\theta_p - t\sinh\theta_p - \xi_p) \pm \frac{i\pi}{2} \quad [16]$$

The rapidity of the pth soliton is θ_p, and the choice of sign corresponds to the kink and antikink, respectively. The "interaction coefficient" is

$$\exp\Upsilon^{(pq)} = \tanh^2\left(\tfrac{1}{2}(\theta_p - \theta_q)\right) \quad [17]$$

For example, the two-soliton solution is

$$\tau = 1 + e^{\Phi^{(1)}} + e^{\Phi^{(2)}} + e^{\Upsilon+\Phi^{(1)}+\Phi^{(2)}} \quad [18]$$

The multisoliton solutions have a natural physical interpretation as the histories of a set of solitons which scatter off each other. To make this more precise, consider the two-soliton solution [18] in more detail. Suppose that $\xi_1 < \xi_2, v_1 > v_2$. Focus on the solution in the vicinity of the first soliton, that is, $x \sim v_1 t + \xi_1$. In the limit $t \to -\infty$, the solution is approximately

$$\tau \simeq 1 + e^{\Phi^{(1)}} \quad [19]$$

while, as $t \to \infty$, it is approximately

$$\tau \simeq e^{\Phi^{(2)}}\left(1 + e^{\Upsilon+\Phi^{(1)}}\right) \quad [20]$$

In both the limits, the solution represents an isolated soliton, the only difference is that the final "position offset" has been displaced: $\xi_1 \longmapsto \xi_1 - \Upsilon$. It is a consequence of integrability that the solitons interact in such a simple way. There were two solitons in the initial configuration and two in the final configuration traveling with the same velocities. The only effect is to introduce a time delay of

$$\Delta t = -\frac{\Upsilon(\theta)}{m\sinh(\theta/2)} \quad [21]$$

in the center-of-mass frame with $\theta_1 = -\theta_2 = \theta/2$, which we illustrate in **Figure 1**. We shall see that this kind of simple scattering is a characteristic feature of integrable field theories which extends to the quantum theory. It reflects the enormous restriction that the existence of the infinite set of integrals of motion puts on the dynamics.

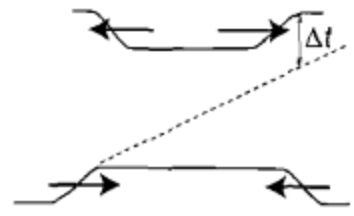


Figure 1 Classical scattering of a kink and an antikink. The final velocities equal the initial velocities and the only effect is to introduce a velocity-dependent time delay as shown.

Integrability at the Quantum Level

In this section we turn to the particular implications of integrability for the field theories at the quantum level. In discussing theories in $1+1$ dimensions it is convenient, as in [12], to use the rapidity. The energy and momentum of a particle of mass m are $E = m\cosh\theta$ and $p = m\sinh\theta$, respectively.

The sinh- and sine-Gordon theory, and their affine Toda generalizations, are scalar field theories with a well-behaved potential and as such they can be quantized in the conventional manner. It can be shown that integrability survives quantization and we now address its consequences. The key observation is that having an infinite set of higher-spin conserved quantities is very restrictive on the possible quantum processes. Assuming that the theory has a mass gap, the asymptotic states $|a,\theta\rangle$ are particles with rapidity

θ and additional quantum numbers needed to specify the state are indicated by the label a. These states are eigenstates of the conserved charges,

$$Q_s|a,\theta\rangle = q_s(a)e^{s\theta}|a,\theta\rangle \qquad [22]$$

Here, s is the spin of the charge which ranges over some infinite subset of the integers. Since the charges must commute with the S-matrix, it follows immediately that if an incoming state of n particles has a set of rapidities $\{\theta_1,\ldots,\theta_n\}$ then the outgoing state must also have n particles with the same set $\{\theta_1,\ldots,\theta_n\}$: there is consequently no particle creation! For example, we have illustrated the scattering of two particles in **Figure 2**. The two-particle S-matrix element will be denoted as

$$S_{ab}^{cd}(\theta_1-\theta_2):\ |a,\theta_1;b,\theta_2\rangle \longrightarrow |c,\theta_2;d,\theta_1\rangle \qquad [23]$$

Note that masses of the incoming particles must match the outgoing ones: $m_a = m_d$ and $m_b = m_c$. We have already seen this kind of behavior with the classical scattering of solitons in the sine-Gordon theory. In spite of the fact that the scattering is purely elastic, it can be nontrivial for two reasons: if there are mass degeneracies in the theory, the quantum numbers $\{a_1,\ldots,a_n\}$ can change and, in addition, the S-matrix element can depend nontrivially on the momenta.

The fact that the incoming and outgoing states have the same set of momenta leads to the notion of factorizability. To see what this means, consider the case of three particles. Let us imagine that we prepare the initial state to consist of three fairly narrow wave packets in position space with momenta smeared in accordance with the uncertainty principle. The key to the following argument is the fact that the infinite set of higher-spin conserved charges (with commute with the S-matrix) allow one to move the positions of the three particles relative to each other in an arbitrary way. In addition, the theory has a mass gap, so interactions have a finite range. By using this freedom, we can arrange for particles 1 and 2 to interact first,

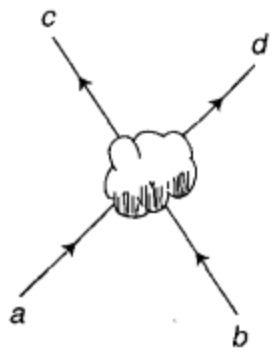


Figure 2 The two particle S-matrix with particles a and b in the initial state and c and d in the final state. For consistency, $m_a = m_d$ and $m_b = m_c$.

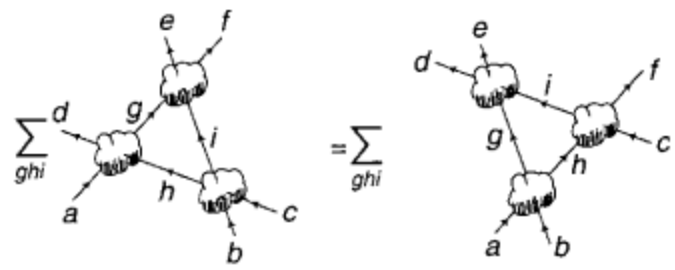


Figure 3 The scattering of three particles can factorize in two distinct ways as illustrated, leading to a nontrivial condition: the Yang–Baxter equation.

well before they come within interaction range of the third. Subsequently, the first two particles interact with the third as on the right-hand side of **Figure 3**. This ability to move the wave packets around using the symmetries means that the three-particle S-matrix element must "factorize" into a product of three two-particle elements:

$$S_{abc}^{def}(\theta_1,\theta_2,\theta_3) = \sum_{ghi} S_{ab}^{gh}(\theta_1-\theta_2)S_{hc}^{if}(\theta_1-\theta_3)S_{gi}^{de}(\theta_2-\theta_3) \qquad [24]$$

However, we could also use the symmetries afforded by the conserved charges to shift the positions of the particles so that particle 2 and 3 interact first, as on the left-hand side of **Figure 3**. Since the charges commute with the S-matrix, the result must the same; hence, there is a nontrivial consistency condition:

$$\sum_{ghi} S_{bc}^{hi}(\theta_2-\theta_2)S_{ah}^{dg}(\theta_1-\theta_3)S_{gi}^{ef}(\theta_2-\theta_3) = \sum_{ghi} S_{ab}^{gh}(\theta_1-\theta_2)S_{hc}^{if}(\theta_1-\theta_3)S_{gi}^{de}(\theta_2-\theta_3) \qquad [25]$$

This is the celebrated Yang–Baxter equation. Notice that it is only nontrivial if there are mass degeneracies, otherwise the particles on internal lines are determined by the external particles.

The factorization of the S-matrix extends readily to the case of more particles in an obvious way. An n-body element factorizes into a two-body element for each pair of particles. One might think that considerations of the n-particle S-matrix would lead to additional constraints; however, it can readily be shown that this is not the case and that the Yang–Baxter equation acts as a basic "move" which allows one to reorder the n-particle S-matrix into an arbitrary order. Further conditions on the S-matrix come from the axioms of analytic S-matrix theory:

(i) *Unitarity*

$$\sum_{ef} S_{ab}^{ef}(\theta)S_{ef}^{cd}(-\theta) = \delta_{ac}\delta_{bd} \qquad [26]$$

(ii) *Crossing symmetry* Each particle a has an antiparticle $\bar{a}$ and

$$S_{ab}^{cd}(\theta) = S_{b\bar{d}}^{\bar{a}c}(\pi i - \theta) \qquad [27]$$

(iii) *Analyticity* The S-matrix is a meromorphic function of θ on the physical strip, $0 \leq \mathrm{Im}\,\theta \leq \pi$. Singularities in most instances occur along the imaginary axis and the simple poles correspond to direct or cross-channel resonances. In this case, if $S_{ab}^{de}(\theta)$ has a simple pole at $\theta = iu_{ab}^{c}$ (necessarily a nonphysical rapidity difference) in the direct channel there exists a bound state of a and b of mass

$$m_c^2 = m_a^2 + m_b^2 + 2m_a m_b \cos u_{ab}^{c} \qquad [28]$$

The situation is illustrated in **Figure 4**. The new particle must itself be included in the particle spectrum. The S-matrix elements at the pole have the form

$$S_{ab}^{de}(\theta) = \sum_c P_{ab}^{c} \frac{ir_{ab}^{c}}{\theta - iu_{ab}^{c}} P_{\bar{d}\bar{e}}^{\bar{c}} + \cdots \qquad [29]$$

where P_{ab}^{c} can be thought of as a kind of projection operator with

$$\sum_{ab} P_{ab}^{c} P_{\bar{b}\bar{a}}^{\bar{d}} = \delta^{cd} \qquad [30]$$

Unitarity of the QFT requires that r_{ab}^{c} is real and positive, although there are also examples of nonunitarity theories with exact S-matrices. If $ab \to c$ can occur then so can $a\bar{c} \to \bar{b}$ and $b\bar{c} \to \bar{a}$. From [28], we deduce the following identity:

$$u_{ab}^{c} + u_{a\bar{c}}^{\bar{b}} + u_{b\bar{c}}^{\bar{a}} = 2\pi \qquad [31]$$

The data $\{u_{ab}^{c}\}$ for any given scattering theory are known as the fusing angles.

(iv) *The Bootstrap equations* These give a nonlinear relation between S-matrix elements. The basic idea is that if particle c appears as a resonance in the scattering of a and b then the S-matrix element of c with another state d can be deduced in terms of the scattering of d with a and b. This is illustrated in **Figure 5**. Using [30], we can write the resulting equation for the S-matrix element of c and d directly:

$$S_{cd}^{ef}(\theta) = \sum_{ghi} P_{\bar{a}\bar{b}}^{\bar{c}} S_{ab}^{eg}\left(\theta - i\bar{u}_{a\bar{c}}^{\bar{b}}\right) S_{bd}^{hi}\left(\theta + i\bar{u}_{b\bar{c}}^{\bar{a}}\right) P_{gi}^{f} \qquad [32]$$

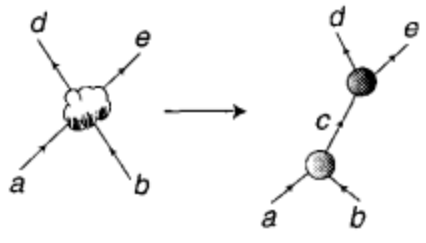


Figure 4 Near a direct channel pole, the scattering of a and b is dominated by the bound state c.

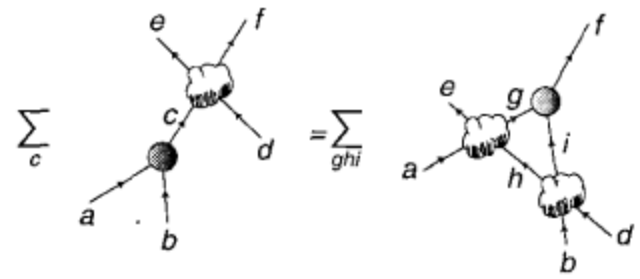


Figure 5 The bootstrap equations result from considering the interaction of a particle d with the bound state c of a and b in two distinct ways as illustrated.

The bootstrap constraints are very powerful because they allow one to extract the S-matrix elements of new particles that appear as bound states. This leads to the philosophy of the "bootstrap program" where one attempts to build consistent S-matrices starting from the S-matrix for a subset of particles which act as a seed for the algorithm. The process is quite an art, but at the end one has to be satisfied that the complete analytic structure is consistent with all the axioms. The key is to be able to account for all the poles in a consistent way, either in terms of bound states, as above, or in terms of the Coleman–Thun mechanism. This allows some poles to be interpreted in ways other than the existence of a bound state. The bootstrap algorithm is very complicated in general and at the present time a complete classification of solutions is not known. However, there are a large number of known solutions which appear to be intimately related to Lie algebras and associated structures known as Yangians and quantum groups. Below we describe some of the simplest known solutions.

Minimal S-Matrices

These scattering theories are in some sense the simplest. The particle spectrum is generally nondegenerate and so the Yang–Baxter equation is trivial. As is ubiquitous in the subject of IQFT, the classification of the theories is related to Lie algebras, although what seems to be important is not so much the algebra in question but rather the details of the associated root system. In this case the appropriate algebras are the simply laced algebras of ADE type. The number of particles is equal to the rank r of the Lie algebra and the masses are given by the r elements of one of the eigenvectors of the Cartan matrix of the algebra $\mathfrak{g}$:

$$\sum_{b=1}^{r} C_{ab} m_b = \left(2 - 2\cos\frac{\pi}{h}\right) m_a \qquad [33]$$

where h is the Coxeter number of $\mathfrak{g}$. The conserved charges have spins corresponding to the exponents of $\mathfrak{g}$ modulo h. We briefly explain how the complete

S-matrix can be written down in terms of properties of the root system of $\mathfrak{g}$. Let Φ be the set of roots of $\mathfrak{g}$, and $\alpha_a, a=1,\ldots,r$, a set of simple roots, as in the last section. In terms of these, $C_{ab}=2\alpha_a\cdot\alpha_b/\alpha_b^2$. Let $\omega_a, a=1,\ldots,r$, be a corresponding set of fundamental weights, $\alpha_a\cdot\omega_b=\delta_{ab}$.

Key to defining the theories is the notation of the Weyl group of $\mathfrak{g}$, the group generated by reflections in the simple roots:

$$R_a(\alpha)=\alpha-\frac{2\alpha\cdot\alpha_a}{\alpha_a^2}\alpha_a \qquad [34]$$

The element $w=R_1R_2\cdots R_r$ is known as a Coxeter element of the Weyl group, and it has special properties that are significant in the present context. In particular, its eigenvalues are of the form $\exp(2\pi i s_a/h)$, where h is the Coxeter number of $\mathfrak{g}$ and the integers s_a are the exponents of the algebra as in [8]. Note that there is always a pair with $s_1=1$ and $s_r=h-1$. Clearly, w acts as a rotation in the two-dimensional space spanned by the two corresponding eigenvectors. We can define an antisymmetric function $u(\alpha,\beta)$ on roots to be h/π times the (signed) angle between the projections of α and β onto this two-dimensional eigenspace. In preparation for what follows, it is useful to also define the roots

$$\phi_a=R_rR_{r-1}\cdots R_{a+1}(\alpha_a) \qquad [35]$$

We can now present P Dorey's amazingly compact formula for the complete S-matrix. For the scattering of particle a with particle b,

$$S_{ab}(\theta)=\prod_{\beta\in\Gamma_b}\{1+u(\phi_a,\beta)\}^{\omega_a\cdot\beta} \qquad [36]$$

In this formula Γ_b is the set of positive roots of $\mathfrak{g}$ which lie in the orbit of ϕ_b under w. We have also defined the building block

$$\{x\}=(x+1)(x-1)$$
$$(x)=\frac{\sinh\left(\frac{\theta}{2}+\frac{i\pi x}{h}\right)}{\sinh\left(\frac{\theta}{2}-\frac{i\pi x}{h}\right)} \qquad [37]$$

The fusing rules are also particularly elegant in the language of root systems. There is a three-point coupling between $a_i, i=1,2,3$, if there exist three roots $\alpha^{(i)}\in\Gamma_{a_i}$ such that $\alpha^{(1)}+\alpha^{(2)}+\alpha^{(3)}=0$. Furthermore, the fusing occurs in the a_1,a_2 channel at rapidity difference

$$iu_{a_1a_2}^{\bar{a}_3}=\frac{i\pi}{h}u(\alpha^{(1)},\alpha^{(2)}) \qquad [38]$$

This is Dorey's fusing rule.

For the case of A_{n-1}, the S-matrices are particularly simple. The mass spectrum is

$$m_a=m\sin\frac{\pi a}{n},\quad a=1,\ldots,n-1 \qquad [39]$$

and Dorey's rule gives the possible fusings as $ab\to(a+b)\bmod n$, which occur at the rapidity values

$$\theta=iu_{ab}=\begin{cases} i\dfrac{a+b}{n}\pi & a+b<n \\ i\left(2-\dfrac{a+b}{n}\right)\pi & a+b\geq n \end{cases} \qquad [40]$$

The charge conjugation operator maps $a\to\bar{a}=n-a$ and the explicit form for the S-matrix elements is

$$S_{ab}(\theta)=\{a+b-1\}\{a+b-3\}\cdots\{|a-b|+1\} \qquad [41]$$

The element $S_{ab}(\theta)$ has one direct channel pole at $\theta=iu_{ab}$ corresponding to the exchange of the particle $a+b \bmod n$, and a cross-channel pole at $\theta=iu_{a\bar{b}}$ corresponding to the exchange of particle $a-b \bmod n$.

Affine Toda Theories

The bootstrap program has been solved for all the affine Toda theories. For the simply laced theories described earlier, the result is directly related to the minimal S-matrices constructed above. The only difference is that there are additional factors which depend on the coupling β of the Toda theory but which do not introduce any additional poles onto the physical strip. These CDD factors are included by simply changing the basic building block [37]:

$$\{x\}\to\{x\}_{\text{Toda}}=\frac{(x+1)(x-1)}{(x-1+B)(x+1-B)} \qquad [42]$$

where

$$B=\frac{1}{2\pi}\cdot\frac{\beta^2}{1+\beta^2/4\pi} \qquad [43]$$

The S-matrix structure for the Toda theories based on the nonsimply laced algebras is a good deal more complicated. Integrability is only maintained in the quantum theory if the ratios of the physical masses of the particles depend on the coupling constant β is some very special way.

The Sine-Gordon Theory

We have seen that the sine-Gordon theory has solitons at the classical level. At the quantum level,

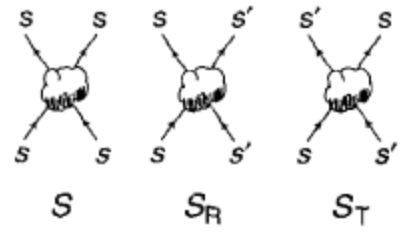


Figure 6 Soliton scattering processes. s and s' are the kink and antikink, respectively, or vice versa.

we expect that these kinks become *bona fide* particle states, in addition to the particle corresponding to the quantum fluctuations of the field ϕ. Focusing on the solitons, we expect a degenerate doublet corresponding to the kink and antikink. For the scattering of two solitons, there are six allowed processes illustrated in Figure 6. Unitarity [26] leads to the constraints

$$\begin{aligned} &S(\theta)S(-\theta) = 1 \\ &S_T(\theta)S_T(-\theta) + S_R(\theta)S_R(-\theta) = 1 \\ &S_T(\theta)S_R(-\theta) + S_R(\theta)S_T(-\theta) = 0 \end{aligned} \qquad [44]$$

while crossing symmetry [27] (using the fact that the soliton and antisoliton are antiparticles) gives

$$S(i\pi - \theta) = S_T(\theta), \qquad S_R(i\pi - \theta) = S_R(\theta) \qquad [45]$$

By themselves, these constraints are rather mild; however, the complete soliton S-matrix must also satisfy the Yang–Baxter equation [25]. The solution to all the constraints is not unique, however, the Zamolodchikovs conjectured that the exact answer is

$$\begin{aligned} S(\theta) &= \frac{1}{i\pi}\sinh\left(\frac{8\pi}{\gamma}(i\pi - \theta)\right)U(\theta) \\ S_T(\theta) &= \frac{1}{i\pi}\sinh\left(\frac{8\pi}{\gamma}\theta\right)U(\theta) \\ S_R(\theta) &= \frac{1}{\pi}\sin\left(\frac{8\pi^2}{\gamma}\right)U(\theta) \end{aligned} \qquad [46]$$

with

$$\begin{aligned} U(\theta) = \Gamma\left(\frac{8\pi}{\gamma}\right)\Gamma\left(1 + i\frac{8\theta}{\gamma}\right)\Gamma\left(1 - \frac{8\pi}{\gamma} - i\frac{8\theta}{\gamma}\right) \\ \times \prod_{n=1}^{\infty}\frac{R_n(\theta)R(i\pi - \theta)}{R_n(0)R_n(i\pi)} \end{aligned}$$

$$\begin{aligned} R_n(\theta) = &\frac{\Gamma\left(2n\frac{8\pi}{\gamma} + i\frac{8\theta}{\gamma}\right)}{\Gamma\left((2n+1)\frac{8\pi}{\gamma} + i\frac{8\theta}{\gamma}\right)} \\ &\times \frac{\Gamma\left(1 + 2n\frac{8\pi}{\gamma} + i\frac{8\theta}{\gamma}\right)}{\Gamma\left(1 + (2n-1)\frac{8\pi}{\gamma} + i\frac{8\theta}{\gamma}\right)} \end{aligned} \qquad [47]$$

where $\gamma = \beta^2(1 - \beta^2/8\pi)^{-1}$. The reason for confidence in the conjecture is that from the soliton S-matrix one can complete the bootstrap program and account for all the poles in terms of particles in the theory. In particular, there is a finite set of bound states of the soliton and antisoliton, called breathers, with masses

$$m_k = 2M\sin\frac{k\gamma}{16}, \quad k = 1, 2, \ldots < \frac{8\pi}{\gamma} \qquad [48]$$

Here, M is the soliton mass. The bootstrap equations give the S-matrix for the scattering of a soliton or antisoliton with the kth breather,

$$\begin{aligned} S_k(\theta) = &\frac{\sinh\theta + i\cos\frac{k\gamma}{16}}{\sinh\theta - i\cos\frac{k\gamma}{16}} \\ &\times \prod_{j=1}^{k-1}\frac{\sin^2\left(\frac{k-2j}{32}\gamma - \frac{\pi}{4} + i\frac{\theta}{2}\right)}{\sin^2\left(\frac{k-2j}{32}\gamma - \frac{\pi}{4} - i\frac{\theta}{2}\right)} \end{aligned} \qquad [49]$$

while, for the scattering of breather k with l,

$$\begin{aligned} S_{kl}(\theta) = &\frac{\sinh^2\theta + i\sin\left(\frac{k+l}{16}\gamma\right)\sinh\theta + i\sin\left(\frac{k-l}{16}\gamma\right)}{\sinh^2\theta - i\sin\left(\frac{k+l}{16}\gamma\right)\sinh\theta - i\sin\left(\frac{k-l}{16}\gamma\right)} \\ &\times \prod_{j=1}^{l-1}\frac{\sin^2\left(\frac{k-l-2j}{32}\gamma + i\frac{\theta}{2}\right)\cos^2\left(\frac{k+l-2j}{32}\gamma + i\frac{\theta}{2}\right)}{\sin^2\left(\frac{k-l-2j}{32}\gamma - i\frac{\theta}{2}\right)\cos^2\left(\frac{k+l-2j}{32}\gamma - i\frac{\theta}{2}\right)} \end{aligned} \qquad [50]$$

where we assume, without loss of generality, that $k \geq l$. The remarkable thing is that the scattering of the lowest-mass breather m_1 with itself,

$$S_{11}(\theta) = \frac{\sinh\theta + \mathrm{i}\sin\frac{\gamma}{8}}{\sinh\theta - \mathrm{i}\sin\frac{\gamma}{8}} \qquad [51]$$

is precisely the Toda S-matrix for A_1 with $\beta \to \mathrm{i}\beta/\sqrt{2}$ (the origin of the factor of $\sqrt{2}$ is mentioned after eqn [9]). This uniquely identifies the lowest-mass breather as being the quantum of the ϕ field.

The quantum structure that we have described above can be directly related to the classical scattering of solitons. In order to implement the classical limit, we can reintroduce $\hbar$ which is achieved by replacing β^2 by $\beta^2\hbar$. In this limit, the S-matrix elements have the form

$$S(\theta) = \exp\frac{2\mathrm{i}}{\hbar}(\delta(\theta) + \mathcal{O}(\hbar)) \qquad [52]$$

The phase $\delta(\theta)$ is related via the WKB approximation to the time delay in the classical theory of soliton scattering via

$$\delta(\theta) = \text{const.} + \int_0^\theta \mathrm{d}\theta' M \sinh(\theta/2)\Delta t(\theta) \qquad [53]$$

where $\Delta t(\theta)$ is the time delay in the center of mass (21). It is possible to verify [53] for the processes $S(\theta)$ and $S_T(\theta)$. Note that the reflection process has no classical analogue.

IQFT, Conformal Field Theories and Statistical Systems

We have described some IQFTs and their factorizable S-matrices in theories with a mass gap. We can ask the question, "what happens at very high energies compared with all the mass scales?" For a generic QFT such a limit may not exist, however, for a special class of theories the limit is a massless scale-invariant theory corresponding to a fixed point of the renormalization group. The massive theory can be thought of as a deformation of the massless theory by a particular relevant operator. At the fixed point, the Poincaré symmetry is enhanced to the full conformal group in the appropriate number of dimensions and the resulting theory is known as a conformal field theory (CFT). In $1+1$ dimensions the conformal group is infinite dimensional and so many CFTs are themselves integrable, in the sense that the complete spectrum of fields is known and their correlation functions can be constructed. Hence, an alternative way of thinking about many IQFTs is as a perturbation of a CFT by a specific relevant operator:

$$S_{\mathrm{IQFT}} = S_{\mathrm{CFT}} + g\int \mathrm{d}^2x\mathcal{O}(x) \qquad [54]$$

We will suppose that the operator has conformal dimensions $(\Delta, \bar{\Delta})$. This description of the theory can be turned around to ask the following question: which relevant deformations of a given CFT lead to IQFTs? Remarkably, since CFTs are so well understood, the question can often be answered exactly. The idea is that the conserved quantities of a CFT are all (anti-)holomorphic with respect to a holomorphic coordinate $z = x + it$. Conserved quantities include the stress tensor of spin 2 but include, in addition, an infinite tower of currents of ever increasing spin $\{T_s\}$. After perturbation, one has

$$\bar{\partial}T_s = gR^{(1)} + \cdots + g^nR^{(n)} + \cdots \qquad [55]$$

The conformal dimensions of the $R^{(n)}$ are $(s - n(1-\Delta), 1 - n(1-\Delta))$. Since the conformal dimensions of fields in a CFT are bounded below by zero, it follows that the series on the right-hand side truncates. The question of whether T_s remains conserved away from the CFT boils down to the question as to whether the right-hand side has the form $\partial\Theta$, for some Θ. Zamolodchikov found an ingenious counting argument which showed in certain circumstances that the right-hand side has precisely this form for some $s > 2$. This is sufficient to establish that the perturbed theory is an IQFT. In certain cases the spectrum of spins of the conserved quantities that are established by the counting argument is enough to make a connection with a known factorizable S-matrix.

This way of viewing IQFT as perturbations of CFTs is especially fruitful when we make the connection of the Euclidean QFT with the classical statistical mechanics of a two-dimensional system. In this connection, the Feynman path integral is reinterpreted as the sum over the configurations in the canonical ensemble with the Euclidean action interpreted as the energy. Usually, we consider statistical systems which are discrete, so typically defined on a lattice. The Euclidean QFTs are to be thought of as these statistical systems in the continuum limit where the lattice spacing is taken to zero keeping the long-range physics fixed. CFTs which have no massive degrees of freedom are identified with points of second-order phase transitions in the statistical system where correlation lengths are infinite. Perturbations of CFTs by relevant operators correspond to taking the statistical system away from criticality by changing some external parameter.

The prototypical example of such a statistical system is the Ising model. In the lattice version of

this model, there are a set spins $\{\sigma_i\}$ at each lattice site which can take the discrete values ± 1. The partition function of the theory is

$$Z(H,T) = \sum_{\{\sigma_i\}} \exp\left(-T^{-1}\sum_{\langle i,j\rangle}\sigma_i\sigma_j - H\sum_i \sigma_i\right) \quad [56]$$

The Ising model is the simplest model of a ferromagnet, where T is the temperature and H is the external applied field. The theory has a second-order phase transition for $T = T_c$, the Curie temperature, and $H = 0$ when the competition between the energy, which favors aligning the spins, and entropy, which favors disorder, exactly balance. In the two-dimensional neighborhood of the critical point, the lattice theory admits a continuum limit which can be described as the perturbation of a CFT, describing the critical Ising model, by a pair of relevant operators with couplings $T - T_c$ and H. In the case of the Ising model, the CFT is simply the theory of a free massless fermion in two-dimensional Euclidean space.

It turns out that in the two-dimensional space of relevant perturbations, there are two directions which lead to IQFTs. The most obvious is changing the temperature away from T_c while keeping $H = 0$. This leads to a particularly simple IQFT, that of a free massive fermion. More unexpectedly, the direction for which H varies away from 0, but $T = T_c$, also leads to an IQFT. In this case, Zamolodchikov's counting argument shows that there are higher-spin conserved charges of spin including

$$s = 1, 7, 11, 13, 17, 19, \ldots \quad [57]$$

This is remarkable because, as we have described previously, there is a minimal solution of the bootstrap program that describes the scattering of eight particles which has a spectrum of conserved charges that includes these spins. It is the minimal scattering theory related to the algebra E_8.

The fact that the scattering theory of the off-critical Ising model in the magnetic field direction has been identified is remarkable. From the S-matrix one can proceed to investigate the off-critical correlation functions using a technique known as the "form factor programe." Detailed simulation of the original lattice model [56] has provided strong support for the veracity of the E_8 scattering theory. For instance, the two lightest masses in the scattering theory determine the ratio of the two longest correlation lengths $m_2/m_1 = 2\cos(\pi/5)$.

In general, the identification of an IQFT and the CFT at its ultraviolet limit can be more difficult to establish. One way to proceed is to use the thermodynamic Bethe ansatz. This technique involves considering the thermodynamics of a gas of the particles in a periodic box. Since the scattering is purely elastic, thermodynamic quantities can be calculated, albeit in terms of a set of coupled nonlinear integral equations. If the box is small enough, ultraviolet effects dominate and various features of the CFT can be recovered.

Other IQFTs

There is a rich menagerie of other IQFTs that we have no space to discuss in detail. One is sigma models, whose fields take values in a Riemannian target space $\mathfrak{M}$ with an action

$$S = \int d^2x g_{ab}\partial_\mu X^a \partial^\mu X^b \quad [58]$$

where $g_{ab}dX^a\,dX^b$ is the metric of $\mathfrak{M}$. These theories are integrable at the classical level if the target space is either a group manifold of a compact simple group G or a symmetric space coset G/H, where H is a suitable subgroup of G. The former are known as the "principal chiral models." There are two kinds of conserved quantities, both local and nonlocal. At the quantum level, the conserved currents which imply classical integrability can be subject to quantum anomalies. An analysis of these anomalies proves that the principal chiral models are all integrable at the quantum level, while only the subset of symmetric space coset models, namely

$$\begin{aligned} &SO(n+1)/SO(n), \qquad SU(n)/SO(n) \\ &SU(2n)/Sp(n), \qquad SO(2n)/SO(n)\times SO(n) \\ &Sp(2n)/Sp(n)\times Sp(n) \end{aligned} \quad [59]$$

are quantum integrable. S-matrices have been proposed for all these integrable sigma models. They have a more complicated structure than most of the cases discussed here, because the particles fall into representations of the associated Lie groups and the Yang–Baxter equation, such as for the sine-Gordon solitons, is now nontrivial. Remarkably, gross features of the S-matrices, such as the mass spectrum fusing rules, are identical to the Toda theories or the minimal S-matrices.

Returning to IQFTs that are associsted with deformations of CFTs, there are more general classes which are associated with the renormalization group trajectories between two nontrivial fixed points. These theories have both massless and massive degrees of freedom. Even more remarkable are the staircase models of Zamolodchikov that exhibit an infinite series of crossover behavior where the renormalization group trajectory passes close to an infinite series of fixed points in sequence.

For all of the theories described above, one might have thought more generally that integrability is a very rigid property of a theory. In general, for example, the number of external coupling constants is very limited and the mass ratios are all fixed. For

example, in Toda theories there is only an overall mass scale m and the coupling β. If the form of the potential is altered in any way then integrability is lost. However, in certain circumstances, integrability appears to be a looser constraint that allows more flexibility. One class of such theories is known as the homogeneous sine-Gordon theories. These are integrable deformations of gauged WZW models associated with the coset $G/U(1)^r$, where r is the rank of a simple compact group G. In these theories there is a rich spectrum of both stable and unstable particles with masses and an S-matrix that depends continuously on a set of r coupling constants.

See also: Algebraic Approach to Quantum Field Theory; Bethe Ansatz; Constructive Quantum Field Theory; Eight Vertex and Hard Hexagon Models; Functional Equations and Integrable Systems; Integrable Systems: Overview; Quantum Field Theory: A Brief Introduction; Quantum Field Theory in Curved Spacetime; Sine-Gordon Equation; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Two-Dimensional Models; Yang–Baxter Equations.

Further Reading

Arinshtein AE, Fateev VA, and Zamolodchikov AB (1979) Quantum S matrix of the $(1+1)$-dimensional Todd chain. *Physics Letters B* 87: 389.

Braden HW, Corrigan E, Dorey PE, and Sasaki R (1990) Affine Toda field theory and exact S matrices. *Nuclear Physics B* 338: 689.

Delfino G (2004) Integrable field theory and critical phenomena: the Ising model in a magnetic field. *Journal of Physics A* 37: R45 (arXiv:hep-th/0312119).

Delius GW, Grisaru MT, and Zanon D (1992) *Nuclear Physics B* 382: 365 (arXiv:hep-th/9201067).

Dorey P (1992) Root systems and purely elastic S matrices. 2. *Nuclear Physics B* 374: 741 (arXiv:hep-th/9110058).

Dorey P (1998) Exact S matrices, arXiv:hep-th/9810026.

Dorey PE and Ravanini F (1993) Staircase models from affine Toda field theory. *International Journal of Modern Physics A* 8: 873 (arXiv:hep-th/9206052). (A B Zamolodchikov's original paper on the staircase models 'Resonance factorized scattering and roaming trajectories' is unpublished.).

Evans JM, Kagan D, MacKay NJ, and Young CAS (2005) Quantum, higher-spin, local charges in symmetric space sigma models. *Journal of High Energy Physics* 0501: 020 (arXiv:hep-th/0408244).

Jackiw R and Woo G (1975) Semiclassical scattering of quantized nonlinear waves. *Physical Review D* 12: 1643.

Miramontes JL and Fernandez-Pousa CR (2000) Integrable quantum field theories with unstable particles. *Physics Letters B* 472: 392 (arXiv:hep-th/9910218).

Mussardo G (1992) Off critical statistical models: factorized scattering theories and bootstrap program. *Physics Reports* 218: 215.

Olive DI and Turok N (1985) Local conserved densities and zero curvature conditions for Toda lattice field theories. *Nuclear Physics B* 257: 277.

Olshanetsky MA and Perelomov AM (1981) Classical integrable finite dimensional systems related to Lie algebras. *Physics Reports* 71: 313.

Zamolodchikov AB (1989) Integrals of motion and S matrix of the (scaled) $T=T(C)$ Ising model with magnetic field. *International Journal of Modern Physics A* 4: 4235.

Zamolodchikov AB and Zamolodchikov AB (1979) Factorized S-matrices in two dimensions as the exact solutions of certain relativistic quantum field models. *Annals of Physics* 120: 253.

Zamolodchikov AB (1990) Thermodynamic Bethe ansatz in relativistic models. Scaling three state Potts and Lee–Yang models. *Nuclear Physics B* 342: 695.

Operator Product Expansion in Quantum Field Theory

H Osborn, University of Cambridge, Cambridge, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The operator product expansion (OPE) provides an algebraic structure in quantum field theory. In a sense it supercedes or rather transcends the equal-time commutation relations, which provide the traditional starting point for the canonical quantization of any quantum field theory. The essential idea is that for any two local operator quantum fields at spacetime points x_1, x_2 their product may be expressed in terms of a series of other local quantum fields at a point x, which may be identified with x_1 or x_2, times c-number coefficient functions which depend on $x_1 - x_2$. The set of operators which may appear depends on the particular quantum field theory and must of course be in accord with any requirements of conserved quantum numbers. The coefficient functions depend on $x_1 - x_2$ in a fashion which depends on the dimensions of the various operators involved, at least up to renormalization group corrections. The most singular contributions are those for the operators appearing in the OPE with lowest scale dimension. From a phenomenological point of view, only the first few terms in the OPE are of relevance. However, theoretically, especially for conformal field theories, it is desirable to know the full expansion to all orders in powers of $x_1 - x_2$ in such a way that the operator product may

Notes:

be replaced by the full expansion in appropriate correlation functions. We first discuss the OPE for free theories and then the interacting case.

Free Field Theory

The OPE is most straightforward in free field theory when it almost reduces to a Taylor series expansion. For a simple free massless scalar field $\phi(x)$ then in four dimensions we may write

$$\phi(x)\phi(0) = \frac{C}{x^2} + :\phi(x)\phi(0): \qquad [1]$$

where :: denotes normal ordering (moving all annihilation operators to the right of creation operators) and C is just a normalization numerical constant (for canonical normalization $C = 1/4\pi^2$). The $1/x^2$ term proportional to the identity operator reflects the leading singular behavior at short distances of $\phi(x)\phi(0)$, the power being determined by ϕ having dimension 1. For the normal-ordered term we may expand in terms of an infinite set of local operators by using the Taylor expansion

$$:\phi(x)\phi(0): = \sum_{n=0}^{\infty} \frac{1}{n!} x^{\mu_1} \cdots x^{\mu_n} :\partial_{\mu_1} \cdots \partial_{\mu_n} \phi(0)\phi(0): \qquad [2]$$

where the operator appearing in the nth term has dimension $n+2$. Manifestly at short distances only the leading terms are relevant. Equation [1] also provides a point splitting definition of the local composite operator $:\phi^2(0):$ in terms of limit of $\phi(x)\phi(0)$ as $x \to 0$ after subtraction of the singular C/x^2 term.

The OPE can be easily generalized to composite operators defined by normal ordering. For $:\phi^2:$ we have, by applying Wick's theorem,

$$:\phi^2(x)::\phi^2(0): = \frac{2C^2}{x^4} + \frac{4C}{x^2} :\phi(x)\phi(0): + :\phi^2(x)\phi^2(0): \qquad [3]$$

where Taylor series expansion may be applied to both $:\phi(x)\phi(0):$ and also $:\phi^2(x)\phi^2(0):$ to give an infinite sequence of local operators of increasing dimensions.

The expansion in terms of local operators may be reordered. For instance, from [1] we may write, using $\partial^2\phi = 0$,

$$\phi(x)\phi(0) = \frac{C}{x^2} + \left(1 + \tfrac{1}{2}x^\mu\partial_\mu + \tfrac{1}{4}x^\mu x^\nu \partial_\mu\partial_\nu + \tfrac{1}{16}x^2\partial^2\right) :\phi^2(0): - \tfrac{1}{2}x^\mu x^\nu T_{\mu\nu} + O(x^3) \qquad [4]$$

where

$$T_{\mu\nu} = :\partial_\mu\phi\partial_\nu\phi: - \tfrac{1}{4}\eta_{\mu\nu} :\partial\phi\cdot\partial\phi: \qquad [5]$$

is the energy–momentum tensor. In [4], and also in a similar context subsequently, we define $\partial :\phi^2(0): = \partial_y :\phi^2(y): |_{y=0}$. The expansion [4] provides a point splitting definition of $T_{\mu\nu}$ and also demonstrates that many operators appearing in the OPE are expressible in terms of overall derivatives of lower-dimension operators. We may also note that without further input there is an ambiguity in the definition of $T_{\mu\nu}$ of the form

$$T_{\mu\nu} \sim T_{\mu\nu} + a(\partial_\mu\partial_\nu - \tfrac{1}{4}\eta_{\mu\nu}\partial^2) :\phi^2: \qquad [6]$$

In a conformal theory, however, we require $a = -1/6$.

Interacting Theories

The OPE becomes an essential tool in the context of interacting quantum field theories. For renormalizable quantum field theories various results can be proved to all orders in the standard perturbative expansion and are naturally assumed to be properties of the complete theory. In interacting theories we may no longer use normal ordering to define composite operators which, in general, have anomalous dimensions. The coefficient functions appearing in the OPE also gain perturbative corrections but these are constrained by renormalization group (RG) Callan–Symanzik equations.

Again if we consider the simplest case of a massless scalar theory as above but now with a renormalized coupling constant g the leading terms in the expansion of $\phi(x)\phi(0)$ are of the form (here we assume a Z_2 symmetry under $\phi \to -\phi$, otherwise the operator ϕ would be expected to appear in the OPE)

$$\phi(x)\phi(0) = \frac{C(g, \mu^2x^2)}{x^2} + D(g, \mu^2x^2)\phi^2(0) + \cdots \qquad [7]$$

where μ is an arbitrary renormalization scale. This arbitrariness is reflected in the RG equation

$$\left(\mu\frac{\partial}{\partial\mu} + \beta(g)\frac{\partial}{\partial g} + 2\gamma_\phi(g)\right) C(g, \mu^2x^2) = 0 \qquad [8]$$

At a fixed point $\beta(g_*) = 0$ this equation may be solved with an arbitrary choice of normalization to give $C(g_*, \mu^2x^2) = (\mu^2x^2)^{-\gamma_\phi(g_*)}$, which corresponds to the fields ϕ having a modified scale dimension $1 + \gamma_\phi(g_*)$. In a similar fashion the coefficient $D(g, \mu^2x^2)$ in [7] satisfies

$$\left(\mu\frac{\partial}{\partial\mu} + \beta(g)\frac{\partial}{\partial g} + 2\gamma_\phi(g) - \gamma_{\phi^2}(g)\right) \times D(g, \mu^2x^2) = 0 \qquad [9]$$

where it is necessary to introduce a new anomalous dimension function $\gamma_{\phi^2}(g)$ related to the composite

operator ϕ^2. Although it is natural to label the operator as ϕ^2 its definition in terms of the elementary field ϕ is essentially only as given in terms of the OPE [9]. At a fixed point again $D(g_*, \mu^2 x^2) = k(\mu^2 x^2)^{-\gamma_\phi(g_*)+(1/2)\gamma_{\phi^2}(g_*)}$, where the coefficient k is determined by the scale of the three-point function $\langle\phi(x)\phi(y)\phi^2(0)\rangle$. In asymptotically free theories the RG equations show that at short distances the coefficient functions tend to those of free field theory but with calculable logarithmic corrections. More generally, for a set of operators $\{O_i\}$ the OPE has the form

$$O_i(x)O_j(0) \sim \frac{1}{(x^2)^p}\sum_k C_{ijk}(g, \mu^2 x^2)O_k(0) \qquad [10]$$

where p is determined by the free scale dimensions of the O_i and

$$\left(\mu\frac{\partial}{\partial\mu} + \beta(g)\frac{\partial}{\partial g}\right)C_{ijk}(g, \mu^2 x^2)$$
$$= \sum_n \left(\gamma_{kn}(g)C_{ijn}(g, \mu^2 x^2) - \gamma_{in}(g)C_{njk}(g, \mu^2 x^2)\right.$$
$$\left. - \gamma_{jn}(g)C_{ink}(g, \mu^2 x^2)\right) \qquad [11]$$

with $\gamma_{in}(g)$ the anomalous dimension matrix arising from the mixing of composite operators.

An important aspect of the OPE is that the coefficient functions may be calculated perturbatively, essentially by applying the OPE in some suitable correlation function. Essentially the OPE provides a factorization between short-distance UV singularities and nonperturbative effects. In a Feynman graph the short distances in an operator product correspond to the large-momentum behavior and power-counting theorems allow a factorization up to calculable logarithmic corrections. A detailed analysis depends on the detailed technicalities of the proofs of renormalization to all orders of perturbation theory.

The coefficient functions in the OPE should be independent of any infrared or nonperturbative long-distance effects (such as confinement in QCD). However, the operators which appear in the OPE, such as ϕ^2 above, may have nonzero expectation values which are absent to all orders in perturbation theory.

Perturbative Example

The general considerations can be illustrated by considering a scalar field theory to lowest order in a perturbative expansion. We consider a four dimensional theory with a single scalar field and a potential $V(\phi) = \frac{1}{2}m^2\phi^2 + \frac{1}{24}g\phi^4$. Using dimensional regularization m^2, as well as g, is treated as a coupling with an associated β-function $\gamma_{\phi^2}(g)m^2$. With a mass term the operator ϕ^2 mixes with the identity operator so that

$$(\mathcal{D} + \gamma_{\phi^2}(g))\langle\phi^2(0)\rangle = -\gamma_{\phi^2 I}(g)m^2$$
$$\mathcal{D} = \mu\frac{\partial}{\partial\mu} + \beta(g)\frac{\partial}{\partial g} + \gamma_{\phi^2}(g)m^2\frac{\partial}{\partial m^2} \qquad [12]$$

where $\gamma_{\phi^2 I}$ reflects the mixing. At one loop order we have

$$\beta(g) = \frac{3g^2}{16\pi^2}, \quad \gamma_{\phi^2}(g) = \frac{g}{16\pi^2}, \quad \gamma_{\phi^2 I}(g) = \frac{1}{8\pi^2} \qquad [13]$$

and we may also set $\gamma_\phi(g) = 0$. In this case in the operator product expansion (7) the coefficient C also depends on m^2x^2 and the RG equations [8] and [9] are now modified to include the effects of mixing

$$\mathcal{D}C(g, m^2x^2, \mu^2x^2) = m^2x^2\gamma_{\phi^2 I}(g)D(g, \mu^2x^2)$$
$$(\mathcal{D} - \gamma_{\phi^2}(g))D(g, \mu^2x^2) = 0 \qquad [14]$$

From lowest order perturbation theory with [13], and using [14] to include all orders in $g\ln\mu^2x^2$, we have in this approximation

$$C(g, m^2x^2, \mu^2x^2)$$
$$= \frac{1}{4\pi^2} - \frac{2m^2x^2}{g}\left(1 + \frac{3g}{32\pi^2}\ln\mu^2x^2\right)^{2/3}$$
$$\times\left(\left(1 + \frac{3g}{32\pi^2}\ln\mu^2x^2\right)^{-1/3} - 1\right) \qquad [15]$$
$$D(g, \mu^2x^2) = \left(1 + \frac{3g}{32\pi^2}\ln\mu^2x^2\right)^{1/3}$$

The operator product expansion then reproduces the small x behavior of the two point function $\langle\phi(x)\phi(0)\rangle$ at one loop, expanding C, D to first order in g, if we take

$$\langle\phi^2(0)\rangle = -\frac{m^2}{8\pi^2}\ln\frac{\mu}{m} + O(g) \qquad [16]$$

which is in accord with [12]. If $m^2 < 0$ the symmetry $\phi \leftrightarrow -\phi$ is broken and it is necessary to shift the field $\phi = \nu + f$, with $\nu^2 = -6m^2/g$ and the field f has a mass m_f with $m_f^2 = -2m^2$. The operator product expansion [7] with the same coefficient functions as in [15] remains valid. The two point function $\langle\phi(x)\phi(0)\rangle$, which includes a nonperturbative term ν^2, is again reproduced for small x at one loop now if

$$\langle\phi^2(0)\rangle = -\frac{6m^2}{g} - \frac{m^2}{2\pi^2}\ln\frac{\mu}{m_f} + O(g) \qquad [17]$$

but in this case it is necessary to expand $D(g, \mu^2x^2)$ to $O(g^2)$ as a consequence of the leading $1/g$ term in [17]. Note that both [16] and [17] contain the nonperturbative dependance on $\ln m$ and $\ln m_f$ which is present in the two point function.

Conformal Field Theories

When the β-function vanishes and a quantum field theory enjoys conformal invariance the operator product expansion is a potentially convergent expansion. It is natural to restrict to conformal quasiprimary operators which do not mix with lower scale dimensions under conformal transformations. If we consider, for instance, two scalar operators ϕ with scale dimension Δ_ϕ then the OPE has the generic form

$$\phi(x)\phi(0) = \frac{1}{x^{2\Delta_\phi}} + \sum_I C_{\phi\phi O^I} \frac{1}{(x^2)^{1/2}(2\Delta_\phi - \Delta^I + \ell)} \times C^{(\ell)}_{\Delta^I}(x,\partial)^{\mu_1,\ldots,\mu_\ell} O^I_{\mu_1,\ldots,\mu_\ell}(0) \qquad [18]$$

where there is a sum over quasiprimary operators $O^I_{\mu_1,\ldots,\mu_\ell}$ with scale dimension Δ^I and spin ℓ, so they are symmetric traceless tensors of rank ℓ. In the first term in [18] the coefficient is chosen to be 1 by a choice of normalization. The coefficients $C_{\phi\phi O^I}$, with a standard normalization for O^I, are then determined by the coefficients of the corresponding three-point functions involving $\phi\phi$ and O^I. In [18] $C^{(\ell)}_{\Delta^I}$ are differential operators which sum up the contributions of all derivatives or descendants of the quasiprimary operator O^I. They can be explicitly given in terms of an integral representation, for any spacetime dimension, where the scale is fixed by requiring for the leading term $C^{(\ell)}_{\Delta^I}(x,0)^{\mu_1,\ldots,\mu_\ell} = x^{\mu_1}\cdots x^{\mu_\ell}$ – traces. The spectrum of operators which appear is obviously a property of the particular conformal field theory.

Ward Identities

If the theory has a symmetry with corresponding conserved currents then there are Ward identities which constrain the OPE of fields with the conserved current. For a current $J_{\mu a}$ then we have, in d dimensions, the singular contribution in the OPE is given by

$$J_{\mu a}(x)O(0) \sim -\frac{1}{S_d}\frac{x_\mu}{(x^2)^{(1/2)d}} t_a O(0) \qquad [19]$$

where t_a are a set of matrix generators corresponding to the symmetry acting on the fields O and S_d is the volume of the unit $(d-1)$-dimensional sphere, $S_4 = 2\pi^2$. For a conserved current there are no anomalous dimensions and the coefficient in [19], which depends on the normalization for the current $J_{\mu a}$, is chosen so that $[Q_a, O(0)] = -t_a O(0)$ with Q_a the charge formed from $J_{\mu a}$. For the energy–momentum tensor the operator there is an analogous result. We consider the simpler case of a conformal theory when the energy–momentum tensor is both conserved and traceless and

$$T_{\mu\nu}(x)O(0) \sim A_{\mu\nu}(x)O(0) + B_{\mu\nu\lambda}(x)\partial^\lambda O(0) + \cdots \qquad [20]$$

where $A_{\mu\nu}(x) = O(x^{-d})$ and $B_{\mu\nu\lambda}(x) = O(x^{-d+1})$. As a distribution $A_{\mu\nu}(x)$ is ambiguous up to terms proportional to $\delta^d(x)$. If Δ is the scale dimension of O and $s_{\mu\nu}$ are the Lorentz spin generators acting on O the Ward identities then give

$$\partial^\mu A_{\mu\nu}(x) = \left(\frac{\Delta}{d}\eta_{\nu\lambda} + C_{\nu\lambda} + \tfrac{1}{2}s_{\nu\lambda}\right)\partial^\lambda \delta^d(x)$$
$$A_{\mu\mu}(x) = C_{\mu\mu}\delta^d(x)$$
$$\partial^\mu B_{\mu\nu\lambda}(x) = -\eta_{\nu\lambda}\delta^d(x) \qquad [21]$$

where $C_{\mu\nu}$ is a constant tensor reflecting the arbitrariness in $A_{\mu\nu}$, it is immaterial as far as Ward identities are concerned. We may choose

$$\frac{\Delta}{d}\eta_{\nu\lambda} + C_{\nu\lambda} = 0 \qquad [22]$$

(If desired, we might also take $A'_{\mu\nu}(x) = A_{\mu\nu}(x) + (1/2)s_{\mu\nu}\delta^d(x)$ in which case $\partial_\mu A'_{\mu\nu}(x) = 0, A'_{[\mu\nu]}(x) = (1/2)s_{\mu\nu}\delta^d(x)$ but such an antisymmetric piece seems unnatural). In general there is no unique form for $A_{\mu\nu}(x)$, as a consequence of the freedom of choice for $C_{\nu\lambda}$ in [21]. However, for a scalar field O we must have, for $x \neq 0$,

$$A_{\mu\nu}(x) = \frac{\Delta}{d-1}\frac{1}{S_d}\left(\eta_{\mu\nu} - d\frac{x_\mu x_\nu}{x^2}\right)\frac{1}{(x^2)^{(1/2)d}} = -\frac{\Delta}{(d-1)(d-2)}\frac{1}{S_d}\partial_\mu\partial_\nu\frac{1}{(x^2)^{(1/2)d-1}} \qquad [23]$$

with the overall scale determined by [21].

For the operator product of the current $J_{\mu a}$ with itself there is an additional term proportional to the identity operator of the form

$$J_{\mu a}(x)J_{\nu b}(0) \sim C_J\delta_{ab}\left(\eta_{\mu\nu} - 2\frac{x_\mu x_\nu}{x^2}\right)\frac{1}{x^{2(d-1)}} \qquad [24]$$

where the coefficient C_J, which determines the scale of the two-point function for $J_{\mu a}$, is well defined since the normalization of the current is determined through the Ward identity. A similar result also holds for the operator product of the energy–momentum tensor with itself, with an overall coefficient C_T. In general, we may also write for the operator product of two scalar fields O:

$$O(x)O(0) \sim C_O\frac{1}{x^{2\Delta}} - \frac{C_O}{C_T S_d}\frac{d\Delta}{d-1}\frac{1}{(x^2)^{\Delta-(1/2)d+1}} \times x^\mu x^\nu T_{\mu\nu}(0) \qquad [25]$$

neglecting other contributions. The contribution of the energy–momentum tensor does not therefore introduce any new coefficient.

Two Dimensions

In two dimensions the OPE plays an essential role in the discussion of conformal field theories. For a Euclidean metric it is natural to use complex variables z and $\bar{z}$. The energy–momentum tensor in this case reduces to a chiral field $T(z)$ and its conjugate $\bar{T}(\bar{z})$. For the operator product with a chiral field $\phi(z)$ with scale dimension Δ,

$$T(z)\phi(0) \sim \frac{\Delta}{z^2}\phi(0) + \frac{1}{z}\phi'(0) \qquad [26]$$

and, for the operator product of T with itself,

$$T(z)T(0) \sim \frac{c}{2z^4} + \frac{2}{z^2}T(0) + \frac{1}{z}T'(0) \qquad [27]$$

Here c is the Virasoro central charge, which plays a critical role in the discussion of two-dimensional conformal field theories, it is given by the two-point function which follows from [27], $\langle T(z)T(0)\rangle = (1/2)cz^{-4}$.

In simple rational conformal field theories the operators are organized into conformal blocks by the infinite-dimensional extended conformal symmetry in two dimensions. This allows the full spectrum of operators and their dimensions to be determined and in consequence complete results for the OPE to be found in many cases.

Further Remarks

The OPE reflects the locality properties of quantum field theories and can be extended without difficulty to curved space backgrounds. For a product $\phi(x)\phi(0)$, the separation x^2 may be replaced by a biscalar at x and 0 but it is necessary to include in the OPE contributions involving the background Riemann tensor as well as the operator fields present in flat space. There is also a generalization of the OPE for superfields on superspace.

At a fundamental level although the OPE can be derived to all orders in perturbation theory the contribution of nonperturbative effects such as instantons to the coefficients is not entirely clear. Issues of associativity have yet to be fully analyzed.

There are also important applications to the phemenonological analysis of QCD when assumptions about the OPE and saturation of sum rules can lead to results for the vacuum expectation value of gauge-invariant operators such as $F^{\mu\nu}F_{\mu\nu}$.

See also: Boundary Conformal Field Theory; Effective Field Theories; Quantum Chromodynamics; Renormalization: General Theory; Renormalization: Statistical Mechanics and Condensed Matter; Two-Dimensional Models.

Further Reading

Cardy J (1987) Anisotropic corrections to correlation functions in finite size systems. *Nuclear Physics B* 290: 355.

Collins JC (1984) *Renormalization: An Introduction to Renormalization, The Renormalization Group and the Operator-Product Expansion*. Cambridge: Cambridge University Press.

David F (1984) On the ambiguity of composite operators, I.R. renormalons and the status of the operator product expansion. *Nuclear Physics B* 234: 237.

Di Francesco P, Mathieu P, and Sénéchal D (1997) *Conformal Field Theory*. New York: Springer.

Erdmenger J and Osborn H (1996) Conserved currents and the energy–momentum tensor in conformally invariant theories for general dimensions. *Nuclear Physics B* 483: 431.

Kadanoff LP (1969) Operator algebra and the determination of critical indices. *Physical Review Letters* 23: 1430.

Novikov VA, Shifman MA, Voinshtein AJ, and Zakharov VI (1985) Wilson's Operator product expansion: can it fail? *Nuclear Physics* B249: 445.

Wilson KG (1961) Non-Lagrangian models of current algebra. *Physical Review* 179: 1499.

Wilson KG (1971) Renormalization group and strong interactions. *Physical Review D* 3: 1818.

Quantum Fields with Indefinite Metric: Non-Trivial Models

S Albeverio and H Gottschalk, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The nonperturbative construction of quantum field models with nontrivial scattering in arbitrary dimension d of the underlying Minkowski spacetime is much simpler in the framework of quantum field theory with indefinite metric than in the positive-metric case. In particular, there exist a number of solutions in the physical dimension $d = 4$, where up to now no positive-metric solutions are known. The reasons why this is so are reviewed in this article, and some examples obtained by analytic continuation from the solutions of Euclidean covariant stochastic partial differential equations (SPDEs) driven by non-Gaussian white noise are discussed.

The Hilbert Space Structure Condition

It has been proved by F Strocchi that a quantum gauge field in a local, covariant gauge cannot act on a Hilbert space with a positive-definite inner product. But it is possible to overcome this obstacle by passing from a Hilbert space representation of the algebra of the quantum field to Krein space representations in order to preserve locality and covariance under the Poincaré group.

A Krein space $\mathcal{K}$ is an inner-product space which also is a Hilbert space with respect to some auxiliary scalar product. The relation between the inner product $\langle .,. \rangle$ and the auxiliary scalar product $(.,.)$ is given by a self-adjoint linear operator $J: \mathcal{K} \to \mathcal{K}$ with $J^2 = \mathbf{1}_{\mathcal{K}}$ and $\langle .,. \rangle = (.,J.)$. J is called the metric operator. A quantum field acting on such a space is called a quantum field with indefinite metric. The formal definition is as follows.

Let $\mathcal{D} \subseteq \mathcal{K}$ be a dense linear space and $\Omega \in \mathcal{D}$ a distinguished vector (henceforth called the vacuum). Let $\mathcal{S} = \mathcal{S}(\mathbb{R}^d, \mathbb{C}^N)$ be the space of Schwartz test functions with values in $\mathbb{C}^N$. A quantum field ϕ by definition is a linear mapping from $\mathcal{S}$ to the linear operators on $\mathcal{D}$. One usually assumes that $\mathcal{D}$ is generated as the linear span of vectors generated by repeated application of field operators to the vacuum. The following properties should hold for the quantum field ϕ:

1. *Temperedness*: $f_n \to f$ in $\mathcal{S} \Rightarrow \langle \Psi, \phi(f_n)\Phi \rangle \to \langle \Psi, \phi(f)\Psi \rangle \; \forall \Psi, \Phi \in \mathcal{S}$.
2. *Covariance*: There exists a weakly continuous representation U of the covering of the orthochronous, proper Poincaré group $\tilde{P}_+^{\uparrow}$ by linear operators on $\mathcal{D}$ which is J-unitary, that is, $U^{|*|} = U^{-1}$ with $U^{|*|} = JU^*J|_{\mathcal{D}}$ and leaves Ω invariant. ϕ is said to be covariant with respect to U and a representation τ of the covering of the orthochronous, proper Lorentz group $\tilde{L}_+^{\uparrow}$ if $U(g)\phi(f)U(g)^{-1} = \phi(f_g)$, where $f_g(x) = \tau(\Lambda)f(\Lambda^{-1}(x-a))$, $g = \{\Lambda, a\}$, $\Lambda \in \tilde{L}_+^{\uparrow}$, $a \in \mathbb{R}^d$.
3. *Spectrality*: Let $U(a)$, $a \in \mathbb{R}^d$, be the representation of the translation group and let $\sigma = \cup_{\Psi,\Phi\in\mathcal{D}} \mathrm{supp}\mathcal{F}(\langle \Psi, U(.)\Phi \rangle)$ with $\mathcal{F}$ the Fourier transform (in the sense of tempered distributions). Formally, σ is the joint spectrum of the generators of spacetime translations $U(a)$. The spectral condition then demands that $\sigma \subseteq \bar{V}_0^+$, the closed forward light cone in energy–momentum space.
4. *Locality*: There is a decomposition $\mathbb{C}^N = \oplus_\kappa V_\kappa$ such that for each $f, h \in \mathcal{S}$ taking values in a V_κ and having spacelike separated supports one has either $[\phi(f), \phi(h)] = 0$ or $\{\phi(f), \phi(h)\} = 0$, where $[.,.]$ is the commutator and $\{.,.\}$ the anticommutator.
5. *Hermiticity*: There is an involution $*$ on $\mathcal{S}$ such that $\phi(f)^{|*|} = \phi(f^*)$.

The quantum-mechanical interpretation of the inner product of two vectors in $\mathcal{K}$ as a probability amplitude, however, gets lost. It has to be restored by the construction of a physical subspace of $\mathcal{K}$ where the restriction of the inner product is non-negative. This is called the Gupter–Bleuler gauge procedure. Typically, one first considers the problem of constructing quantum fields with indefinite metric, that is, the dynamical problem is addressed. This is often followed by the construction of the physical states, which involves implementation of quantum constraints.

Notes:

The vacuum expectation values (VEVs), also called Wightman functions, of the quantum field theory with indefinite metric (IMQFT) are defined as

$$W_n(f_1 \otimes \cdots \otimes f_n) = \langle \Omega, \phi(f_1) \cdots \phi(f_n)\Omega \rangle \quad f_1, \ldots, f_n \in \mathcal{S} \qquad [1]$$

An axiomatic framework for (unconstrained) IMQFT has been suggested by G Morchio and F Strocchi in terms of the Wightman functions $W_n \in \mathcal{S}'$, $n \in \mathbb{N}_0$. Previous work on the topic had been done by J Yngvason. These generalized Wightman axioms of Morchio and Strocchi replace the positivity condition on the Wightman functions by a so-called Hilbert space structure condition (HSSC): for $n \in \mathbb{N}_0$ there exist p_n a Hilbert seminorm on $\mathcal{S}^{\otimes n}$ such that

$$|W_{n+m}(f \otimes h)| \leq p_n(f) p_m(h) \quad \forall n, m \in \mathbb{N}_0 \quad f \in \mathcal{S}^{\otimes n},\ h \in \mathcal{S}^{\otimes m} \qquad [2]$$

This condition makes sure that a field algebra on a Krein space with VEVs equal to the given set of Wightman functions can be constructed. The remaining axioms of the Wightman framework – temperedness, covariance, spectral condition, locality, and Hermiticity – remain the same. Clustering of Wightman functions is assumed at least for massive theories:

$$\lim_{t \to \infty} W_{n+m}(f \otimes h_{ta}) = W_n(f) W_m(h) \quad \forall n, m \in \mathbb{N}_0 \quad f \in \mathcal{S}^{\otimes n},\ h \in \mathcal{S}^{\otimes m} \qquad [3]$$

for spacelike $a \in \mathbb{R}^d$. It fails to hold in certain physical contexts where multiple vacua (also called Θ-vacua) accompanied with massless Goldstone bosons occur due to spontaneous symmetry breaking.

In the original Wightman axioms, there are essentially two nonlinear axioms: positivity and clustering. Here nonlinear means that checking that condition involves more than one VEV with a given number of field operators. The cluster condition can be linearized by an operation on the Wightman functions called "truncation." The equations

$$W_n(f_1 \otimes \cdots \otimes f_n) = \sum_{I \in \mathcal{P}^{(n)}} \prod_{\substack{\{j_1, \ldots, j_l\} \in I \\ j_1 < j_2 < \cdots < j_l}} W_n^T(f_{j_1} \otimes \cdots \otimes f_{j_l}) \qquad [4]$$

recursively define the truncated Wightman functions W_n^T for $n \in \mathbb{N}$. Here $\mathcal{P}^{(n)}$ stands for the set of all partitions of $\{1, \ldots, n\}$ into disjoint, nonempty sets. Unfortunately, the positivity condition (at least when combined with nontrivial scattering) becomes highly nonlinear for truncated Wightman functions. This can be seen as one explanation why it is so difficult to find nontrivial (i.e., corresponding to nontrivial interactions) solutions to the Wightman axioms.

But it turns out that, in contrast to positivity, the HSSC is essentially linear for truncated Wightman functions.

Theorem 1 *If there exists a Schwartz norm $\|\cdot\|$ on $\mathcal{S}$ such that W_n^T is continuous with respect to $\|\cdot\|^{\otimes n}$ for $n \in \mathbb{N}$ then the associated sequence of Wightman functions $\{W_n\}$ fulfills the HSSC* [2].

Note that $\|\cdot\|^{\otimes n}$ is well defined as $\mathcal{S}$ is a nuclear space. This theorem makes it much easier to construct IMQFTs. In particular, all known solutions of the linear program for truncated Wightman functions lead to an abundance of mathematical solutions to the axioms of IMQFT, as long as the singularities of truncated Wightman functions in position and energy–momentum space do not become increasingly stronger with growing n. For example, the perturbative solutions to Wightman functions of Ostendorf and Steinmann provide solutions when the perturbation series is truncated at a given order.

Relativistic Fields from Euclidean Stochastic Equations

In the classical work on constructive quantum field theory, relativistic fields in spacetime dimensions $d = 2$ and 3 have been constructed by analytic continuation from Euclidean random fields. This, in particular, has led to firm connections between quantum field theory and equilibrium statistical mechanics. Let us discuss one specific class of solutions of the axioms of IMQFT for arbitrary d which also stem from random fields related to an ensemble of statistical mechanics of classical, continuous particles. Mathematically, this is connected with using random fields with Poisson distribution. As in constructive QFT, the moments, also called Schwinger functions, of the random field can be analytically continued from Euclidean imaginary time to relativistic real time. That this is possible results from an explicit calculation. Axiomatic results cannot be used, as they depend on positivity or reflection positivity in the Euclidean spacetime, respectively.

By definition, a mixing Euclidean covariant random field φ is an almost surely linear mapping from $\mathcal{S}_{\mathrm{R}} = \mathcal{S}(\mathbb{R}^d, \mathbb{R}^N)$ to the space of real-valued

measurable functions (random variables) on some probability space that fulfills the following properties:

1. *Temperedness*: $f_n \to f$ in $\mathcal{S}_\mathbb{R} \Rightarrow \varphi(f_n) \xrightarrow{\mathcal{L}} \varphi(f)$.
2. *Covariance*: $\varphi(f) \stackrel{\mathcal{L}}{=} \varphi(f_g)\ \forall f \in S_\mathbb{R}, g = \{\Lambda, a\}$, $\Lambda \in \mathrm{SO}(d)$, $a \in \mathbb{R}^d$, $f_g(x) = \tau(\Lambda) f(\Lambda^{-1}(x-a))$ for some continuous representation $\tau : \mathrm{SO}(d) \to \mathrm{GL}(N)$.
3. *Mixing*: $\lim_{t\to\infty} \mathbb{E}[AB_{ta}] = \mathbb{E}[A]\mathbb{E}[B]$ for all square-integrable random variables $A = A(\varphi)$, $B = B(\varphi)$, and $B_{ta} = B(\varphi_{ta})$, $\varphi_{ta}(f) = \varphi(f_{ta})\ \forall f \in S_\mathbb{R}$, $a \in \mathbb{R}^d \setminus \{0\}$.

The mixing condition in the Euclidean spacetime plays the same role as the cluster property in the generalized Wightman axioms.

In particular, we consider random fields φ obtained as solutions of the SPDE $D\varphi = \eta$. In this equation, η is a noise field, that is, η is τ-covariant for some representation of $\mathrm{SO}(d)$, $\eta(f)$ has infinitely divisible probability law and $\eta(f)$, $\eta(h)$ are independent $\forall f, h \in \mathcal{S}_\mathbb{R}$ with $\mathrm{supp}\, f \cap \mathrm{supp}\, h = \emptyset$. D is a τ-covariant (i.e., $\tau(\Lambda) D \tau(\Lambda)^{-1} = D \quad \forall \Lambda \in \mathrm{SO}(d)$) partial differential operator with constant coefficients (also pseudodifferential operators D could be considered). From the classification of infinitely divisible probability laws, it is known that η essentially consists of Gaussian white noise and Poisson fields and derivatives thereof. Such a Gauss–Poisson noise field by the Bochner–Minlos theorem is characterized by its Fourier transform. Direct relations with QFT arise if one chooses

$$\mathbb{E}[e^{i\eta(f)}] = \exp\left\{ \int_{\mathbb{R}^d} \psi(f) - f \cdot \bar{\sigma}^2 p(-\Delta) f \, dx \right\} \qquad f \in \mathcal{S}_\mathbb{R} \qquad [5]$$

where $\psi : \mathbb{R}^N \to \mathbb{C}$ is a Lévy function,

$$\psi(t) = ia \cdot t - \frac{t \cdot \sigma^2 t}{2} + z \int_{\mathbb{R}^N \setminus \{0\}} (e^{it\cdot s} - 1)\, dr(s) \qquad t \in \mathbb{R}^N \qquad [6]$$

Here the centered dot represents a τ-invariant scalar product on $\mathbb{R}^N$, σ a positive-semidefinite τ-invariant $N \times N$ matrix, $z \geq 0$ a real number and r is a τ-invariant probability measure on $\mathbb{R}^n \setminus \{0\}$ with all moments. Further, $\bar{\sigma}^2_{\alpha,\beta} = (\partial^2 \psi(t)/\partial t_\alpha \partial t_\beta)|_{t=0}$, and $p : [0,\infty) \to [0,\infty)$ is a polynomial depending on D. If $\hat{D}^{-1}$, the Fourier-transformed inverse of D, exists, it can be represented by

$$\hat{D}^{-1}(k) = \frac{Q_E(k)}{\prod_{l=1}^P (|k|^2 + m_l^2)^{\nu_l}} \qquad [7]$$

Here $Q_E(k)$ is a complex $N \times N$ matrix with polynomial entries being τ-covariant, $\tau(\Lambda) Q_E (\Lambda^{-1}k) \tau(\Lambda)^{-1} = Q_E(k) \quad \forall \Lambda \in \mathrm{SO}(d)$, $k \in \mathbb{R}^d$. $\nu_l \in \mathbb{N}$ and $m_1 \in \mathbb{C}\setminus(-\infty, 0)$ are parameters with the interpretation of the mass spectrum $(m_1, \ldots, m_P)$ and $(\nu_1, \ldots, \nu_P)$ the dipole degrees of the related masses. We restrict ourselves to the case of positive mass spectrum where $m_l > 0$, and in this case

$$p(t) = p(t, D) = \frac{\prod_{l=1}^P (t + m_l^2)^{\nu_l}}{\prod_{l=1}^P m_l^{2\nu_l}}, \quad t > 0 \qquad [8]$$

One can show that φ obtained as the unique solution of the SPDE $D\varphi = \eta$ is a Euclidean covariant, mixing random field. The Schwinger functions (moments) of φ are given by

$$S_n(f_1 \otimes \cdots \otimes f_n) = E[\varphi(f_1) \cdots \varphi(f_n)], \quad f_1, \ldots, f_n \in \mathcal{S}_\mathbb{R} \qquad [9]$$

Now the Schwinger functions can be calculated explicitly. They are determined by the truncated Schwinger functions, cf. [4], as follows: for $n = 2$,

$$S^\mathrm{T}_{2,\alpha_1,\alpha_2}(x_1, x_2) = \frac{Q^\mathrm{E}_{2,\alpha_1,\alpha_2}(-i\nabla_2)}{\prod_{l=1}^N m_l^{2\nu_l}} \left[\prod_{l=1}^N (-\Delta + m_l^2)^{-\nu_l} \right](x_1 - x_2) \qquad [10]$$

and for $n \geq 3$

$$S^\mathrm{T}_{n,\alpha_1\cdots\alpha_n}(x_1, \ldots, x_n) = Q^\mathrm{E}_{n,\alpha_1\cdots\alpha_n}(-i\nabla_n) \times \int_{\mathbb{R}^d} \prod_{j=1}^n \left[\prod_{l=1}^N (-\Delta + m_l^2)^{-\nu_l} \right](x_j - x)\, dx \qquad [11]$$

where

$$Q^\mathrm{E}_{n,\alpha_1\cdots\alpha_n}(-i\nabla_n) = C^{\beta_1\cdots\beta_n} \prod_{l=1}^n Q_{E,\beta_l,\alpha_l}\left(-i\frac{\partial}{\partial x_l}\right) \qquad [12]$$

with

$$C_{\beta_1\cdots\beta_n} = (-i)^n \frac{\partial^n \psi(t)}{\partial t_{\beta_1} \cdots \partial t_{\beta_n}}\bigg|_{t=0} \qquad [13]$$

and the Einstein convention of summation and raising/lowering of indices on $\mathbb{R}^N$ with respect to the invariant inner product $\cdot$ is applied. The Schwinger functions fulfill the requirements of τ-covariance, symmetry, clustering, and Hermiticity from the Osterwalder–Schrader axioms of Euclidean QFT.

While there is no known general reason why a relativistic QFT should exist for a given set of Schwinger functions, one can take advantage of the explicit formulas [10]–[13] in order to calculate the analytic continuation from Euclidean to relativistic times explicitly.

It simplifies the considerations to exclude dipole fields, that is, one assumes that $\nu_l = 1$ for $l = 1, \ldots, n$. In physical terms, the no-dipole condition guarantees that the asymptotic fields in Minkowski spacetime fulfill the Klein–Gordon equation and thus generate particles in the usual sense if applied to the vacuum. If this condition is not imposed, asymptotic fields might only fulfill a dipole equation $(\Box + m^2)^2 \phi^{\text{in/out}} = 0$ or a related hyperbolic equation of even higher order, and the particle states generated by application of such fields to the vacuum require a gauge fixing (constraints) in order to obtain a physical interpretation. Given the no-dipole condition, one obtains by expansion into partial fractions

$$\frac{1}{\prod_{l=1}^{P} (|k|^2 + m_l^2)} = \sum_{l=1}^{N} \frac{b_l}{(|k|^2 + m_l^2)} \qquad [14]$$

with $b_l \in (0, \infty)$ uniquely determined and $b_l \neq 0$. For the truncated Schwinger functions, this implies $(n \geq 3)$ that

$$S^{\mathrm{T}}_{n,\alpha_1 \cdots \alpha_n}(x_1, \ldots, x_n) = Q^{\mathrm{E}}_{n,\alpha_1 \cdots \alpha_n}(-\mathrm{i}\nabla_n) \sum_{l_1, \ldots, l_n = 1}^{P} \times \prod_{r=1}^{n} b_{l_r} \int_{\mathbb{R}^d} \prod_{j=1}^{n} (-\Delta + m_{l_j}^2)^{-1}(x - x_j)\, \mathrm{d}x \qquad [15]$$

At this point, a lengthy calculation yields a representation of the functions $\int_{\mathbb{R}^d} \prod_{j=1}^{n} (-\Delta + m_j^2)^{-1} (x - x_j)\, \mathrm{d}x$ as the Fourier–Laplace transform of a distribution $\hat{W}^{\mathrm{T}}_{n,m_1,\ldots,m_n}$ that fulfills the spectral condition. This is equivalent to the statement that the analytic continuation of such functions to relativistic times yields $W^{\mathrm{T}}_{n,m_1,\ldots,m_n}$, where the latter distribution is the inverse Fourier transform of $\hat{W}^{\mathrm{T}}_{n,m_1,\ldots,m_n}$. This distribution up to a constant that can be integrated into Q^{E} is given by

$$\left\{ \sum_{j=1}^{n} \prod_{l=1}^{j-1} \delta^-_{m_l}(k_l) \frac{(-1)}{k^2 - m_j^2} \prod_{l=j+1}^{n} \delta^+_{m_l}(k_l) \right\} \delta\left(\sum_{l=1}^{n} k_l \right) \qquad [16]$$

Here $\delta^{\pm}_m(k) = \theta(\pm k^0)\delta(k^2 - m^2)$, where θ is the Heaviside step function and $k^2 = k^{0^2} - |k|^2$. On the other hand, the partial differential operator Q^{E}_n can be analytically continued in momentum space:

$$Q^{\mathrm{M}}_n((k_1^0, k_1), \ldots, (k_n^0, k_n)) = Q^{\mathrm{E}}_n((\mathrm{i}k_1^0, k_1), \ldots, (\mathrm{i}k_n^0, k_n)) \qquad [17]$$

$k_1, \ldots, k_n \in \mathbb{R}^d$. With the definition

$$\hat{W}^{\mathrm{T}}_{2,\alpha_1\alpha_2}(k_1, k_2) = (2\pi)^{(d+1)} \frac{Q^{\mathrm{M}}_{2,\alpha_1\alpha_2}(k_1, k_2)}{\prod_{l=1}^{N} m_l^2} \times \sum_{l=1}^{N} b_l\, \delta^-_{m_l}(k_1)\delta(k_1 + k_2) \qquad [18]$$

and

$$\hat{W}^{\mathrm{T}}_{n,\alpha_1 \cdots \alpha_n}(k_1, \ldots, k_n) = Q^{\mathrm{M}}_{n,\alpha_1 \cdots \alpha_n}(k_1, \ldots, k_n) \times \sum_{l_1, \ldots, l_n = 1}^{N} \prod_{j=1}^{n} b_{l_j}\, \hat{W}^{\mathrm{T}}_{n,m_{l_1},\ldots,m_{l_n}}(k_1, \ldots, k_n) \qquad [19]$$

the analytic continuation of Schwinger functions can be summarized as follows:

Theorem 2 *The truncated Schwinger functions S^{T}_n have a Fourier–Laplace representation with $\hat{W}^{\mathrm{T}}_n$ defined in eqns* [18] *and* [19]. *Equivalently, S^{T}_n is the analytic continuation of W^{T}_n from purely real relativistic time to purely imaginary Euclidean time. The truncated Wightman functions W^{T}_n fulfill the requirements of temperedness, relativistic covariance with respect to the representation of the orthochronous, proper Lorentz group $\tilde{\tau}: \mathrm{L}^{\uparrow}_{+}(d) \to \mathrm{Gl}(\mathrm{L})$, locality, spectral property, and cluster property. Here $\tilde{\tau}$ is obtained by analytic continuation of τ to a representation of the proper complex Lorentz group over $\mathbb{C}^d$ (which contains $\mathrm{SO}(d)$ as a real submanifold) and restriction of this representation to the real orthochronous proper Lorentz group.*

Again making use of the explicit formula in Theorem 2, the condition of Theorem 1 can be verified. This proves the existence of IMQFT models associated with the class of random fields under discussion.

Theorem 3 *The Wightman functions defined in Theorem 2 fulfill the HSSC* [2]. *In particular, there exists a QFT with indefinite metric such that the Wightman functions are given as the VEVs of that IMQFT.*

Nontrivial Scattering

Theories as described in Theorem 2 obviously have trivial scattering behavior if the noise field η is Gaussian, that is, if, in [7], $z = 0$. In the case where there is also a Poisson component in η, that is, $z > 0$, higher-order truncated Wightman functions do not vanish and such relativistic theories have nontrivial scattering.

Before the scattering of the models can be discussed, some comments about scattering in IMQFT in general are in order. The scattering

theory in axiomatic QFT, Haag–Ruelle theory, relies on positivity. In fact, one can show that in the class of models under discussion, the LSZ asymptotic condition is violated if dipole degrees of freedom are admitted. In that case more complicated asymptotic conditions have to be used. In any case, the Haag–Ruelle theory cannot be adapted to IMQFT.

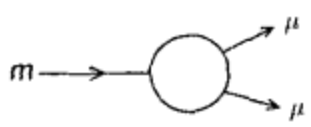
Nevertheless, asymptotic fields and states can be constructed in IMQFT if one imposes a no-dipole condition in a mathematically precise way. Then the LSZ asymptotic condition leads to the construction of mixed VEVs of asymptotic in- and out-fields with local fields. The collection of such VEVs is called the form-factor functional. After constructing this collection of mixed VEVs, one can try to check the HSSC for this functional and obtains a Krein space representation for the algebra generated by in- local and out-fields.

Following this line, asymptotic in- and out-particle states can be constructed for the given mass spectrum $(m_1, \ldots, m_P)$. If $a^{\text{in/out}\dagger}_{\alpha,l}(k)$, $l = 1, \ldots, P$, denotes the creation operator for an incoming/outgoing particle with mass m_l, spin component α, and energy–momentum k, the following scattering amplitude can be derived for r incoming particles with masses $m_{l_1}, \ldots, m_{l_r}$ and $n - r$ outgoing particles with masses $m_{l_{r+1}}, \ldots, m_{l_n}$:

$$\left\langle a^{\text{in}\dagger}_{\alpha_1,l_1}(k_1)\cdots a^{\text{in}\dagger}_{\alpha_r,l_r}(k_r)\Omega, a^{\text{out}\dagger}_{\alpha_{r+1},l_{r+1}}(k_{r+1})\cdots a^{\text{out}\dagger}_{\alpha_n,l_n}(k_n)\Omega\right\rangle^{\mathsf{T}}$$
$$= -(2\pi)\mathrm{i}Q^{\mathrm{M}}_{\alpha_1,\ldots,\alpha_n}(-k_1,\ldots,-k_r,k_{r+1},\ldots,k_n)$$
$$\times \prod_{j=1}^{n} \delta^{+}_{m_{l_j}}(k_j)\,\delta(K^{\text{in}} - K^{\text{out}}) \qquad [20]$$

$K^{\text{in/out}}$ stand for the total energy–momentum of in- and out-particles, that is, $K^{\text{in}} = \sum_{j=1}^{r} k_j$ and $K^{\text{out}} = \sum_{j=r+1}^{n} k_j$.

Two immediate consequences can be drawn from [20]. First, choosing a model with nonvanishing Poisson part such that $C_{\beta_1\beta_2\beta_3} \neq 0$ and a differential operator D containing in its mass spectrum the masses m and μ with $m > 2\mu$, one gets a nonvanishing scattering amplitude for the process



[21]

even though in- and out-particle states consist of particles with well-defined sharp masses. Thus, for the incoming particle, the energy uncertainty, which for a particle at rest is proportional to the mass uncertainty, vanishes but still the particle undergoes a nontrivial decay and must have a finite decay time. This appears to be a contradiction to the energy–time uncertainty relation, which therefore seems to have an unclear status in IMQFT (i.e., in QFT including gauge fields). The origin of this inequality, which of course is experimentally very well tested, apparently has to be located in the constraints, that is, in the procedure of implementing a gauge, of the theory and not in the unconstrained IMQFT.

Second, one can replace somewhat artificially the polynomials Q^{M}_n in [17] by any other symmetric and relativistically covariant polynomial. If the sequence of the "new" Q^{M}_n is of uniformly bounded degree in any of the arguments $k_1, \ldots, k_n$, the redefined Wightman functions in [17] still fulfill the requirements of Theorem 1 and thus define a new relativistic, local IMQFT. The scattering amplitudes of such a theory are again well defined and given by [20]. For example, in the case of only one scalar particle with mass m, one can show that arbitrary Lorentz-invariant scattering behavior of bosonic particles can be reproduced by such theories for energies below an arbitrary maximal energy up to arbitrary precision. This kind of interpolation theorem shows that the outcome of an arbitrary scattering experiment can be reproduced within the formalism of (unconstrained) IMQFT as long as it is in agreement with the general requirements of Poincaré invariance and statistics.

List of Symbols

$\to$	converges to
$\xrightarrow{\mathcal{L}}$	convergence in law
$\mathbb{N}$	set of natural numbers
$\mathbb{N}_0$	set of natural numbers and zero
$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\mathbf{1}$	identity mapping
$\vert_{\mathcal{D}}$	restricted to $\mathcal{D}$
x^0 and $\boldsymbol{x}$	time and spatial part of $x = (x^0, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^{d-1}$
∇_n	gradient operator on $\mathbb{R}^{dn}$

See also: Algebraic Approach to Quantum Field Theory; Euclidean Field Theory; Indefinite Metric; Perturbative Renormalization Theory and BRST; Quantum Field Theory in Curved Spacetime; Quantum Field Theory: A Brief Introduction; Stochastic Differential Equations.

Further Reading

Albeverio S and Gottschalk H (2001) Scattering theory for quantum fields with indefinite metric. *Communications in Mathematical Physics* 216: 491–513.

Albeverio S, Gottschalk H, and Wu J-L (1997) Models of local relativistic quantum fields with indefinite metric (in all dimensions). *Communications in Mathematical Physics* 184: 509–531.

Gottschalk H (2001) On the stability of one particle states generated by quantum fields fulfilling Yang-Feldman equations. *Reports on Mathematical Physics* 47(2): 241–246.

Grothaus M and Streit L (1999) Construction of relativistic quantum fields in the framework of white noise analysis. *Journal of Mathematical Physics* 40(11): 5387.

Morchio G and Strocchi F (1980) Infrared singularities, vacuum structure and pure phases in local quantum field theory. *Ann. Inst. H. Poincaré* 33: 251.

Steinmann O (2000) *Perturbative Quantum Electrodynamics and Axiomatic Field Theory*. Berlin: Springer.

Strocchi F (1993) *Selected Topics on the General Properties of Quantum Field Theory*. Lecture Notes in Physics, vol. 51. Singapore: World Scientific.

Perturbation Theory and Its Techniques

R J Szabo, Heriot-Watt University, Edinburgh, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

There are several equivalent formulations of the problem of quantizing an interacting field theory. The list includes canonical quantization, path-integral (or functional) techniques, stochastic quantization, "unified" methods such as the Batalin–Vilkovisky formalism, and techniques based on the realizations of field theories as low-energy limits of string theory. The problem of obtaining an exact nonperturbative description of a given quantum field theory is most often a very difficult one. Perturbative techniques, on the other hand, are abundant, and common to all of the quantization methods mentioned above is that they admit particle interpretations in this formalism.

The basic physical quantities that one wishes to calculate in a relativistic $(d+1)$-dimensional quantum field theory are the S-matrix elements

$$S_{ba} = {}_{\text{out}}\langle\psi_b(t)|\psi_a(t)\rangle_{\text{in}} \qquad [1]$$

between in and out states at large positive time t. The scattering operator S is then defined by writing [1] in terms of initial free-particle (descriptor) states as

$$S_{ba} =: \langle\psi_b(0)|\mathsf{S}|\psi_a(0)\rangle \qquad [2]$$

Suppose that the Hamiltonian of the given field theory can be written as $H = H_0 + H'$, where H_0 is the free part and H' the interaction Hamiltonian. The time evolutions of the in and out states are governed by the total Hamiltonian H. They can be expressed in terms of descriptor states which evolve in time with H_0 in the interaction picture and correspond to free-particle states. This leads to the Dyson formula

$$\mathsf{S} = \mathsf{T}\exp\left(-\mathrm{i}\int_{-\infty}^{\infty}\mathrm{d}t\,H_I(t)\right) \qquad [3]$$

where T denotes time ordering and $H_I(t) = \int \mathrm{d}^d x\,\mathcal{H}_{\text{int}}(x,t)$ is the interaction Hamiltonian in the interaction picture, with $\mathcal{H}_{\text{int}}(x,t)$ the interaction Hamiltonian density, which deals with essentially free fields. This formula expresses S in terms of interaction-picture operators acting on free-particle states in [2] and is the first step towards Feynman perturbation theory.

Notes:

For many analytic investigations, such as those which arise in renormalization theory, one is interested instead in the Green's functions of the quantum field theory, which measure the response of the system to an external perturbation. For definiteness, let us consider a free real scalar field theory in $d+1$ dimensions with Lagrangian density

$$\mathcal{L} = \tfrac{1}{2}\partial_\mu\phi\partial^\mu\phi - \tfrac{1}{2}m^2\phi^2 + \mathcal{L}_{\text{int}} \qquad [4]$$

where $\mathcal{L}_{\text{int}}$ is the interaction Lagrangian density which we assume has no derivative terms. The interaction Hamiltonian density is then given by $\mathcal{H}_{\text{int}} = -\mathcal{L}_{\text{int}}$. Introducing a real scalar source $J(x)$, we define the normalized "partition function" through the vacuum expectation values,

$$Z[J] = \frac{\langle 0|\mathsf{S}[J]|0\rangle}{\langle 0|\mathsf{S}[0]|0\rangle} \qquad [5]$$

where $|0\rangle$ is the normalized perturbative vacuum state of the quantum field theory given by (4) (defined to be destroyed by all field annihilation operators), and

$$\mathsf{S}[J] = \mathrm{T}\exp\left(\mathrm{i}\int \mathrm{d}^{d+1}x(\mathcal{L}_{\text{int}} + J(x)\phi(x))\right) \qquad [6]$$

from the Dyson formula. This partition function is the generating functional for all Green's functions of the quantum field theory, which are obtained from [5] by taking functional derivatives with respect to the source and then setting $J(x)=0$. Explicitly, in a formal Taylor series expansion in J one has

$$Z[J] = \sum_{n=0}^{\infty}\frac{\mathrm{i}^n}{n!}\prod_{i=1}^{n}\int \mathrm{d}^{d+1}x_i\, J(x_i)\, G^{(n)}(x_1,\ldots,x_n) \qquad [7]$$

whose coefficients are the Green's functions

$$G^{(n)}(x_1,\ldots,x_n) := \frac{\langle 0|\mathrm{T}[\exp\left(\mathrm{i}\int \mathrm{d}^{d+1}x\mathcal{L}_{\text{int}}\right)\phi(x_1)\cdots\phi(x_n)]|0\rangle}{\langle 0|\mathrm{T}\exp\left(\mathrm{i}\int \mathrm{d}^{d+1}x\mathcal{L}_{\text{int}}\right)|0\rangle} \qquad [8]$$

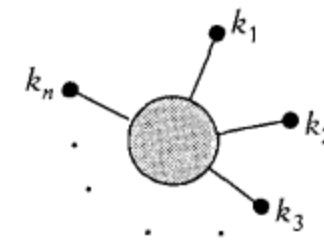
It is customary to work in momentum space by introducing the Fourier transforms

$$\begin{aligned}\tilde{J}(k) &= \int \mathrm{d}^{d+1}x\, \mathrm{e}^{\mathrm{i}k\cdot x}J(x)\\ \tilde{G}^{(n)}(k_1,\ldots,k_n) &= \prod_{i=1}^{n}\int \mathrm{d}^{d+1}x_i\, \mathrm{e}^{\mathrm{i}k_i\cdot x_i}G^{(n)}(x_1,\ldots,x_n)\end{aligned} \qquad [9]$$

in terms of which the expansion [7] reads

$$Z[J] = \sum_{n=0}^{\infty}\frac{\mathrm{i}^n}{n!}\prod_{i=1}^{n}\int \frac{\mathrm{d}^{d+1}k_i}{(2\pi)^{d+1}}\,\tilde{J}(-k_i) \times \tilde{G}^{(n)}(k_1,\ldots,k_n) \qquad [10]$$

The generating functional [10] can be written as a sum of Feynman diagrams with source insertions. Diagrammatically, the Green's function is an infinite series of graphs which can be represented symbolically as

$\tilde{G}^{(n)}(k_1,\ldots,k_n) =$ 

[11]

where the n external lines denote the source insertions of momenta k_i and the bubble denotes the sum over all Feynman diagrams constructed from the interaction vertices of $\mathcal{L}_{\text{int}}$.

This procedure is, however, rather formal in the way that we have presented it, for a variety of reasons. First of all, by Haag's theorem, it follows that the interaction representation of a quantum field theory does not exist unless a cutoff regularization is introduced into the interaction term in the Lagrangian density (this regularization is described explicitly below). The addition of this term breaks translation covariance. This problem can be remedied via a different definition of the regularized Green's functions, as we discuss below. Furthermore, the perturbation series of a quantum field theory is typically divergent. The expansion into graphs is, at best, an asymptotic series which is Borel summable. These shortcomings will not be emphasized any further in this article. Some mathematically rigorous approaches to perturbative quantum field theory can be found in the bibliography.

The Green's functions can also be used to describe scattering amplitudes, but there are two important differences between the graphs [11] and those which appear in scattering theory. In the present case, external lines carry propagators, that is, the free-field Green's functions

$$\begin{aligned}\Delta(x-y) &= \langle 0|\,\mathrm{T}[\phi(x)\,\phi(y)]\,|0\rangle\\ &= \left\langle x\middle|(\Box + m^2)^{-1}\middle|y\right\rangle\\ &= \int \frac{\mathrm{d}^{d+1}p}{(2\pi)^{d+1}}\frac{\mathrm{i}}{p^2 - m^2 + \mathrm{i}\epsilon}\mathrm{e}^{-\mathrm{i}p\cdot(x-y)}\end{aligned} \qquad [12]$$

where $\epsilon \to 0^+$ regulates the mass shell contributions, and their momenta k_i are off-shell in general ($k_i^2 \neq m^2$). By the LSZ theorem, the S-matrix element

is then given by the multiple on-shell residue of the Green's function in momentum space as

$$\langle k'_1,\ldots,k'_n|\,\mathsf{S}-\mathbb{1}|k_1,\ldots,k_l\rangle = \lim_{\substack{k'_1,\ldots,k'_n\to m^2\\ k_1,\ldots,k_l\to m^2}} \prod_{i=1}^{n}\frac{1}{\mathrm{i}\sqrt{c'_i}}\left(k'^2_i-m^2\right)\prod_{j=1}^{l}\frac{1}{\mathrm{i}\sqrt{c_j}}\left(k_j^2-m^2\right) \times \tilde{G}^{(n+m)}\left(-k'_1,\ldots,-k'_n,k_1,\ldots,k_l\right) \qquad [13]$$

where $\mathrm{i}c'_i, \mathrm{i}c_j$ are the residues of the corresponding particle poles in the exact two-point Green's function.

This article deals with the formal development and computation of perturbative scattering amplitudes in relativistic quantum field theory, along the lines outlined above. Initially we deal only with real scalar field theories of the sort [4] in order to illustrate the concepts and technical tools in as simple and concise a fashion as possible. These techniques are common to most quantum field theories. Fermions and gauge theories are then separately treated afterwards, focusing on the methods which are particular to them.

Diagrammatics

The pinnacle of perturbation theory is the technique of Feynman diagrams. Here we develop the basic machinery in a quite general setting and use it to analyze some generic features of the terms comprising the perturbation series.

Wick's Theorem

The Green's functions [8] are defined in terms of vacuum expectation values of time-ordered products of the scalar field $\phi(x)$ at different spacetime points. Wick's theorem expresses such products in terms of normal-ordered products, defined by placing each field creation operator to the right of each field annihilation operator, and in terms of two-point Green's functions [12] of the free-field theory (propagators). The consequence of this theorem is the Haffnian formula

$$\langle 0|\,\mathrm{T}[\phi(x_1)\cdots\phi(x_n)]|0\rangle = \begin{cases} 0 & \\ & n=2k-1 \\ \displaystyle\sum_{\pi\in S_{2k}}\prod_{i=1}^{k}\langle 0|\,\mathrm{T}\left[\phi(x_{\pi(2i-1)})\,\phi(x_{\pi(2i)})\right]|0\rangle & \\ & n=2k \end{cases} \qquad [14]$$

The formal Taylor series expansion of the scattering operator S may now be succinctly summarized into a diagrammatic notation by using Wick's theorem. For each spacetime integration $\int \mathrm{d}^{d+1}x_i$ we introduce a vertex with label i, and from each vertex there emanate some lines corresponding to field insertions at the point x_i. If the operators represented by two lines appear in a two-point function according to [14], that is, they are contracted, then these two lines are connected together. The S operator is then represented as a sum over all such Wick diagrams, bearing in mind that topologically equivalent diagrams correspond to the same term in S. Two diagrams are said to have the same pattern if they differ only by a permutation of their vertices. For any diagram $\mathfrak{D}$ with $n(\mathfrak{D})$ vertices, the number of ways of interchanging vertices is $n(\mathfrak{D})!$. The number of diagrams per pattern is always less than this number. The symmetry number $S(\mathfrak{D})$ of $\mathfrak{D}$ is the number of permutations of vertices that give the same diagram. The number of diagrams with the pattern of $\mathfrak{D}$ is then $n(\mathfrak{D})!/S(\mathfrak{D})$.

In a given pattern, we write the contribution to S of a single diagram $\mathfrak{D}$ as

$$\frac{1}{n(\mathfrak{D})!}:\theta(\mathfrak{D}):$$

where the combinatorial factor comes from the Taylor expansion of S, the large colons denote normal ordering of quantum operators, and $:\theta(\mathfrak{D}):$ contains spacetime integrals over normal-ordered products of the fields. Then all diagrams with the pattern of $\mathfrak{D}$ contribute $:\theta(\mathfrak{D}):/S(\mathfrak{D})$ to S. Only the connected diagrams $\mathfrak{D}_r, r\in\mathbb{N}$ (those in which every vertex is connected to every other vertex) contribute and we can write the scattering operator in a simple form which eliminates contributions from all disconnected diagrams as

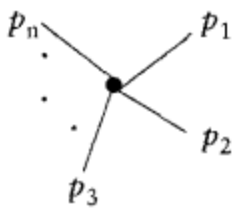
$$\mathsf{S} =: \exp\left(\sum_{r=1}^{\infty}\frac{\theta(\mathfrak{D}_r)}{S(\mathfrak{D}_r)}\right): \qquad [15]$$

Feynman Rules

Feynman diagrams in momentum space are defined from the Wick diagrams above by dropping the labels on vertices (and also the symmetry factors $S(\mathfrak{D})^{-1}$), and by labeling the external lines by the momenta of the initial and final particles that the corresponding field operators annihilate. In a spacetime interpretation, external lines

represent on-shell physical particles while internal lines of the graph represent off-shell virtual particles ($k^2 \neq m^2$). Physical particles interact via the exchange of virtual particles. An arbitrary diagram is then calculated via the Feynman rules:

$$\underset{p}{\text{———}} = \int \frac{\mathrm{d}^{d+1}p}{(2\pi)^{d+1}} \frac{\mathrm{i}}{p^2 - m^2 + \mathrm{i}\epsilon}$$



$$= \mathrm{i}g\,(2\pi)^{d+1}\,\delta^{(d+1)}(p_1 + \cdots + p_n) \qquad [16]$$

for a monomial interaction $\mathcal{L}_{\text{int}} = (g/n!)\phi^n$.

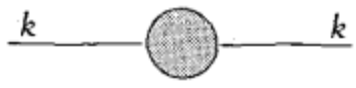
Irreducible Green's Functions

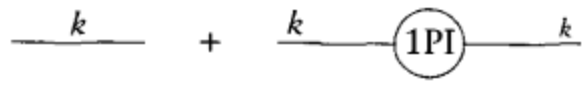
A one-particle irreducible (1PI) or proper Green's function is given by a sum of diagrams in which each diagram cannot be separated by cutting one internal line. In momentum space, it is defined without the overall momentum conservation delta-function factors and without propagators on external lines. For example, the two particle 1PI Green's function

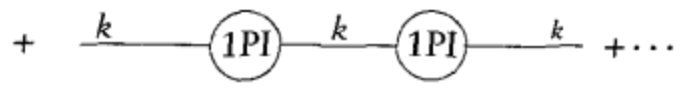


$$=: \Sigma(k) \qquad [17]$$

is called the self-energy. If $G(k)$ is the complete two-point function in momentum space, then one has

$G(k) :=$ 

$=$ 

$+$ 

$+\cdots$

$$= \frac{\mathrm{i}}{k^2 - m^2 - \Sigma(k)} \qquad [18]$$

and thus it suffices to calculate only 1PI diagrams.

The 1PI effective action, defined by the Legendre transformation $\Gamma[\phi] := -\mathrm{i}\ln Z[J] - \int \mathrm{d}^{d+1}x J(x)\phi(x)$ of [5], is the generating functional for proper vertex functions and it can be represented as a functional of only the vacuum expectation value of the field ϕ, that is, its classical value. In the semiclassical (WKB) approximation, the one-loop effective action is given by

$$\begin{aligned}\Gamma[\phi] &= S[\phi] + \frac{\mathrm{i}\hbar}{2}\mathrm{Tr}\ln(1 + \Delta V''[\phi]) + O(\hbar^2) \\ &= S[\phi] + \mathrm{i}\hbar \sum_{n=1}^{\infty} \frac{(-1)^n}{2n} \\ &\quad \times \prod_{i=1}^{n} \int \mathrm{d}^{d+1}x_i \Delta(x_i - x_{i+1}) V''[\phi(x_{i+1})] \\ &\quad + O(\hbar^2)\end{aligned} \qquad [19]$$

where we have denoted $S[\phi] = \int \mathrm{d}^{d+1}x\mathcal{L}$ and $V[\phi] = -\mathcal{L}_{\text{int}}$, and for each term in the infinite series we define $x_{n+1} := x_1$. The first term in [19] is the classical contribution and it can be represented in terms of connected tree diagrams. The second term is the sum of contributions of one-loop diagrams constructed from n propagators $-\mathrm{i}\Delta(x - y)$ and n vertices $-\mathrm{i}V''[\phi]$. The expansion may be carried out to all orders in terms of connected Feynman diagrams, and the result of the above Legendre transformation is to select only the one-particle irreducible diagrams and to replace the classical value of ϕ by an arbitrary argument. All information about the quantum field theory is encoded in this effective action.

Parametric Representation

Consider an arbitrary proper Feynman diagram $\mathfrak{D}$ with n internal lines and v vertices. The number, ℓ, of independent loops in the diagram is the number of independent internal momenta in $\mathfrak{D}$ when conservation laws at each vertex have been taken into account, and it is given by $\ell = n + 1 - v$. There is an independent momentum integration variable k_i for each loop, and a propagator for each internal line as in [16]. The contribution of $\mathfrak{D}$ to a proper Green's function with r incoming external momenta p_i, with $\sum_{i=1}^{r} p_i = 0$, is given by

$$\begin{aligned}\tilde{I}_{\mathfrak{D}}(p) = \frac{V(\mathfrak{D})}{S(\mathfrak{D})} \prod_{i=1}^{n} \int \frac{\mathrm{d}^{d+1}k_i}{(2\pi)^{d+1}} \frac{\mathrm{i}}{k_i^2 - m^2 + \mathrm{i}\epsilon} \\ \times \prod_{j=1}^{v} (2\pi)^{d+1} \delta^{(d+1)}(P_j - K_j)\end{aligned} \qquad [20]$$

where $V(\mathfrak{D})$ contains all contributions from the interaction vertices of $\mathcal{L}_{\text{int}}$, and P_j (resp. K_j) is the sum of incoming external momenta p_{l_j} (resp. internal momenta k_{l_j}) at vertex j with respect to a fixed chosen orientation of the lines of the graph. After resolving the delta-functions in terms of independent internal loop momenta $k_1, \ldots, k_\ell$ and dropping the overall momentum conservation

delta-function along with the symmetry and vertex factors in [20], one is left with a set of momentum space integrals

$$I_{\mathcal{D}}(p) = \prod_{i=1}^{\ell} \int \frac{\mathrm{d}^{d+1}k_i}{(2\pi)^{d+1}} \prod_{j=1}^{n} \frac{\mathrm{i}}{a_j(k,p) + \mathrm{i}\epsilon} \qquad [21]$$

where $a_j(k,p)$ are functions of both the internal and external momenta.

It is convenient to exponentiate propagators using the Schwinger parametrization

$$\frac{\mathrm{i}}{a_j + \mathrm{i}\epsilon} = \int_0^\infty \mathrm{d}\alpha_j \, \mathrm{e}^{\mathrm{i}\alpha_j(a_j + \mathrm{i}\epsilon)} \qquad [22]$$

and after some straightforward manipulations one may write the Feynman parametric formula

$$\prod_{j=1}^{n} \frac{\mathrm{i}}{a_j(k,p) + \mathrm{i}\epsilon} = (n-1)! \prod_{j=1}^{n} \int_0^1 \mathrm{d}\alpha_j \frac{\delta\left(1 - \sum_j \alpha_j\right)}{D_{\mathcal{D}}(k;\alpha,p)^n} \qquad [23]$$

where $D_{\mathcal{D}}(k;\alpha,p) := \sum_j \alpha_j [a_j(k,p) + \mathrm{i}\epsilon]$ is generically a quadratic form

$$D_{\mathcal{D}}(k;\alpha,p) = \frac{1}{2} \sum_{i,j=1}^{\ell} k_i \cdot Q_{ij}(\alpha) k_j + \sum_{i=1}^{\ell} L_i(p) \cdot k_i + \lambda(p^2) \qquad [24]$$

The positive symmetric matrix Q_{ij} is independent of the external momenta p_l, invertible, and has nonzero eigenvalues $Q_1, \ldots, Q_\ell$. The vectors L_i^μ are linear combinations of the p_j^μ, while $\lambda(p^2)$ is a function of only the Lorentz invariants p_i^2. After some further elementary manipulations, the loop diagram contribution [21] may be written as

$$I_{\mathcal{D}}(p) = (n-1)! \prod_{j=1}^{n} \int_0^1 \mathrm{d}\alpha_j \prod_{i=1}^{\ell} \frac{1}{Q_i(\alpha)^2} \int \frac{\mathrm{d}^{d+1}k_i}{(2\pi)^{d+1}} \delta\left(1 - \sum_j \alpha_j\right) \times \left(\frac{1}{2}\sum_i k_i^2 + \lambda(p^2) - \frac{1}{2}\sum_{i,j} L_i(p) \cdot Q^{-1}(\alpha)_{ij} L_j(p)\right)^{-n} \qquad [25]$$

Finally, the integrals over the loop momenta k_i may be performed by Wick-rotating them to Euclidean space and using the fact that the combination of ℓ integrations in $\mathbb{R}^{d+1}$ has $O((d+1)\ell)$ rotational invariance. The contribution from the entire Feynman diagram $\mathcal{D}$ thereby reduces to the calculation of the parametric integrals:

$$I_{\mathcal{D}}(p) = \frac{\Gamma\left(n - \frac{(d+1)\ell}{2}\right)}{(2\pi)^{\frac{(d+1)\ell}{2}} \mathrm{i}^{d\ell}} \prod_{j=1}^{n} \int_0^1 \mathrm{d}\alpha_j \prod_{i=1}^{\ell} \frac{1}{Q_i(\alpha)^2} \times \frac{\delta\left(1 - \sum_j \alpha_j\right)}{\left(\lambda(p^2) - \frac{1}{2}\sum_{i,j} L_i(p) \cdot Q^{-1}(\alpha)_{ij} L_j(p)\right)^{\frac{n-(d+1)\ell}{2}}} \qquad [26]$$

where $\Gamma(s)$ is the Euler gamma-function.

Regularization

The parametric representation [26] is generically convergent when $2n - (d+1)\ell > 0$. When divergent, the infinities arise from the lower limits of integration $\alpha_j \to 0$. This is just the parametric representation of the large-k divergence of the original Feynman amplitude [20]. Such ultraviolet divergences plague the very meaning of a quantum field theory and must be dealt with in some way. We will now quickly tour the standard methods of ultraviolet regularization for such loop integrals, which is a prelude to the renormalization program that removes the divergences (in a renormalizable field theory). Here we consider regularization simply as a means of justification for the various formal manipulations that are used in arriving at expressions such as [26].

Momentum Cutoff

Cutoff regularization introduces a mass scale Λ into the quantum field theory and throws away the Fourier modes of the fields for spatial momenta $\boldsymbol{k}$ with $|\boldsymbol{k}| > \Lambda$. This regularization spoils Lorentz invariance. It is also nonlocal. For example, if we restrict to a hypercube in momentum space, so that $|k_i| < \Lambda$ for $i = 1, \ldots, d$, then

$$\int_{|\boldsymbol{k}|>\Lambda} \frac{\mathrm{d}^d\boldsymbol{k}}{(2\pi)^d} \mathrm{e}^{\mathrm{i}k\cdot x} = \prod_{i=1}^{d} \frac{\sin(\Lambda x^i)}{\pi x^i}$$

which is a delta-function in the limit $\Lambda \to \infty$ but is nonlocal for $\Lambda < \infty$. The regularized field theory is finite order by order in perturbation theory and depends on the cutoff Λ.

Lattice Regularization

We can replace the spatial continuum by a lattice $\mathfrak{L}$ of rank d and define a Lagrangian on $\mathfrak{L}$ by

$$L_{\mathfrak{L}} = \frac{1}{2}\sum_{i\in S(\mathfrak{L})} \dot{\phi}_i^2 + J \sum_{\langle i,j\rangle \in L(\mathfrak{L})} \phi_i \phi_j + \sum_{i\in S(\mathfrak{L})} V(\phi_i) \quad [27]$$

where $S(\mathfrak{L})$ is the set of sites i of the lattice on each of which is situated a time-dependent function ϕ_i, and $L_{\mathfrak{L}}$ is the collection of links connecting pairs $\langle i,j\rangle$ of nearest-neighbor sites i,j on $\mathfrak{L}$. The regularized field theory is now local, but still has broken Lorentz invariance. In particular, it suffers from broken rotational symmetry. If $\mathfrak{L}$ is hypercubic with lattice spacing a, that is, $\mathfrak{L}=(\mathbb{Z}a)^d$, then the momentum cutoff is at $\Lambda = a^{-1}$.

Pauli–Villars Regularization

We can replace the propagator $\mathrm{i}(k^2 - m^2 + \mathrm{i}\epsilon)^{-1}$ by $\mathrm{i}(k^2 - m^2 + \mathrm{i}\epsilon)^{-1} + \mathrm{i}\sum_{j=1}^{N} c_j (k^2 - M_j^2 + \mathrm{i}\epsilon)^{-1}$, where the masses $M_j \gg m$ are identified with the momentum cutoff as $\min\{M_j\} = \Lambda \to \infty$. The mass-dependent coefficients c_j are chosen to make the modified propagator decay rapidly as $(k^2)^{-N-1}$ at $k \to \infty$, which gives the N equations $(m^2)^i + \sum_j c_j (M_j^2)^i = 0, i = 0, 1, \ldots, N-1$. This regularization preserves Lorentz invariance (and other symmetries that the field theory may possess) and is local in the following sense. The modified propagator can be thought of as arising through the alteration of the Lagrangian density [4] by N additional scalar fields φ_j of masses M_j with

$$\mathcal{L}_{\mathrm{PV}} = \tfrac{1}{2}\partial_\mu\phi\partial^\mu\phi - \tfrac{1}{2}m^2\phi^2 + \sum_{j=1}^{N}\left(\tfrac{1}{2}\partial_\mu\varphi_j\partial^\mu\varphi_j - \tfrac{1}{2}M_j^2\varphi_j^2\right) + \mathcal{L}_{\mathrm{int}}[\Phi] \quad [28]$$

where $\Phi := \phi + \sum_j \sqrt{c_j}\varphi_j$. The contraction of the Φ field thus produces the required propagator. However, the c_j's as computed above are generically negative numbers and so the Lagrangian density [28] is not Hermitian (as $\Phi \neq \Phi^\dagger$). It is possible to make [28] formally Hermitian by redefining the inner product on the Hilbert space of physical states, but this produces negative-norm states. This is no problem at energy scales $E \ll M_j$ on which the extra particles decouple and the negative probability states are invisible.

Dimensional Regularization

Consider a Euclidean space integral $\int \mathrm{d}^4k (k^2 + a^2)^{-r}$ arising after Wick rotation from some loop diagram in $(3+1)$-dimensional scalar field theory. We replace this integral by its D-dimensional version

$$\int \frac{\mathrm{d}^D k}{(k^2 + a^2)^r} = \frac{\pi^{D/2}(a^2)^{D/2-r}}{(r-1)!}\Gamma\left(r - \frac{D}{2}\right) \quad [29]$$

This integral is absolutely convergent for $D < 2r$. We can analytically continue the result of this integration to the complex plane $D \in \mathbb{C}$. As an analytic function, the only singularities of the Euler function $\Gamma(z)$ are poles at $z = 0, -1, -2, \ldots$. In particular, $\Gamma(z)$ has a simple pole at $z = 0$ of residue 1. If we write $D = 4 + \epsilon$ with $|\epsilon| \to 0$, then the integral [29] is proportional to $\Gamma(r - 2 - \epsilon/2)$ and ϵ plays the role of the regulator here. This regularization is Lorentz invariant (in D dimensions) and is distinguished as having a dimensionless regularization parameter ϵ. This parameter is related to the momentum cutoff Λ by $\epsilon^{-1} = \ln(\Lambda/m)$, so that the limit $\epsilon \to 0$ corresponds to $\Lambda \to \infty$.

Infrared Divergences

Thus far we have only considered the ultraviolet behavior of loop amplitudes in quantum field theory. When dealing with massless particles ($m = 0$ in [4]) one has to further worry about divergences arising from the $k \to 0$ regions of Feynman integrals. After Wick rotation to Euclidean momenta, one can show that no singularities arise in a given Feynman diagram as some of its internal masses vanish provided that all vertices have superficial degree of divergence $d + 1$, the external momenta are not exceptional (i.e., no partial sum of the incoming momenta p_i vanishes), and there is at most one soft external momentum. This result assumes that renormalization has been carried out at some fixed Euclidean point. The extension of this property when the external momenta are continued to physical on-shell values is difficult. The Kinoshita–Lee–Nauenberg theorem states that, as a consequence of unitarity, transition probabilities in a theory involving massless particles are finite when the sum over all degenerate states (initial and final) is taken. This is true order by order in perturbation theory in bare quantities or if minimal subtraction renormalization is used (to avoid infrared or mass singularities in the renormalization constants).

Fermion Fields

We will now leave the generalities of our pure scalar field theory and start considering the extensions of our previous considerations to other types of particles. Henceforth we will primarily deal with the case of $(3+1)$-dimensional spacetime. We begin by indicating how the rudiments of perturbation

theory above apply to the case of Dirac fermion fields. The Lagrangian density is

$$\mathcal{L}_F = \overline{\psi}(\mathrm{i}\partial\!\!\!/ - m)\psi + \mathcal{L}' \qquad [30]$$

where ψ are four-component Dirac fermion fields in $3+1$ dimensions, $\overline{\psi} := \psi^\dagger\gamma^0$ and $\partial\!\!\!/ = \gamma^\mu\partial_\mu$ with γ^μ the generators of the Clifford algebra $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$. The Lagrangian density $\mathcal{L}'$ contains couplings of the Dirac fields to other field theories, such as the scalar field theories considered previously.

Wick's theorem for anticommuting Fermi fields leads to the Pfaffian formula

$$\langle 0|\mathrm{T}[\psi(1)\cdots\psi(n)]|0\rangle = \begin{cases} 0, & n = 2k-1 \\ \dfrac{1}{2^k k!}\displaystyle\sum_{\pi\in S_{2k}} \mathrm{sgn}(\pi) & \\ \times \displaystyle\prod_{i=1}^{k} \langle 0|\mathrm{T}[\psi(\pi(2i-1))\psi(\pi(2i))]|0\rangle & \\ n = 2k & \end{cases} \qquad [31]$$

where for compactness we have written in the argument of $\psi(i)$ the spacetime coordinate, the Dirac index, and a discrete index which distinguishes ψ from $\overline{\psi}$. The nonvanishing contractions in [31] are determined by the free-fermion propagator

$$\begin{aligned}\Delta_F(x-y) &= \langle 0|\mathrm{T}[\psi(x)\overline{\psi}(y)]|0\rangle \\ &= \left\langle x\left|(\mathrm{i}\partial\!\!\!/ - m)^{-1}\right|y\right\rangle \\ &= \mathrm{i}\int \frac{\mathrm{d}^4p}{(2\pi)^4}\frac{p\!\!\!/ + m}{p^2 - m^2 + \mathrm{i}\epsilon}\mathrm{e}^{-\mathrm{i}p\cdot(x-y)} \end{aligned} \qquad [32]$$

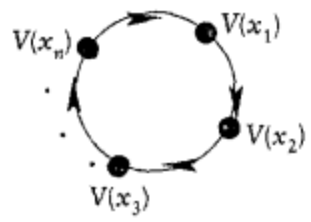
Perturbation theory now proceeds exactly as before. Suppose that the coupling Lagrangian density in [30] is of the form $\mathcal{L}' = \overline{\psi}(x)V(x)\psi(x)$. Both the Dyson formula [3] and the diagrammatic formula [15] are formally the same in this instance. For example, in the formal expansion in powers of $\int \mathrm{d}^4x\mathcal{L}'$, the vacuum-to-vacuum amplitude (the denominator in [5]) will contain field products of the form

$$\prod_{i=1}^{n}\int \mathrm{d}^4x_i\langle 0|\mathrm{T}[\overline{\psi}(x_i)V(x_i)\psi(x_i)]|0\rangle$$

which correspond to fermion loops. Before applying Wick's theorem, the fields must be rearranged as

$$\mathrm{tr}\prod_{i=1}^{n} V(x_i)\psi(x_i)\overline{\psi}(x_{i+1})$$

(with $x_{n+1} := x_1$), where tr is the 4×4 trace over spinor indices. This reordering introduces the familiar minus sign for a closed fermion loop, and one has



$$\begin{aligned} = (-)\prod_{i=1}^{n}\int \mathrm{d}^4x_i & \\ \times \mathrm{tr}\prod_{j=1}^{n-1}\Delta_F(x_j - x_{j+1}) & \\ \times V(x_{j+1})\,\Delta_F(x_{j+1} - x_{j+2}) & \end{aligned} \qquad [33]$$

Feynman rules are now described as follows. Fermion lines are oriented to distinguish a particle from its corresponding antiparticle, and carry both a four-momentum label p as well as a spin polarization index $r = 1, 2$. Incoming fermions (resp. antifermions) are described by the wave functions $u_p^{(r)}$ (resp. $\overline{v}_p^{(r)}$), while outgoing fermions (resp. antifermions) are described by the wave functions $\overline{u}_p^{(r)}$ (resp. $v_p^{(r)}$). Here $u_p^{(r)}$ and $v_p^{(r)}$ are the classical spinors, that is, the positive and negative-energy solutions of the Dirac equation $(p\!\!\!/ - m)u_p^{(r)} = (p\!\!\!/ + m)v_p^{(r)} = 0$. Matrices are multiplied along a Fermi line, with the head of the arrow on the left. Closed fermion loops produce an overall minus sign as in [33], and the multiplication rule gives the trace of Dirac matrices along the lines of the loop. Unpolarized scattering amplitudes are summed over the spins of final particles and averaged over the spins of initial particles using the completeness relations for spinors

$$\sum_{r=1,2} u_p^{(r)}\overline{u}_p^{(r)} = p\!\!\!/ + m, \quad \sum_{r=1,2} v_p^{(r)}\overline{v}_p^{(r)} = p\!\!\!/ - m \qquad [34]$$

leading to basis-independent results. Polarized amplitudes are computed using the spinor bilinears $\overline{u}_p^{(r)}\gamma^\mu u_p^{(s)} = \overline{v}_p^{(r)}\gamma^\mu v_p^{(s)} = 2p^\mu\delta^{rs}$, $\overline{u}_p^{(r)}u_p^{(s)} = -\overline{v}_p^{(r)}v_p^{(s)} = 2m\,\delta^{rs}$, and $\overline{u}_p^{(r)}v_p^{(s)} = 0$.

When calculating fermion loop integrals using dimensional regularization, one utilizes the Dirac algebra in D dimensions

$$\begin{aligned} \gamma^\mu\gamma_\mu &= \eta^\mu_\mu = D \\ \gamma_\mu p\!\!\!/\,\gamma^\mu &= (2-D)p\!\!\!/ \\ \gamma^\mu p\!\!\!/\,k\!\!\!/\,\gamma_\mu &= 4p\cdot k + (D-4)p\!\!\!/\,k\!\!\!/ \\ \gamma^\mu p\!\!\!/\,k\!\!\!/\,q\!\!\!/\,\gamma_\mu &= -2q\!\!\!/\,k\!\!\!/\,p\!\!\!/ - (D-4)p\!\!\!/\,k\!\!\!/\,q\!\!\!/ \\ \mathrm{tr}\,\mathbb{1} = 4, \mathrm{tr}\gamma^{\mu_1}\cdots\gamma^{\mu_{2k-1}} &= 0, \mathrm{tr}\gamma^\mu\gamma^\nu = 4\eta^{\mu\nu} \\ \mathrm{tr}\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma &= 4(\eta^{\mu\nu}\eta^{\rho\sigma} - \eta^{\mu\rho}\eta^{\nu\sigma} \\ &\quad + \eta^{\mu\sigma}\eta^{\nu\rho}) \end{aligned} \qquad [35]$$

Specific to $D=4$ dimensions are the trace identities

$$\begin{aligned} \mathrm{tr}\gamma^5 &= \mathrm{tr}\gamma^\mu\gamma^\nu\gamma^5 = 0, \\ \mathrm{tr}\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma\gamma^5 &= -4\mathrm{i}\epsilon^{\mu\nu\rho\sigma} \end{aligned} \qquad [36]$$

where $\gamma^5 := \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3$. Finally, loop diagrams evaluated with the fermion propagator [32] require a generalization of the momentum space integral [29] given by

$$\begin{aligned} &\int \frac{\mathrm{d}^D k}{(2\pi)^D} \frac{1}{(k^2 + 2k\cdot p + a^2 + \mathrm{i}\epsilon)^r} \\ &= \frac{\mathrm{i}(-\pi)^{D/2}\Gamma\left(r - \frac{D}{2}\right)}{(2\pi)^D (r-1)!} \frac{1}{(a^2 - p^2 + \mathrm{i}\epsilon)^{r-D/2}} \end{aligned} \qquad [37]$$

From this formula we can extract expressions for more complicated Feynman integrals which are tensorial, that is, which contain products of momentum components k^μ in the numerators of their integrands, by differentiating [37] with respect to the external momentum p^μ.

Gauge Fields

The issues we have dealt with thus far have interesting difficulties when dealing with gauge fields. We will now discuss some general aspects of the perturbation expansion of gauge theories using as prototypical examples quantum electrodynamics (QED) and quantum chromodynamics (QCD) in four spacetime dimensions.

Quantum Electrodynamics

Consider the QED Lagrangian density

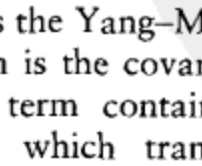
$$\begin{aligned} \mathcal{L}_{\mathrm{QED}} = &-\tfrac{1}{4} F_{\mu\nu}F^{\mu\nu} \\ &+ \tfrac{1}{2}\mu^2 A_\mu A^\mu + \bar{\psi}(\mathrm{i}\partial\!\!\!/ - eA\!\!\!/ - m)\psi \end{aligned} \qquad [38]$$

where A_μ is a U(1) gauge field in $3+1$ dimensions and $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is its field strength tensor. We have added a small mass term $\mu \to 0$ for the gauge field, which at the end of calculations should be taken to vanish in order to describe real photons (as opposed to the soft photons described by [38]). This is done in order to cure the infrared divergences generated in scattering amplitudes due to the masslessness of the photon, that is, the long-range nature of the electromagnetic interaction. The Bloch–Nordsieck theorem in QED states that infrared divergences cancel for physical processes, that is, for processes with an arbitrary number of undetectable soft photons.

Perturbation theory proceeds in the usual way via the Dyson formula, Wick's theorem, and Feynman diagrams. The gauge field propagator is given by

$$\begin{aligned} &\langle 0|\mathrm{T}[A_\mu(x)A_\nu(y)]|0\rangle \\ &= \langle x|\left[\eta_{\mu\nu}(\Box + \mu^2) - \partial_\mu\partial_\nu\right]^{-1}|y\rangle \\ &= \mathrm{i}\int \frac{\mathrm{d}^4 p}{(2\pi)^4} \frac{-\eta_{\mu\nu} + \frac{p_\mu p_\nu}{\mu^2}}{p^2 - \mu^2 + \mathrm{i}\epsilon} \mathrm{e}^{-\mathrm{i}p\cdot(x-y)} \end{aligned} \qquad [39]$$

and is represented by a wavy line. The fermion–fermion–photon vertex is



$$= -\mathrm{i}e\gamma_\mu \qquad [40]$$

An incoming (resp. outgoing) soft photon of momentum k and polarization r is described by the wave function $e_\mu^{(r)}(k)$ (resp. $e_\mu^{(r)}(k)^*$), where the polarization vectors $e_\mu^{(r)}(k), r = 1, 2, 3$ solve the vector field wave equation $(\Box + \mu^2)A_\mu = \partial_\mu A^\mu = 0$ and obey the orthonormality and completeness conditions

$$\begin{aligned} e^{(r)}(k)^* \cdot e^{(s)}(k) &= -\delta^{rs} \\ \sum_{r=1}^{3} e_\mu^{(r)}(k) e_\nu^{(r)}(k)^* &= -\eta_{\mu\nu} + \frac{k_\mu k_\nu}{\mu^2} \end{aligned} \qquad [41]$$

along with $k \cdot e^{(r)}(k) = 0$. All vector indices are contracted along the lines of the Feynman graph. All other Feynman rules are as previously.

Quantum Chromodynamics

Consider nonabelian gauge theory in $3+1$ dimensions minimally coupled to a set of fermion fields $\psi^A, A = 1, \ldots, N_f$, each transforming in the fundamental representation of the gauge group G whose generators T^a satisfy the commutation relations $[T^a, T^b] = f^{abc}T^c$. The Lagrangian density is given by

$$\begin{aligned} \mathcal{L}_{\mathrm{QCD}} = &-\frac{1}{4} F^a_{\mu\nu} F^{a\,\mu\nu} + \frac{1}{2\alpha}\left(\partial_\mu A^{a\,\mu}\right)^2 + \partial_\mu\bar{\eta} D^\mu \eta \\ &+ \sum_{A=1}^{N_f} \bar{\psi}^A (\mathrm{i}D\!\!\!/ - m_A)\psi^A \end{aligned} \qquad [42]$$

where $F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + f^{abc}A^b_\mu A^c_\nu$ and $D_\mu = \partial_\mu + \mathrm{i}eR(T^a)A^a_\mu$, with R the pertinent representation of G ($R(T^a)_{bc} = f^a_{bc}$ for the adjoint representation and $R(T^a) = T^a$ for the fundamental representation). The first term is the Yang–Mills Lagrangian density, the second term is the covariant gauge-fixing term, and the third term contains the Faddeev–Popov ghost fields η which transform in the adjoint representation of the gauge group.

Feynman rules are straightforward to write down and are given in **Figure 1** where wavy lines

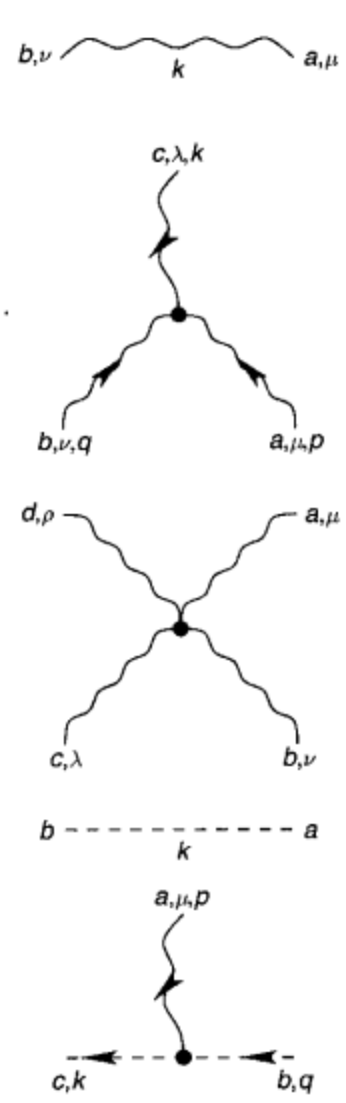


Figure 1 Feynman rules.

represent gluons and dashed lines represent ghosts. Feynman rules for the fermions are exactly as before, except that now the vertex [40] is multiplied by the color matrix T^a. All color indices are contracted along the lines of the Feynman graph. Color factors may be simplified by using the identities

$$\begin{aligned} &\mathrm{Tr}\, R^a R^b = \frac{\dim R}{\dim G} C_2(R)\delta^{ab}, \quad R^a R^a = C_2(R) \\ &R^a R^b R^a = \left(C_2(R) - \frac{1}{2} C_2(G)\right) R^b \end{aligned} \qquad [43]$$

where $R^a := R(T^a)$ and $C_2(R)$ is the quadratic Casimir invariant of the representation R (with value $C_2(G)$ in the adjoint representation). For $G = \mathrm{SU}(N)$, one has $C_2(G) = N$ and $C_2(N) = (N^2 - 1)/2N$ for the fundamental representation.

The cancellation of infrared divergences in loop amplitudes of QCD is far more delicate than in QED, as there is no analog of the Bloch–Nordsieck theorem in this case. The Kinoshita–Lee–Nauenberg theorem guarantees that, at the end of any perturbative calculation, these divergences must cancel for any appropriately defined physical quantity. However, at a given order of perturbation theory, a physical quantity typically involves both virtual and real emission contributions that are separately infrared divergent. Already at two-loop level these divergences have a highly intricate structure. Their precise form is specified by the Catani color-space factorization formula, which also provides an efficient way of organizing amplitudes into divergent parts, which ultimately drop out of physical quantities, and finite contributions.

The computation of multigluon amplitudes in nonabelian gauge theory is rather complicated when one uses polarization states of vector bosons. A much more efficient representation of amplitudes is provided by adopting a helicity (or circular polarization) basis for external gluons. In the spinor–helicity formalism, one expresses positive and negative-helicity polarization vectors in terms of massless Weyl spinors $|k^\pm\rangle := \frac{1}{2}(1 \pm \gamma_5)u_k = \frac{1}{2}(1 \pm \gamma_5)v_k$ through

$$e_\mu^\pm(k;q) = \pm \frac{\langle q^\mp | \gamma_\mu | k^\mp \rangle}{\sqrt{2}\langle q^\mp | k^\pm \rangle} \qquad [44]$$

where q is an arbitrary null reference momentum which drops out of the final gauge-invariant amplitudes. The spinor products are crossing symmetric, antisymmetric in their arguments, and satisfy the identities

$$\left\langle k_i^- | k_j^+ \right\rangle \left\langle k_j^+ | k_i^- \right\rangle = 2k_i \cdot k_j$$

$$\left\langle k_i^- \middle| k_j^+ \right\rangle \langle k_l^- | k_r^+ \rangle = \langle k_i^- | k_r^+ \rangle \left\langle k_l^- | k_j^+ \right\rangle + \langle k_i^- | k_l^+ \rangle \left\langle k_j^- | k_r^+ \right\rangle \quad [45]$$

Any amplitude with massless external fermions and vector bosons can be expressed in terms of spinor products. Conversely, the spinor products offer the most compact representation of helicity amplitudes which can be related to more conventional amplitudes described in terms of Lorentz invariants. For loop amplitudes, one uses a dimensional regularization scheme in which all helicity states are kept four dimensional and only internal loop momenta are continued to $D = 4 + \epsilon$ dimensions.

Computing Loop Integrals

At the very heart of perturbative quantum field theory is the problem of computing Feynman integrals for multiloop scattering amplitudes. The integrations typically involve serious technical challenges and for the most part are intractable by straightforward analytical means. We will now survey some of the computational techniques that have been developed for calculating quantum loop amplitudes which arise in the field theories considered previously.

Asymptotic Expansion

In many physical instances one is interested in scattering amplitudes in certain kinematical limits. In this case one may perform an asymptotic expansion of multiloop diagrams whose coefficients are typically nonanalytic functions of the perturbative expansion parameter $\hbar$. The main simplification which arises comes from the fact that the expansions are done before any momentum integrals are evaluated. In the limits of interest, Taylor series expansions in different selected regions of each loop momentum can be interpreted in terms of subgraphs and co-subgraphs of the original Feynman diagram.

Consider a Feynman diagram $\mathfrak{D}$ which depends on a collection $\{Q_i\}$ of large momenta (or masses), and a collection $\{m_i, q_i\}$ of small masses and momenta. The prescription for the large-momentum asymptotic expansion of $\mathfrak{D}$ may be summarized in the diagrammatic formula

$$\lim_{Q\to\infty} \mathfrak{D}(Q; m, q) = \sum_{\mathfrak{d}\subset\mathfrak{D}} (\mathfrak{D}/\mathfrak{d})(m, q) \star (\mathcal{T}_{\{m_\mathfrak{d}, q_\mathfrak{d}\}}\mathfrak{d})(Q; m_\mathfrak{d}, q_\mathfrak{d}) \quad [46]$$

where the sum runs through all subgraphs $\mathfrak{d}$ of $\mathfrak{D}$ which contain all vertices where a large momentum enters or leaves the graph and is one-particle irreducible after identifying these vertices. The operator $\mathcal{T}_{\{m_\mathfrak{d}, q_\mathfrak{d}\}}$ performs a Taylor series expansion before any integration is carried out, and the notation $(\mathfrak{D}/\mathfrak{d}) \star (\mathcal{T}_{\{m_\mathfrak{d}, q_\mathfrak{d}\}}\mathfrak{d})$ indicates that the subgraph $\mathfrak{d} \subset \mathfrak{D}$ is replaced by its Taylor expansion in all masses and external momenta of $\mathfrak{d}$ that do not belong to the set $\{Q_i\}$. The external momenta of $\mathfrak{d}$ which become loop momenta in $\mathfrak{D}$ are also considered to be small. The loop integrations are then performed only after all these expansions have been carried out. The diagrams $\mathfrak{D}/\mathfrak{d}$ are called co-subgraphs.

The subgraphs become massless integrals in which the scales are set by the large momenta. For instance, in the simplest case of a single large momentum Q one is left with integrals over propagators. The co-subgraphs may contain small external momenta and masses, but the resulting integrals are typically much simpler than the original one. A similar formula is true for large-mass expansions, with the vertex conditions on subdiagrams replace by propagator conditions. For example, consider the asymptotic expansion of the two-loop double bubble diagram (**Figure 2**) in the region $q^2 \ll m^2$, where m is the mass of the inner loop. The subgraphs (to the right of the stars) are expanded in all external momenta including q and reinserted into the fat vertices of the co-subgraphs (to the left of the stars). Once such asymptotic expansions are carried out, one may attempt to reconstruct as much information as possible about the given scattering amplitude

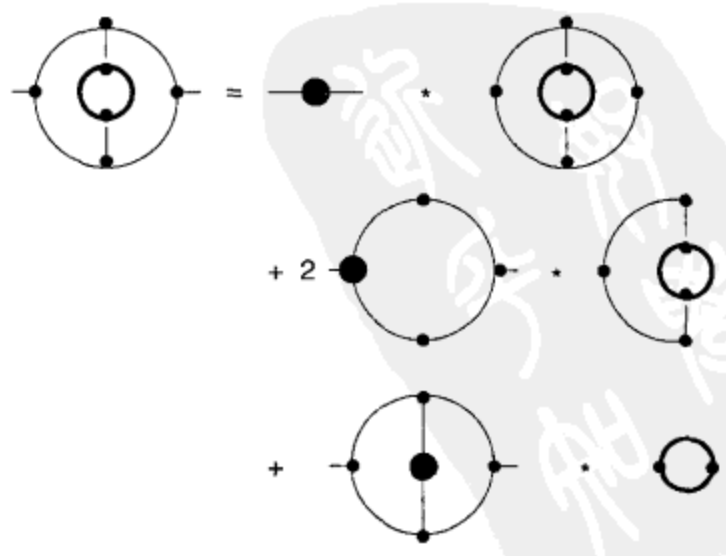


Figure 2 Asymptotic expansion of the two-loop double bubble diagram.

by using the method of Padé approximation which requires knowledge of only part of the expansion of the diagram. By construction, the Padé approximation has the same analytic properties as the exact amplitude.

Brown–Feynman Reduction

When considering loop diagrams which involve fermions or gauge bosons, one encounters tensorial Feynman integrals. When these involve more than three distinct denominator factors (propagators), they require more than two Feynman parameters for their evaluation and become increasingly complicated. The Brown–Feynman method simplifies such higher-rank integrals and effectively reduces them to scalar integrals which typically require fewer Feynman parameters for their evaluation.

To illustrate the idea behind this method, consider the one-loop rank-3 tensor Feynman integral

$$J^{\mu\nu\lambda} = \int \frac{d^D k}{(2\pi)^D} \times \frac{k^\mu k^\nu k^\lambda}{k^2(k^2-\mu^2)(q-k)^2((k-q)^2+\mu^2)(k^2+2k\cdot p)} \quad [47]$$

where p and q are external momenta with the mass-shell conditions $p^2=(p-q)^2=m^2$. By Lorentz invariance, the general structure of the integral [47] will be of the form

$$J^{\mu\nu\lambda} = a^{\mu\nu}p^\lambda + b^{\mu\nu}q^\lambda + c^\mu s^{\nu\lambda} + c^\nu s^{\mu\lambda} \quad [48]$$

where $a^{\mu\nu}, b^{\mu\nu}$ are tensor-valued functions and c^μ a vector-valued function of p and q. The symmetric tensor $s^{\mu\nu}$ is chosen to project out components of vectors transverse to both p and q, i.e., $p_\mu s^{\mu\nu} = q_\mu s^{\mu\nu} = 0$, with the normalization $s_\mu{}^\mu = D-2$. Solving these constraints leads to the explicit form

$$s^{\mu\nu} = \eta^{\mu\nu} - \frac{m^2 q^\mu q^\nu + q^2 p^\mu p^\nu - (p\cdot q)(q^\mu p^\nu + p^\mu q^\nu)}{m^2 q^2 - (p\cdot q)^2} \quad [49]$$

To determine the as yet unknown functions $a^{\mu\nu}, b^{\mu\nu}$ and c^μ above, we first contract both sides of the decomposition [48] with p^μ and q^μ to get

$$\begin{aligned} 2p_\lambda J^{\mu\nu\lambda} &= 2m^2 a^{\mu\nu} + 2(p\cdot q)b^{\mu\nu} \\ 2q_\lambda J^{\mu\nu\lambda} &= 2(p\cdot q)a^{\mu\nu} + 2q^2 b^{\mu\nu} \end{aligned} \quad [50]$$

Inside the integrand of [47], we then use the trivial identities

$$\begin{aligned} 2k\cdot p &= (k^2 + 2k\cdot p) - k^2 \\ 2q\cdot k &= k^2 + q^2 - (k-q)^2 \end{aligned} \quad [51]$$

to write the left-hand sides of [50] as the sum of rank-2 Feynman integrals which, with the exception of the one multiplied by q^2 from [51], have one less denominator factor. This formally determines the coefficients $a^{\mu\nu}$ and $b^{\mu\nu}$ in terms of a set of rank-2 integrations. The vector function c^μ is then found from the contraction

$$J^{\mu\nu}{}_\nu = p_\nu a^{\mu\nu} + q_\nu b^{\mu\nu} + (D-2)c^\mu \quad [52]$$

This contraction eliminates the k^2 denominator factor in the integrand of [47] and produces a vector-valued integral. Solving the system of algebraic equations [50] and [52] then formally determines the rank-3 Feynman integral [47] in terms of rank-1 and rank-2 Feynman integrals. The rank-2 Feynman integrals thus generated can then be evaluated in the same way by writing a decomposition for them analogous to [48] and solving for them in terms of vector-valued and scalar-valued Feynman integrals. Finally, the rank-1 integrations can be solved for in terms of a set of scalar-valued integrals, most of which have fewer denominator factors in their integrands.

Generally, any one-loop amplitude can be reduced to a set of basic integrals by using the Passarino–Veltman reduction technique. For example, in supersymmetric amplitudes of gluons any tensor Feynman integral can be reduced to a set of scalar integrals, that is, Feynman integrals in a scalar field theory with a massless particle circulating in the loop, with rational coefficients. In the case of $\mathcal{N}=4$ supersymmetric Yang–Mills theory, only scalar box integrals appear.

Reduction to Master Integrals

While the Brown–Feynman and Passarino–Veltman reductions are well suited for dealing with one-loop diagrams, they become rather cumbersome for higher-loop computations. There are other more powerful methods for reducing general tensor integrals into a basis of known integrals called master integrals. Let us illustrate this technique on a scalar example. Any scalar massless two-loop Feynman integral can be brought into the form

$$I(p) = \int \frac{d^D k}{(2\pi)^D} \int \frac{d^D k'}{(2\pi)^D} \prod_{j=1}^{t} \Delta_j^{-l_j} \prod_{i=1}^{q} \Sigma_i^{n_i} \quad [53]$$

where Δ_j are massless scalar propagators depending on the loop momenta k, k' and the external momenta $p_1, \ldots, p_n$, and Σ_i are scalar products of a loop momentum with an external momentum or of the two loop momenta. The topology of the corresponding Feynman diagram is uniquely determined by specifying the set $\Delta_1, \ldots, \Delta_t$ of t distinct

propagators in the graph, while the integral itself is specified by the powers $l_j \geq 1$ of all propagators, by the selection $\Sigma_1, \ldots, \Sigma_q$ of q scalar products and by their powers $n_i \geq 0$.

The integrals in a class of diagrams of the same topology with the same denominator dimension $r = \sum_j l_j$ and same total scalar product number $s = \sum_i n_i$ are related by various identities. One class follows from the fact that the integral over a total derivative with respect to any loop momentum vanishes in dimensional regularization as

$$\int \frac{\mathrm{d}^D k}{(2\pi)^D} \frac{\partial J(k)}{\partial k^\mu} = 0$$

where $J(k)$ is any tensorial combination of propagators, scalar products and loop momenta. The resulting relations are called integration-by-parts identities and for two-loop integrals can be cast into the form

$$\int \frac{\mathrm{d}^D k}{(2\pi)^D} \int \frac{\mathrm{d}^D k'}{(2\pi)^D} v^\mu \frac{\partial f(k,k',p)}{\partial k^\mu} = 0$$

$$= \int \frac{\mathrm{d}^D k}{(2\pi)^D} \int \frac{\mathrm{d}^D k'}{(2\pi)^D} v^\mu \frac{\partial f(k,k',p)}{\partial k'^\mu} \qquad [54]$$

where $f(k,k',p)$ is a scalar function containing propagators and scalar products, and v^μ is any internal or external momentum. For a graph with ℓ loops and n independent external momenta, this results in a total of $\ell(n+\ell)$ relations.

In addition to these identities, one can also exploit the fact that all Feynman integrals [53] are Lorentz scalars. Under an infinitesimal Lorentz transformation $p^\mu \to p^\mu + \delta p^\mu$, with $\delta p^\mu = p^\nu \delta \epsilon^\mu_\nu$, $\delta \epsilon^\mu_\nu = -\delta \epsilon^\nu_\mu$, one has the invariance condition $I(p+\delta p) = I(p)$, which leads to the linear homogeneous differential equations

$$\sum_{i=1}^{n} \left(p_i^\nu \frac{\partial}{\partial p_{i\mu}} - p_i^\mu \frac{\partial}{\partial p_{i\nu}} \right) I(p) = 0 \qquad [55]$$

This equation can be contracted with all possible antisymmetric combinations of $p_{i\mu} p_{j\nu}$ to yield linearly independent Lorentz invariance identities for (53).

Using these two sets of identities, one can either obtain a reduction of integrals of the type (53) to those corresponding to a small number of simpler diagrams of the same topology and diagrams of simpler topology (fewer denominator factors), or a complete reduction to diagrams with simpler topology. The remaining integrals of the topology under consideration are called irreducible master integrals. These momentum integrals cannot be further reduced and have to be computed by different techniques. For instance, one can apply a Mellin–Barnes transformation of all propagators given by

$$\frac{1}{(k^2+a)^l} = \frac{1}{(l-1)!} \int_{-i\infty}^{i\infty} \frac{\mathrm{d}z}{2\pi i} \frac{a^z}{(k^2)^{l+z}} \Gamma(l+z)\Gamma(-z) \qquad [56]$$

where the contour of integration is chosen to lie to the right of the poles of the Euler function $\Gamma(l+z)$ and to the left of the poles of $\Gamma(-z)$ in the complex z-plane. Alternatively, one may apply the negative-dimension method in which D is regarded as a negative integer in intermediate calculations and the problem of loop integration is replaced with that of handling infinite series. When combined with the above methods, it may be used to derive powerful recursion relations among scattering amplitudes. Both of these techniques rely on an explicit integration over the loop momenta of the graph, their differences occurring mainly in the representations used for the propagators.

The procedure outlined above can also be used to reduce a tensor Feynman integral to scalar integrals, as in the Brown–Feynman and Passarino–Veltman reductions. The tensor integrals are expressed as linear combinations of scalar integrals of either higher dimension or with propagators raised to higher powers. The projection onto a tensor basis takes the form [53] and can thus be reduced to master integrals.

String Theory Methods

The realizations of field theories as the low-energy limits of string theory provides a number of powerful tools for the calculation of multiloop amplitudes. They may be used to provide sets of diagrammatic computational rules, and they also work well for calculations in quantum gravity. In this final part we shall briefly sketch the insights into perturbative quantum field theory that are provided by techniques borrowed from string theory.

String Theory Representation

String theory provides an efficient compact representation of scattering amplitudes. At each loop order there is only a single closed string diagram, which includes within it all Feynman graphs along with the contributions of the infinite tower of massive string excitations. Schematically, at one-loop order, the situation is as shown in **Figure 3**. The terms arising from the heavy string modes are removed by taking the low-energy limit in which all external momenta lie well below the energy scale set by the string tension. This limit picks out the regions of integration in the string diagram corresponding to particle-like graphs, but with different diagrammatic rules.



Figure 3 String theory representation at one-loop order.

Given these rules, one may formulate a purely field-theoretic framework which reproduces them. In the case of QCD, a key ingredient is the use of a special gauge originally derived from the low-energy limit of tree-level string amplitudes. This is known as the Gervais–Neveu gauge and it is defined by the gauge-fixing Lagrangian density

$$\mathcal{L}_{\rm GN} = -\frac{1}{2}\mathrm{Tr}\left(\partial_\mu A^\mu - \frac{ie}{\sqrt{2}}A_\mu A^\mu\right)^2 \qquad [57]$$

This gauge choice simplifies the color factors that arise in scattering amplitudes. The string theory origin of gauge theory amplitudes is then most closely mimicked by combining this gauge with the background field gauge, in which one decomposes the gauge field into a classical background field and a fluctuating quantum field as $A_\mu = A_\mu^{\rm cl} + A_\mu^{\rm qu}$, and imposes the gauge-fixing condition $D_\mu^{\rm cl} A^{{\rm qu}\mu} = 0$, where $D_\mu^{\rm cl}$ is the background field covariant derivative evaluated in the adjoint representation of the gauge group. This hybrid gauge is well suited for computing the effective action, with the quantum part describing gluons propagating around loops and the classical part describing gluons emerging from the loops. The leading loop momentum behavior of one-particle irreducible graphs with gluons in the loops is very similar to that of graphs with scalar fields in the loops.

Supersymmetric Decomposition

String theory also suggests an intimate relationship with supersymmetry. For example, at tree level, QCD is effectively supersymmetric because a multigluon tree amplitude contains no fermion loops, and so the fermions may be taken to lie in the adjoint representation of the gauge group. Thus, pure gluon tree amplitudes in QCD are identical to those in supersymmetric Yang–Mills theory. They are connected by supersymmetric Ward identities to amplitudes with fermions (gluinos) which drastically simplify computations. In supersymmetric gauge theory, these identities hold to all orders of perturbation theory.

At one-loop order and beyond, QCD is not supersymmetric. However, one can still perform a supersymmetric decomposition of a QCD amplitude for which the supersymmetric components of the amplitude obey the supersymmetric Ward identities. Consider, for example, a one-loop multigluon scattering amplitude. The contribution from a fermion propagating in the loop can be decomposed into the contribution of a complex scalar field in the loop plus a contribution from an $\mathcal{N}=1$ chiral supermultiplet consisting of a complex scalar field and a Weyl fermion. The contribution from a gluon circulating in the loop can be decomposed into contributions of a complex scalar field, an $\mathcal{N}=1$ chiral supermultiplet, and an $\mathcal{N}=4$ vector supermultiplet comprising three complex scalar fields, four Weyl fermions and one gluon all in the adjoint representation of the gauge group. This decomposition assumes the use of a supersymmetry-preserving regularization.

The supersymmetric components have important cancellations in their leading loop momentum behavior. For instance, the leading large loop momentum power in an n-point 1PI graph is reduced from $|k|^n$ down to $|k|^{n-2}$ in the $\mathcal{N}=1$ amplitude. Such a reduction can be extended to any amplitude in supersymmetric gauge theory and is related to the improved ultraviolet behavior of supersymmetric amplitudes. For the $\mathcal{N}=4$ amplitude, further cancellations reduce the leading power behavior all the way down to $|k|^{n-4}$. In dimensional regularization, $\mathcal{N}=4$ supersymmetric loop amplitudes have a very simple analytic structure owing to their origins as the low-energy limits of superstring scattering amplitudes. The supersymmetric Ward identities in this way can be used to provide identities among the nonsupersymmetric contributions. For example, in $\mathcal{N}=1$ supersymmetric Yang–Mills theory one can deduce that fermion and gluon loop contributions are equal and opposite for multigluon amplitudes with maximal helicity violation.

Scattering Amplitudes in Twistor Space

The scattering amplitude in QCD with n incoming gluons of the same helicity vanishes, as does the amplitude with $n-1$ incoming gluons of one helicity and one gluon of the opposite helicity for $n \geq 3$. The first nonvanishing amplitudes are the maximal helicity violating (MHV) amplitudes involving $n-2$ gluons of one helicity and two gluons of the opposite helicity. Stripped of the momentum conservation delta-function and the group theory factor, the tree-level amplitude for a pair of gluons of negative helicity is given by

$$\mathcal{A}(k) = e^{n-2}\langle k_r^- | k_s^+ \rangle \prod_{i=1}^{n} \langle k_i^- | k_{i+1}^+ \rangle^{-1} \qquad [58]$$

This amplitude depends only on the holomorphic (negative chirality) Weyl spinors. The full MHV amplitude (with the momentum conservation delta-function) is invariant under the conformal group $\mathrm{SO}(4,2) \cong \mathrm{SU}(2,2)$ of four-dimensional

Minkowski space. After a Fourier transformation of the positive-chirality components, the complexification $SL(4, \mathbb{C})$ has an obvious four-dimensional representation acting on the positive- and negative-chirality spinor products. This representation space is isomorphic to $\mathbb{C}^4$ and is called twistor space. Its elements are called twistors.

Wave functions and amplitudes have a known behavior under the $\mathbb{C}^\times$-action which rescales twistors, giving the projective twistor space $\mathbb{CP}^3$ or $\mathbb{RP}^3$ according to whether the twistors are complex valued or real valued. The Fourier transformation to twistor space yields (due to momentum conservation) the localization of an MHV amplitude to a genus-0 holomorphic curve $\mathbb{CP}^1$ of degree 1 in $\mathbb{CP}^3$ (or to a real line $\mathbb{RP}^1 \subset \mathbb{RP}^3$). It is conjectured that, generally, an ℓ-loop amplitude with p gluons of positive helicity and q gluons of negative helicity is supported on a holomorphic curve in twistor space of degree $q + \ell - 1$ and genus $\leq \ell$. The natural interpretation of this curve is as the world sheet of a string. The perturbative gauge theory may then be described in terms of amplitudes arising from the couplings of gluons to a string. This twistor string theory is a topological string theory which gives the appropriate framework for understanding the twistor properties of scattering amplitudes. This framework has been used to analyze MHV tree diagrams and one-loop $\mathcal{N} = 4$ supersymmetric amplitudes of gluons.

See also: Constructive Quantum Field Theory; Dispersion Relations; Effective Field Theories; Gauge Theories from Strings; Hopf Algebra Structure of Renormalizable Quantum Field Theory; Perturbative Renormalization Theory and BRST; Quantum Chromodynamics; Renormalization: General Theory; Scattering, Asymptotic Completeness and Bound States; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Stationary Phase Approximation; Supersymmetric Particle Models.

Further Reading

Bern Z, Dixon L, and Kosower DA (1996) Progress in one-loop QCD computations. *Annual Review of Nuclear and Particle Science* 46: 109–148.

Brown LM and Feynman RP (1952) Radiative corrections to Compton scattering. *Physical Review* 85: 231–244.

Cachazo F and Svrček P (2005) Lectures on twistor strings and perturbative Yang–Mills theory. Preprint arXiv:hep-th/0504194.

Catani S (1998) The singular behaviour of QCD amplitudes at two-loop order. *Physics Letters B* 427: 161–171.

Chetyrkin KG and Tkachov FV (1981) Integration by parts: the algorithm to calculate beta functions in 4 loops. *Nuclear Physics B* 192: 159–204.

Dixon L (1996) Calculating scattering amplitudes efficiently. In: Soper DE *QCD and Beyond*, pp. 539–584. Singapore: World Scientific.

Fleischer J, Jegerlehner F, and Tarasov OV (2000) Algebraic reduction of one-loop Feynman graph amplitudes. *Nuclear Physics B* 566: 423–440.

Gehrmann T and Remiddi E (2000) Differential equations for two-loop four-point functions. *Nuclear Physics B* 580: 485–518.

Halliday IG and Ricotta RM (1987) Negative dimensional integrals 1: Feynman graphs. *Physics Letters B* 193: 241–246.

Itzykson C and Zuber J-B (1980) *Quantum Field Theory*. New York: McGraw-Hill.

Mangano ML and Parke SJ (1991) Multiparton amplitudes in gauge theories. *Physics Reports* 200: 301–367.

Magnen J and Seneor R (1979) Expansion and summability methods in constructive field theory. *Mathematical Problems in Theoretical Physics, Proc. Lausanne* 1979: 217–223.

Passarino G and Veltman MJG (1979) One-loop corrections for e^+e^- annihilation into $\mu^+\mu^-$ in the Weinberg model. *Nuclear Physics B* 160: 151–207.

Peskin ME and Schroeder DV (1995) *An Introduction to Quantum Field Theory*. Reading, MA: Addison-Wesley.

Pokorski S (1987) *Gauge Field Theories*. Cambridge: Cambridge University Press.

Steinhauser M (2002) Results and techniques of multiloop calculations. *Physics Reports* 364: 247–357.

Steinmann O (1995) Axiomatic approach to perturbative quantum field theory. *Annales de Poincare en Physique Theorique*. 63: 399–409.

Weinberg S (1995) *The Quantum Theory of Fields, Volume 1: Foundations*. Cambridge: Cambridge University Press.

Quantum Fields with Topological Defects

M Blasone and G Vitiello, Università degli Studi di Salerno, Baronissi (SA), Italy
P Jizba, Czech Technical University, Prague, Czech Republic

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The ordered patterns we observe in condensed matter and in high-energy physics are created by the quantum dynamics. Macroscopic systems exhibiting some kind of ordering, such as superconductors, ferromagnets, and crystals, are described by the underlying quantum dynamics. Even the large-scale structures in the universe, as well as the ordering in the biological systems appear to be the manifestation of the microscopic dynamics governing the elementary components of these systems. Thus, we talk of macroscopic quantum systems: these are quantum systems in the sense that, although they behave classically, some of their macroscopic features nevertheless cannot be understood without recourse to quantum theory.

The question then arises how the quantum dynamics generates the observed macroscopic properties. In other words, how it happens that the macroscopic scale characterizing those systems is dynamically generated out of the microscopic scale of the quantum elementary components (Umezawa 1993, Umezawa *et al.* 1982).

Moreover, we also observe a variety of phenomena where quantum particles coexist and interact with extended macroscopic objects which show a classical behavior, for example, vortices in superconductors and superfluids, magnetic domains in ferromagnets, dislocations and other topological defects (grain boundaries, point defects, etc.) in crystals, and so on.

We are thus also faced with the question of the quantum origin of topological defects and their interaction with quanta (Umezawa 1993, Umezawa *et al.* 1982): this is a crucial issue for the understanding of symmetry-breaking phase transitions and structure formation in a wide range of systems from condensed matter to cosmology (Kibble 1976, Zurek 1997, Volovik 2003).

Here, we will review how the generation of ordered structures and extended objects is explained in quantum field theory (QFT). We follow Umezawa (1993) and Umezawa *et al.* (1982) in our presentation. We will consider systems in which spontaneous symmetry breaking (SSB) occurs and show that topological defects originate by inhomogeneous (localized) condensation of quanta. The approach followed here is alternative to the usual one (Rajaraman 1982), in which one starts from the classical soliton solutions and then "quantizes" them, as well as to the QFT method based on dual (disorder) fields (Kleinert 1989).

In the next section we introduce some general features of QFT useful for our discussion and treat some aspects of SSB and the rearrangement of symmetry. Next we discuss the boson transformation theorem and the topological singularities of the boson condensate. We then present, as an example, a model with U(1) gauge invariance in which SSB, rearrangement of symmetry, and topological defects are present (Matsumoto *et al.* 1975a, b). There we show how macroscopic fields and currents are obtained from the microscopic quantum dynamics. The Nielsen–Olesen vortex solution is explicitly obtained as an example. The final section is devoted to conclusions.

Symmetry and Order in QFT: A Dynamical Problem

QFT deals with systems with infinitely many degrees of freedom. The fields used for their description are operator fields whose mathematical significance is fully specified only when the state space where they operate is also assigned. This is the space of the states, or physical phase, of the system under given boundary conditions. A change in the boundary conditions may result in the transition of the system from one phase to another. For example, a change of temperature from above to below the critical temperature may induce the transition from the

Notes:

normal to the superconducting phase in a metal. The identification of the state space where the field operators have to be realized is thus a physically nontrivial problem in QFT. In this respect, the QFT structure is drastically different from the one of quantum mechanics (QM). The reason is the following.

The von Neumann theorem (1955) in QM states that for systems with a finite number of degrees of freedom all the irreducible representations of the canonical commutation relations are unitarily equivalent. Therefore, in QM the physical system can only live in one single physical phase: unitary equivalence means indeed physical equivalence and thus there is no room (no representations) for physically different phases. Such a situation drastically changes in QFT where systems with infinitely many degrees of freedom are treated. In such a case, the von Neumann theorem does not hold and infinitely many unitarily inequivalent representations of the canonical commutation relations do in fact exist (Umezawa 1993, Umezawa *et al.* 1982). It is such richness of QFT that allows the description of different physical phases.

QFT as a Two-Level Theory

In the perturbative approach, any quantum experiment or observation can be schematized as a scattering process where one prepares a set of free (noninteracting) particles (incoming particles or in-fields) which are then made to collide at some later time in some region of space (spacetime region of interaction). The products of the collision are expected to emerge out of the interaction region as free particles (outgoing particles or out-fields). Correspondingly, one has the in-field and the out-field state space. The interaction region is where the dynamics operates: given the in-fields and the in-states, the dynamics determines the out-fields and the out-states.

The incoming particles and the outgoing ones (also called quasiparticles in solid state physics) are well distinguishable and localizable particles only far away from the interaction region, at a time much before ($t=-\infty$) and much after ($t=+\infty$) the interaction time: in- and out-fields are thus said to be asymptotic fields, and for them the interaction forces are assumed not to operate (switched off).

The only regions accessible to observations are those far away (in space and in time) from the interaction region, that is, the asymptotic regions (the in- and out-regions). It is so since, at the quantum level, observations performed in the interaction region or vacuum fluctuations occurring there may drastically interfere with the interacting objects, thus changing their nature. Besides the asymptotic fields, one then also introduces dynamical or Heisenberg fields, that is, the fields in terms of which the dynamics is given. Since the interaction region is precluded from observation, we do not observe Heisenberg fields. Observables are thus solely described in terms of asymptotic fields.

Summing up, QFT is a "two-level" theory: one level is the interaction level where the dynamics is specified by assigning the equations for the Heisenberg fields. The other level is the physical level, the one of the asymptotic fields and of the physical state space directly accessible to observations. The equations for the physical fields are equations for free fields, describing the observed incoming/outgoing particles.

To be specific, let the Heisenberg operator fields be generically denoted by $\psi_{\mathrm{H}}(x)$ and the physical operator fields by $\varphi_{\mathrm{in}}(x)$. For definiteness, we choose to work with the in-fields, although the set of out-fields would work equally well. They are both assumed to satisfy equal-time canonical (anti)-commutation relations.

For brevity, we omit considerations on the renormalization procedure, which are not essential for the conclusions we will reach. The Heisenberg field equations and the free-field equations are written as

$$\Lambda(\partial)\psi_{\mathrm{H}}(x) = \mathcal{J}[\psi_{\mathrm{H}}](x) \qquad [1]$$

$$\Lambda(\partial)\varphi_{\mathrm{in}}(x) = 0 \qquad [2]$$

where $\Lambda(\partial)$ is a differential operator, $x \equiv (t, \boldsymbol{x})$ and $\mathcal{J}$ is some functional of the ψ_{H} fields, describing the interaction.

Equation [1] can be formally recast in the following integral form (Yang–Feldman equation):

$$\psi_{\mathrm{H}}(x) = \varphi_{\mathrm{in}}(x) + \Lambda^{-1}(\partial) * \mathcal{J}[\psi_{\mathrm{H}}](x) \qquad [3]$$

where $*$ denotes convolution. The symbol $\Lambda^{-1}(\partial)$ denotes formally the Green function for $\varphi_{\mathrm{in}}(x)$. The precise form of Green's function is specified by the boundary conditions. Equation [3] can be solved by iteration, thus giving an expression for the Heisenberg fields $\psi_{\mathrm{H}}(x)$ in terms of powers of the $\varphi_{\mathrm{in}}(x)$ fields; this is the Haag expansion in the LSZ formalism (or "dynamical map" in the language of Umezawa 1993 and Umezawa *et al.* 1982), which might be formally written as

$$\psi_{\mathrm{H}}(x) = F[x; \varphi_{\mathrm{in}}] \qquad [4]$$

(A (formal) closed form for the dynamical map is obtained in the closed time path (CTP) formalism (Blasone and Jizba 2002). Then the Haag expansion [4] is directly applicable to both equilibrium and nonequilibrium situations.)

We stress that the equality in the dynamical map [4] is a "weak" equality, which means that it must be understood as an equality among matrix elements computed in the Hilbert space of the physical particles.

We observe that mathematical consistency in the above procedure requires that the set of φ_{in} fields must be an irreducible set; however, it may happen that not all the elements of the set are known from the beginning. For example, there might be composite (bound states) fields or even elementary quanta whose existence is ignored in a first recognition. Then the computation of the matrix elements in physical states will lead to the detection of unexpected poles in the Green's functions, which signal the existence of the ignored quanta. One thus introduces the fields corresponding to these quanta and repeats the computation. This way of proceeding is called the self- consistent method (Umezawa 1993, Umezawa *et al.* 1982). Thus it is not necessary to have a one-to-one correspondence between the sets $\{\psi^i_H\}$ and $\{\varphi^i_{in}\}$, as it happens whenever the set $\{\varphi^i_{in}\}$ includes composite particles.

The Dynamical Rearrangement of Symmetry

As already mentioned, in QFT the Fock space for the physical states is not unique since one may have several physical phases, for example, for a metal the normal phase and the superconducting phase, and so on. Fock spaces describing different phases are unitarily inequivalent spaces and correspondingly we have different expectation values for certain observables and even different irreducible sets of physical quanta. Thus, finding the dynamical map involves singling out the Fock space where the dynamics has to be realized.

Let us now suppose that the Heisenberg field equations are invariant under some group G of transformations of ψ_H:

$$\psi_H(x) \to \psi'_H(x) = g[\psi_H(x)] \qquad [5]$$

with $g \in G$. The symmetry is spontaneously broken when the vacuum state in the Fock space $\mathcal{H}$ is not invariant under the group G but only under one of its subgroups (Umezawa 1993, Umezawa *et al.* 1982).

On the other hand, eqn [4] implies that when ψ_H is transformed as in [5], then

$$\varphi_{in}(x) \to \varphi'_{in}(x) = g'[\varphi_{in}(x)] \qquad [6]$$

with g' belonging to some group of transformations G' and such that

$$g[\psi_H(x)] = F[g'[\varphi_{in}(x)]] \qquad [7]$$

When symmetry is spontaneously broken it is $G' \neq G$, with G' the group contraction of G; when symmetry is not broken then $G' = G$.

Since G is the invariance group of the dynamics, eqn [4] requires that G' is the group under which free fields equations are invariant, that is, also φ'_{in} is a solution of [2]. Since eqn [4] is a weak equality, G' depends on the choice of the Fock space $\mathcal{H}$ among the physically realizable unitarily inequivalent state spaces. Thus, we see that the (same) original invariance of the dynamics may manifest itself in different symmetry groups for the φ_{in} fields according to different choices of the physical state space. Since this process is constrained by the dynamical equations [1], it is called the dynamical rearrangement of symmetry (Umezawa 1993, Umezawa *et al.* 1982).

In conclusion, different ordering patterns appear to be different manifestations of the same basic dynamical invariance. The discovery of the process of the dynamical rearrangement of symmetry leads to a unified understanding of the dynamical generation of many observable ordered patterns. This is the phenomenon of the dynamical generation of order. The contraction of the symmetry group is the mathematical structure controlling the dynamical rearrangement of the symmetry. For a qualitative presentation see Vitiello (2001).

One can now ask which ones are the carriers of the ordering information among the system elementary constituents and how the long-range correlations and the coherence observed in ordered patterns are generated and sustained. The answer is in the fact that SSB implies the appearance of bosons (Goldstone 1961, Goldstone *et al.* 1962, Nambu and Jona-Lasinio 1961), the so-called Nambu–Goldstone (NG) modes or quanta. They manifest as long-range correlations and thus they are responsible of the above-mentioned change of scale, from microscopic to macroscopic. The coherent boson condensation of NG modes turns out to be the mechanism by which order is generated, as we will see in an explicit example in a later section.

The "Boson Transformation" Method

We now discuss the quantum origin of extended objects (defects) and show how they naturally emerge as macroscopic objects (inhomogeneous condensates) from the quantum dynamics. At zero temperature, the classical soliton solutions are then recovered in the Born approximation. This approach is known as the "boson transformation" method (Umezawa 1993, Umezawa *et al.* 1982).

The Boson Transformation Theorem

Let us consider, for simplicity, the case of a dynamical model involving one scalar field ψ_H and one asymptotic field φ_{in} satisfying eqns [1] and [2], respectively.

As already remarked, the dynamical map is valid only in a weak sense, that is, as a relation among matrix elements. This implies that eqn [4] is not unique, since different sets of asymptotic fields and the corresponding Hilbert spaces can be used in its construction. Let us indeed consider a c-number function $f(x)$, satisfying the φ_{in} equations of motion [2]:

$$\Lambda(\partial)f(x) = 0 \qquad [8]$$

The boson transformation theorem (Umezawa 1993, Umezawa *et al.* 1982) states that the field

$$\psi^f_H(x) = F[x; \varphi_{in} + f] \qquad [9]$$

is also a solution of the Heisenberg equation [1]. The corresponding Yang–Feldman equation takes the form

$$\psi^f_H(x) = \varphi_{in}(x) + f(x) + \Lambda^{-1}(\partial) * \mathcal{J}[\psi^f_H](x) \qquad [10]$$

The difference between the two solutions ψ_H and ψ^f_H is only in the boundary conditions. An important point is that the expansion in [9] is obtained from that in [4] by the spacetime-dependent translation

$$\varphi_{in}(x) \to \varphi_{in}(x) + f(x) \qquad [11]$$

The essence of the boson transformation theorem is that the dynamics embodied in eqn [1] contains an internal freedom, represented by the possible choices of the function $f(x)$, satisfying the free-field equation [8].

We also observe that the transformation [11] is a canonical transformation since it leaves invariant the canonical form of commutation relations.

Let $|0\rangle$ denote the vacuum for the free field φ_{in}. The vacuum expectation value of eqn [10] gives

$$\begin{aligned}\phi^f(x) &\equiv \langle 0|\psi^f_H(x)|0\rangle \\ &= f(x) + \left\langle 0\middle|\left[\Lambda^{-1}(\partial) * \mathcal{J}[\psi^f_H](x)\right]\middle|0\right\rangle\end{aligned} \qquad [12]$$

The c-number field $\phi^f(x)$ is the order parameter. We remark that it is fully determined by the quantum dynamics. In the classical or Born approximation, which consists in taking $\langle 0|\mathcal{J}[\psi^f_H]|0\rangle = \mathcal{J}[\phi^f]$, that is, neglecting all the contractions of the physical fields, we define $\phi^f_{cl}(x) \equiv \lim_{\hbar\to 0}\phi^f(x)$. In this limit, we have

$$\Lambda(\partial)\phi^f_{cl}(x) = \mathcal{J}[\phi^f_{cl}](x) \qquad [13]$$

that is, $\phi^f_{cl}(x)$ provides the solution of the classical Euler–Lagrange equation.

Beyond the classical level, in general, the form of this equation changes. The Yang–Feldman equation [10] gives not only the equation for the order parameter, eqn [13], but also, at higher orders in $\hbar$, the dynamics of the physical quanta in the potential generated by the "macroscopic object" $\phi^f(x)$ (Umezawa 1993, Umezawa *et al.* 1982).

One can show (Umezawa 1993, Umezawa *et al.* 1982) that the class of solutions of eqn [8] which lead to topologically nontrivial (i.e., carrying a nonzero topological charge) solutions of eqn [13], are those which have some sort of singularity with respect to Fourier transform. These can be either divergent singularities or topological singularities. The first are associated to a divergence of $f(x)$ for $|x| = \infty$, at least in some direction. Topological singularities are instead present when $f(x)$ is not single-valued, that is, it is path dependent. In both cases, the macroscopic object described by the order parameter, carries a nonzero topological charge.

Topological Singularities and Massless Bosons

An important result is that the boson transformation functions carrying topological singularities are only allowed for massless bosons (Umezawa 1993, Umezawa *et al.* 1982).

Consider a generic boson field χ_{in} satisfying the equation

$$(\partial^2 + m^2)\chi_{in}(x) = 0 \qquad [14]$$

and suppose that the function $f(x)$ for the boson transformation $\chi_{in}(x) \to \chi_{in}(x) + f(x)$ carries a topological singularity. It is then not single-valued and thus path dependent:

$$G^+_{\mu\nu}(x) \equiv [\partial_\mu, \partial_\nu] f(x) \neq 0, \quad \text{for certain } \mu, \nu, x \qquad [15]$$

On the other hand, $\partial_\mu f(x)$, which is related with observables, is single-valued, that is, $[\partial_\rho, \partial_\nu]$ $\partial_\mu f(x) = 0$. Recall that $f(x)$ is solution of the χ_{in} equation:

$$(\partial^2 + m^2)f(x) = 0 \qquad [16]$$

From the definition of $G^+_{\mu\nu}(x)$ and the regularity of $\partial_\mu f(x)$, it follows, by computing $\partial^\mu G^+_{\mu\nu}(x)$, that

$$\partial_\mu f(x) = \frac{1}{\partial^2 + m^2}\partial^\lambda G^+_{\lambda\mu}(x) \qquad [17]$$

This equation and the antisymmetric nature of $G^+_{\mu\nu}(x)$ then lead to $\partial^2 f(x) = 0$, which in turn implies $m = 0$. Thus, we conclude that [15] is only compatible with massless equation for χ_{in}.

The topological charge is defined as

$$N_T = \int_C dl^\mu \partial_\mu f = \int_S dS_\mu \epsilon^{\mu\nu\sigma} \partial_\nu \partial_\sigma f$$
$$= \frac{1}{2} \int_S dS^{\mu\nu} G^+_{\mu\nu} \quad [18]$$

Here C is a contour enclosing the singularity and S a surface with C as boundary. N_T does not depend on the path C provided this does not cross the singularity. The dual tensor $G^{\mu\nu}(x)$ is

$$G^{\mu\nu}(x) \equiv -\tfrac{1}{2} \epsilon^{\mu\nu\lambda\rho} G^+_{\lambda\rho}(x) \quad [19]$$

and satisfies the continuity equation

$$\partial_\mu G^{\mu\nu}(x) = 0$$
$$\Leftrightarrow \quad \partial_\mu G^+_{\lambda\rho}(x) + \partial_\rho G^+_{\mu\lambda}(x) + \partial_\lambda G^+_{\rho\mu}(x) = 0 \quad [20]$$

Equation [20] completely characterizes the topological singularity (Umezawa 1993, Umezawa *et al.* 1982).

An Example: The Anderson–Higgs–Kibble Mechanism and the Vortex Solution

We consider a model of a complex scalar field $\phi(x)$ interacting with a gauge field $A_\mu(x)$ (Anderson 1958, Higgs 1960, Kibble 1967). The lagrangian density $\mathcal{L}[\phi(x), \phi^*(x), A_\mu(x)]$ is invariant under the global and the local U(1) gauge transformations (we do not assume a particular form for the Lagrangian density, so the following results are quite general):

$$\phi(x) \to e^{i\theta} \phi(x), \quad A_\mu(x) \to A_\mu(x) \quad [21]$$

$$\phi(x) \to e^{ie_0 \lambda(x)} \phi(x), \quad A_\mu(x) \to A_\mu(x) + \partial_\mu \lambda(x) \quad [22]$$

respectively, where $\lambda(x) \to 0$ for $|x_0| \to \infty$ and/or $|x| \to \infty$ and e_0 is the coupling constant. We work in the Lorentz gauge $\partial_\mu A^\mu(x) = 0$. The generating functional, including the gauge constraint, is (Matsumoto *et al.* 1975a, b)

$$\mathcal{Z}[J, K] = \frac{1}{\mathcal{N}} \int [dA_\mu][d\phi][d\phi^*][dB]$$
$$\times \exp\left[i\, \mathcal{S}[A_\mu, B, \phi]\right] \quad [23]$$

$$\mathcal{S} = \int d^4x \Big[\mathcal{L}(x) + B(x) \partial^\mu A_\mu(x)$$
$$+ K^*(x)\phi(x) + K(x)\phi^*(x)$$
$$+ J^\mu(x) A_\mu(x) + i\epsilon |\phi(x) - v|^2 \Big]$$

$$\mathcal{N} = \int [dA_\mu][d\phi][d\phi^*][dB]$$
$$\times \exp\left[i \int d^4x \Big(\mathcal{L}(x) + i\epsilon |\phi(x) - v|^2 \Big) \right]$$

$B(x)$ is an auxiliary field which implements the gauge-fixing condition (Matsumoto *et al.* 1975a, b). Notice the ϵ-term where v is a complex number; its rôle is to specify the condition of symmetry breaking under which we want to compute the functional integral and it may be given the physical meaning of a small external field triggering the symmetry breaking (Matsumoto *et al.* 1975a, b). The limit $\epsilon \to 0$ must be made at the end of the computations. We will use the notation

$$\langle F[\phi] \rangle_{\epsilon, J, K} \equiv \frac{1}{\mathcal{N}} \int [dA_\mu][d\phi][d\phi^*][dB] F[\phi]$$
$$\times \exp\left[i\, \mathcal{S}[A_\mu, B, \phi]\right] \quad [24]$$

with $\langle F[\phi] \rangle_\epsilon \equiv \langle F[\phi] \rangle_{\epsilon, J=K=0}$ and $\langle F[\phi] \rangle \equiv \lim_{\epsilon \to 0} \langle F[\phi] \rangle_\epsilon$.

The fields ϕ, A_μ, and B appearing in the generating functional are c-number fields. In the following, the Heisenberg operator fields corresponding to them will be denoted by ϕ_H, $A_{H\mu}$, and B_H, respectively. Thus, the spontaneous symmetry breaking condition is expressed by $\langle 0 | \phi_H(x) | 0 \rangle \equiv \tilde{v} \neq 0$, with $\tilde{v}$ constant.

Since in the functional integral formalism the functional average of a given c-number field gives the vacuum expectation value of the corresponding operator field, for example, $\langle F[\phi] \rangle \equiv \langle 0 | F[\phi_H] | 0 \rangle$, we have $\lim_{\epsilon \to 0} \langle \phi(x) \rangle_\epsilon \equiv \langle 0 | \phi_H(x) | 0 \rangle = \tilde{v}$.

Let us introduce the following decompositions:

$$\phi(x) = \frac{1}{\sqrt{2}} [\psi(x) + i\chi(x)]$$
$$K(x) = \frac{1}{\sqrt{2}} [K_1(x) + iK_2(x)]$$
$$\rho(x) \equiv \psi(x) - \langle \psi(x) \rangle_\epsilon$$

Note that $\langle \chi(x) \rangle_\epsilon = 0$ because of the invariance under $\chi \to -\chi$.

The Goldstone Theorem

Since the functional integral [23] is invariant under the global transformation [21], we have that $\partial \mathcal{Z}[J, K] / \partial \theta = 0$ and subsequent derivatives with respect to K_1 and K_2 lead to

$$\langle \psi(x) \rangle_\epsilon = \sqrt{2} \epsilon v \int d^4y \langle \chi(x) \chi(y) \rangle_\epsilon$$
$$= \sqrt{2} \epsilon v \Delta_\chi(\epsilon, 0) \quad [25]$$

In momentum space the propagator for the field χ has the general form

$$\Delta_\chi(0, p) = \lim_{\epsilon \to 0} \left[\frac{Z_\chi}{p^2 - m_\chi^2 + i\epsilon a_\chi} \right.$$
$$\left. + (\text{continuum contributions}) \right] \quad [26]$$

Here Z_χ and a_χ are renormalization constants. The integration in eqn [25] picks up the pole contribution at $p^2=0$, and leads to

$$\tilde{v}=\sqrt{2}\frac{Z_\chi}{a_\chi}v \Leftrightarrow m_\chi=0, \quad \tilde{v}=0 \Leftrightarrow m_\chi\neq 0 \qquad [27]$$

The Goldstone theorem (Goldstone 1961, Goldstone *et al.* 1962) is thus proved: if the symmetry is spontaneously broken ($\tilde{v}\neq 0$), a massless mode must exist, whose field is $\chi(x)$, that is, the NG boson mode. Since it is massless, it manifests as a long-range correlation mode. (Notice that in the present case of a complex scalar field model, the NG mode is an elementary field. In other models, it may appear as a bound state, for example, the magnon in (anti)ferromagnets.) Note that

$$\frac{\partial}{\partial v}\langle\psi(x)\rangle_\epsilon=\sqrt{2}\epsilon\int \mathrm{d}^4y\langle\rho(x)\rho(y)\rangle_\epsilon \qquad [28]$$

and because $m_\rho\neq 0$, the right-hand side of this equation vanishes in the limit $\epsilon\to 0$; therefore, $\tilde{v}$ is independent of $|v|$, although the phase of $|v|$ determines the one of $\tilde{v}$ (from eqn [25]): as in ferromagnets, once an external magnetic field is switched on, the system is magnetized independently of the strength of the external field.

The Dynamical Map and the Field Equations

Observing that the change of variables [21] (and/or [22]) does not affect the generating functional, we may obtain the Ward–Takahashi identities. Also, using $B(x)\to B(x)+\lambda(x)$ in [23] gives $\langle\partial^\mu A_\mu(x)\rangle_{\epsilon,J,K}=0$. One then finds the following two-point function pole structures (Matsumoto *et al.* 1975a, b):

$$\langle B(x)\chi(y)\rangle=\lim_{\epsilon\to 0}\left\{\frac{-\mathrm{i}}{(2\pi)^4}\int \mathrm{d}^4p\,\mathrm{e}^{-\mathrm{i}p(x-y)}\frac{e_0\tilde{v}}{p^2+\mathrm{i}\epsilon a_\chi}\right\} \qquad [29]$$

$$\langle B(x)A^\mu(y)\rangle=\partial_x^\mu\frac{\mathrm{i}}{(2\pi)^4}\int \mathrm{d}^4p\,\mathrm{e}^{-\mathrm{i}p(x-y)}\frac{1}{p^2} \qquad [30]$$

$$\langle B(x)B(y)\rangle=\lim_{\epsilon\to 0}\left\{\frac{-\mathrm{i}}{(2\pi)^4}\int \mathrm{d}^4p\,\mathrm{e}^{-\mathrm{i}p(x-y)}\frac{(e_0\tilde{v})^2}{Z_\chi}\times\left[\frac{1}{p^2+\mathrm{i}\epsilon a_\chi}-\frac{1}{p^2}\right]\right\} \qquad [31]$$

The absence of branch-cut singularities in propagators [29]–[31] suggests that $B(x)$ obeys a free-field equation. In addition, eqn [31] indicates that the model contains a massless negative-norm state (ghost) besides the NG massless mode χ. Moreover, it can be shown (Matsumoto *et al.* 1975a, b) that a massive vector field U^μ_{in} also exists in the theory. Note that because of the invariance $(\chi, A_\mu, B)\to(-\chi, -A_\mu, -B)$, all the other two-point functions must vanish.

The dynamical maps expressing the Heisenberg operator fields in terms of the asymptotic operator fields are found to be (Matsumoto *et al.* 1975a, b)

$$\phi_{\mathrm{H}}(x)=:\exp\left\{\mathrm{i}\frac{Z_\chi^{1/2}}{\tilde{v}}\chi_{\mathrm{in}}(x)\right\}\left[\tilde{v}+Z_\rho^{1/2}\rho_{\mathrm{in}}(x)+\mathcal{F}[\rho_{\mathrm{in}}, U^\mu_{\mathrm{in}}, \partial(\chi_{\mathrm{in}}-b_{\mathrm{in}})]\right]: \qquad [32]$$

$$A^\mu_{\mathrm{H}}(x)=Z_3^{1/2}U^\mu_{\mathrm{in}}(x)+\frac{Z_\chi^{1/2}}{e_0\tilde{v}}\partial^\mu b_{\mathrm{in}}(x)+:\mathcal{F}^\mu[\rho_{\mathrm{in}}, U^\mu_{\mathrm{in}}, \partial(\chi_{\mathrm{in}}-b_{\mathrm{in}})]: \qquad [33]$$

$$B_{\mathrm{H}}(x)=\frac{e_0\tilde{v}}{Z_\chi^{1/2}}[b_{\mathrm{in}}(x)-\chi_{\mathrm{in}}(x)]+c \qquad [34]$$

where $:\ldots:$ denotes the normal ordering and the functionals $\mathcal{F}$ and $\mathcal{F}^\mu$ are to be determined within a particular model. In eqns [32]–[34], χ_{in} denotes the NG mode, b_{in} the ghost mode, U^μ_{in} the massive vector field, and ρ_{in} the massive matter field. In eqn [34] c is a c-number constant, whose value is irrelevant since only derivatives of B appear in the field equations (see below). Z_3 represents the wave function renormalization for U^μ_{in}. The corresponding field equations are

$$\partial^2\chi_{\mathrm{in}}(x)=0, \qquad \partial^2 b_{\mathrm{in}}(x)=0, \qquad (\partial^2+m_\rho^2)\rho_{\mathrm{in}}(x)=0 \qquad [35]$$

$$(\partial^2+m_V^2)U^\mu_{\mathrm{in}}(x)=0, \quad \partial_\mu U^\mu_{\mathrm{in}}(x)=0 \qquad [36]$$

with $m_V{}^2=(Z_3/Z_\chi)(e_0\tilde{v})^2$. The field equations for B_{H} and $A_{\mathrm{H}\mu}$ read (Matsumoto *et al.* 1975a, b)

$$\partial^2 B_{\mathrm{H}}(x)=0, \quad -\partial^2 A_{\mathrm{H}\mu}(x)=j_{\mathrm{H}\mu}(x)-\partial_\mu B_{\mathrm{H}}(x) \qquad [37]$$

with $j_{\mathrm{H}\mu}(x)=\delta\mathcal{L}(x)/\delta A^\mu_{\mathrm{H}}(x)$. One may then require that the current $j_{\mathrm{H}\mu}$ is the only source of the gauge field $A_{\mathrm{H}\mu}$ in any observable process. This amounts to impose the condition: ${}_p\langle b|\partial_\mu B_{\mathrm{H}}(x)|a\rangle_p=0$, that is,

$$(-\partial^2)_p\langle b|A^0_{\mathrm{H}\mu}(x)|a\rangle_p={}_p\langle b|j_{\mathrm{H}\mu}(x)|a\rangle_p \qquad [38]$$

where $|a\rangle_p$ and $|b\rangle_p$ denote two generic physical states and $A^{0\mu}_{\mathrm{H}}(x)\equiv A^\mu_{\mathrm{H}}(x)-e_0\tilde{v}:\partial^\mu b_{\mathrm{in}}(x):$. Equations [38] are the classical Maxwell equations. The condition ${}_p\langle b|\partial_\mu B_{\mathrm{H}}(x)|a\rangle_p=0$ leads to the Gupta–Bleuler–like condition

$$[\chi^{(-)}_{\mathrm{in}}(x)-b^{(-)}_{\mathrm{in}}(x)]|a\rangle_p=0 \qquad [39]$$

where $\chi^{(-)}_{\mathrm{in}}$ and $b^{(-)}_{\mathrm{in}}$ are the positive-frequency parts of the corresponding fields. Thus, we see that χ_{in} and b_{in} cannot participate in any observable reaction.

This is confirmed by the fact that they are present in the S-matrix in the combination $(\chi_{in} - b_{in})$ (Matsumoto *et al.* 1975a, b). It is to be remarked, however, that the NG boson does not disappear from the theory: we shall see below that there are situations in which the NG fields do have observable effects.

The Dynamical Rearrangement of Symmetry and the Classical Fields and Currents

From eqns [32]–[33] we see that the local gauge transformations of the Heisenberg fields

$$\begin{aligned} &\phi_H(x) \to e^{ie_0\lambda(x)}\phi_H(x) \\ &A^\mu_H(x) \to A^\mu_H(x) + \partial^\mu\lambda(x), \quad B_H(x) \to B_H(x) \end{aligned} \tag{40}$$

with $\partial^2\lambda(x) = 0$, are induced by the in-field transformations

$$\begin{aligned} &\chi_{in}(x) \to \chi_{in}(x) + \frac{e_0\tilde{v}}{Z_\chi^{1/2}}\lambda(x) \\ &b_{in}(x) \to b_{in}(x) + \frac{e_0\tilde{v}}{Z_\chi^{1/2}}\lambda(x) \\ &\rho_{in}(x) \to \rho_{in}(x), \quad U^\mu_{in}(x) \to U^\mu_{in}(x) \end{aligned} \tag{41}$$

On the other hand, the global phase transformation $\phi_H(x) \to e^{i\theta}\phi_H(x)$ is induced by

$$\begin{aligned} &\chi_{in}(x) \to \chi_{in}(x) + \frac{\tilde{v}}{Z_\chi^{1/2}}\theta f(x), \quad b_{in}(x) \to b_{in}(x) \\ &\rho_{in}(x) \to \rho_{in}(x), \quad U^\mu_{in}(x) \to U^\mu_{in}(x) \end{aligned} \tag{42}$$

with $\partial^2 f(x) = 0$ and the limit $f(x) \to 1$ to be performed at the end of computations. Note that under the above transformations, the in-field equations and the S-matrix are invariant and that B_H is changed by an irrelevant c-number (in the limit $f \to 1$).

Consider now the boson transformation $\chi_{in}(x) \to \chi_{in}(x) + \alpha(x)$: in local gauge theories the boson transformation must be compatible with the Heisenberg field equations but also with the physical state condition [39]. Under the boson transformation with $\alpha(x) = \tilde{v}Z_\chi^{-1/2}\theta f(x)$ and $\partial^2 f(x) = 0$, B_H changes as

$$B_H(x) \to B_H(x) - \frac{e_0\tilde{v}^2}{Z_\chi}f(x) \tag{43}$$

eqn [38] is thus violated when the Gupta–Bleuler-like condition is imposed. In order to restore it, the shift in B_H must be compensated by means of the following transformation on U^μ_{in}:

$$U^\mu_{in}(x) \to U^\mu_{in}(x) + Z_3^{-1/2}a^\mu(x), \quad \partial_\mu a^\mu(x) = 0 \tag{44}$$

with a convenient c-number function $a^\mu(x)$. The dynamical maps of the various Heisenberg operators are not affected by [44] since they contain U^μ_{in} and B_H in a combination such that the changes of B_H and of U^μ_{in} compensate each other provided

$$(\partial^2 + m_V^2)a_\mu(x) = \frac{m_V^2}{e_0}\partial_\mu f(x) \tag{45}$$

Equation [45] thus obtained is the Maxwell equation for the massive potential vector a_μ (Matsumoto *et al.* 1975a, b). The classical ground state current j^μ turns out to be

$$j^\mu(x) \equiv \langle 0|j^\mu_H(x)|0\rangle = m_V^2\left[a^\mu(x) - \frac{1}{e_0}\partial^\mu f(x)\right] \tag{46}$$

The term $m_V^2 a^\mu(x)$ is the Meissner current, while $(m_V^2/e_0)\partial^\mu f(x)$ is the boson current. The key point here is that both the macroscopic field and current are given in terms of the boson condensation function $f(x)$.

Two remarks are in order: first, note that the terms proportional to $\partial^\mu f(x)$ are related to observable effects, for example, the boson current which acts as the source of the classical field. Second, note that the macroscopic ground state effects do not occur for regular $f(x)(G^+_{\mu\nu}(x) = 0)$. In fact, from [45] we obtain $a_\mu(x) = (1/e_0)\partial_\mu f(x)$ for regular $f(x)$ which implies zero classical current $(j_\mu = 0)$ and zero classical field $(F_{\mu\nu} = \partial_\mu a_\nu - \partial_\nu a_\mu)$, since the Meissner and the boson current cancel each other.

In conclusion, the vacuum current appears only when $f(x)$ has topological singularities and these can be created only by condensation of massless bosons, that is, when SSB occurs. This explains why topological defects appear in the process of phase transitions, where NG modes are present and gradients in their condensate densities are nonzero (Kibble 1976, Zurek 1997).

On the other hand, the appearance of spacetime order parameter is no guarantee that persistent ground state currents (and fields) will exist: if $f(x)$ is a regular function, the spacetime dependence of $\tilde{v}$ can be gauged away by an appropriate gauge transformation.

Since, as already mentioned, the boson transformation with regular $f(x)$ does not affect observable quantities, the S-matrix is actually given by

$$S = :S\left[\rho_{in}, U^\mu_{in} - \frac{1}{m_V}\partial(\chi_{in} - b_{in})\right]: \tag{47}$$

This is indeed independent of the boson transformation with regular $f(x)$:

$$\begin{aligned} S \to S' = :S\Bigg[&\rho_{in}, U^\mu_{in} - \frac{1}{m_V}\partial(\chi_{in} - b_{in}) \\ &+ Z_3^{-1/2}\left(a^\mu - \frac{1}{e_0}\partial^\mu f\right)\Bigg]: \end{aligned} \tag{48}$$

since $a_\mu(x) = (1/e_0)\partial_\mu f(x)$ for regular $f(x)$. However, $S' \neq S$ for singular $f(x)$: S' includes the interaction of the quanta U^μ_{in} and ϕ_{in} with the classically behaving macroscopic defects (Umezawa 1993, Umezawa *et al.* 1982).

The Vortex Solution

Below we consider the example of the Nielsen–Olesen vortex string solution. We show which one is the boson function $f(x)$ controlling the nonhomogeneous NG boson condensation in terms of which the string solution is described. For brevity, we only report the results of the computations. The detailed derivation as well as the discussion of further examples can be found in (Umezawa 1993, Umezawa *et al.* 1982).

In the present U(1) problem, the electromagnetic tensor and the vacuum current are (Umezawa 1993, Umezawa *et al.* 1982, Matsumoto *et al.* 1975a, b)

$$\begin{aligned} F_{\mu\nu}(x) &= \partial_\mu a_\nu(x) - \partial_\nu a_\mu(x) \\ &= 2\pi \frac{m_V^2}{e_0} \int \mathrm{d}^4x' \Delta_c(x-x') G^+_{\mu\nu}(x') \end{aligned} \quad [49]$$

$$j_\mu(x) = -2\pi \frac{m_V^2}{e_0} \int \mathrm{d}^4x' \Delta_c(x-x') \partial^\nu_{x'} G^+_{\nu\mu}(x') \quad [50]$$

respectively, and satisfy $\partial^\mu F_{\mu\nu}(x) = -j_\nu(x)$. In these equations,

$$\Delta_c(x-x') = \frac{1}{(2\pi)^4} \int \mathrm{d}^4p\, \mathrm{e}^{-\mathrm{i}p(x-x')} \frac{1}{p^2 - m_V^2 + \mathrm{i}\epsilon} \quad [51]$$

The line singularity for the vortex (or string) solution can be parametrized by a single line parameter σ and by the time parameter τ. A static vortex solution is obtained by setting $y_0(\tau,\sigma) = \tau$ and $\boldsymbol{y}(\tau,\sigma) = \boldsymbol{y}(\sigma)$, with y denoting the line coordinate. $G^+_{\mu\nu}(x)$ is nonzero only on the line at y (we can consider more lines but let us limit to only one line, for simplicity). Thus, we have

$$\begin{aligned} &G_{0i}(x) = \int \mathrm{d}\sigma \frac{\mathrm{d}y_i(\sigma)}{\mathrm{d}\sigma} \delta^3[\boldsymbol{x} - \boldsymbol{y}(\sigma)]\; G_{ij}(x) = 0 \\ &G^+_{ij}(x) = -\epsilon_{ijk} G_{0k}(x), \quad G^+_{0i}(x) = 0 \end{aligned} \quad [52]$$

Equation [49] shows that these vortices are purely magnetic. We obtain

$$\begin{aligned} \partial_0 f(x) &= 0 \\ \partial_i f(x) &= \frac{1}{(2\pi)^2} \int \mathrm{d}\sigma \epsilon_{ijk} \frac{\mathrm{d}y_k(\sigma)}{\mathrm{d}\sigma} \partial^x_j \\ &\quad \times \int \mathrm{d}^3p \frac{\mathrm{e}^{\mathrm{i}\boldsymbol{p}\cdot(\boldsymbol{x}-\boldsymbol{y}(\sigma))}}{p^2} \end{aligned} \quad [53]$$

that is, by using the identity $(2\pi)^{-2} \int \mathrm{d}^3p (\mathrm{e}^{\mathrm{i}p\cdot x}/p^2) = 1/2|\boldsymbol{x}|$,

$$\nabla f(x) = -\frac{1}{2} \int \mathrm{d}\sigma \frac{\mathrm{d}y_k(\sigma)}{\mathrm{d}\sigma} \wedge \nabla_x \frac{1}{|\boldsymbol{x} - \boldsymbol{y}(\sigma)|} \quad [54]$$

Note that $\nabla^2 f(x) = 0$ is satisfied.

A straight infinitely long vortex is specified by $y_i(\sigma) = \sigma\delta_{i3}$ with $-\infty < \sigma < \infty$. The only nonvanishing component of $G^{\mu\nu}(x)$ are $G^{03}(x) = G^+_{12}(x) = \delta(x_1)\delta(x_2)$. Equation [54] gives (Umezawa 1993, Umezawa *et al.* 1982, Matsumoto 1975a, b)

$$\begin{aligned} \frac{\partial}{\partial x_1} f(x) &= \frac{1}{2} \int \mathrm{d}\sigma \frac{\partial}{\partial x_2} [x_1^2 + x_2^2 + (x_3 - \sigma)^2]^{-1/2} \\ &= -\frac{x_2}{x_1^2 + x_2^2} \\ \frac{\partial}{\partial x_2} f(x) &= \frac{x_1}{x_1^2 + x_2^2}, \quad \frac{\partial}{\partial x_3} f(x) = 0 \end{aligned} \quad [55]$$

and then

$$f(x) = \tan^{-1}\left(\frac{x_2}{x_1}\right) = \theta(x) \quad [56]$$

We have thus determined the boson transformation function corresponding to a particular vortex solution. The vector potential is

$$\begin{aligned} a_1(x) &= -\frac{m_V^2}{2e_0} \int \mathrm{d}^4x' \Delta_c(x-x') \frac{x'_2}{x'^2_1 + x'^2_2} \\ a_2(x) &= \frac{m_V^2}{2e_0} \int \mathrm{d}^4x' \Delta_c(x-x') \frac{x'_1}{x'^2_1 + x'^2_2} \\ a_3(x) &= a_0(x) = 0 \end{aligned} \quad [57]$$

and the only nonvanishing component of $F_{\mu\nu}$:

$$\begin{aligned} F_{12}(x) &= -2\pi \frac{m_V^2}{e_0} \int \mathrm{d}^4x' \Delta_c(x-x') \delta(x'_1)\delta(x'_2) \\ &= \frac{m_V^2}{e_0} K_0\left(m_V \sqrt{x_1^2 + x_2^2}\right) \end{aligned} \quad [58]$$

Finally, the vacuum current eqn [50] is given by

$$\begin{aligned} j_1(x) &= -\frac{m_V^3}{e_0} \frac{x_2}{\sqrt{x_1^2 + x_2^2}} K_1\left(m_V \sqrt{x_1^2 + x_2^2}\right) \\ j_2(x) &= \frac{m_V^3}{e_0} \frac{x_1}{\sqrt{x_1^2 + x_2^2}} K_1\left(m_V \sqrt{x_1^2 + x_2^2}\right) \\ j_3(x) &= j_0(x) = 0 \end{aligned} \quad [59]$$

We observe that these results are the same of the Nielsen–Olesen vortex solution. Notice that we did not specify the potential in our model but only the invariance properties. Thus, the invariance properties of the dynamics determine the characteristics of the topological solutions. The vortex solution

manifests the original U(1) symmetry through the cylindrical angle θ which is the parameter of the U(1) representation in the coordinate space.

Conclusions

We have discussed how topological defects arise as inhomogeneous condensates in QFT. Topological defects are shown to have a genuine quantum nature. The approach reviewed here goes under the name of "boson transformation method" and relies on the existence of unitarily inequivalent representations of the field algebra in QFT.

Describing quantum fields with topological defects amounts then to properly choose the physical Fock space for representing the Heisenberg field operators. Once the boundary conditions corresponding to a particular soliton sector are found, the Heisenberg field operators embodied with such conditions contain the full information about the defects, the quanta and their mutual interaction. One can thus calculate Green's functions for particles in the presence of defects. The extension to finite temperature is discussed in Blasone and Jizba (2002) and Manka and Vitiello (1990).

As an example we have discussed a model with U(1) gauge invariance and SSB and we have obtained the Nielsen–Olesen vortex solution in terms of localized condensation of Goldstone bosons. These thus appear to play a physical role, although, in the presence of gauge fields, they do not show up in the physical spectrum as excitation quanta. The function $f(x)$ controlling the condensation of the NG bosons must be singular in order to produce observable effects. Boson transformations with regular $f(x)$ only amount to gauge transformations. For the treatment of topological defects in nonabelian gauge theories, see Manka and Vitiello (1990).

Finally, when there are no NG modes, as in the case of the kink solution or the sine-Gordon solution, the boson transformation function has to carry divergence singularity at spatial infinity (Umezawa 1993, Umezawa *et al.* 1982, Blasone and Jizba 2002). The boson transformation has also been discussed in connection with the Bäklund transformation at a classical level and the confinement of the constituent quanta in the coherent condensation domain.

For further reading on quantum fields with topological defects, see Blasone *et al.* (2006).

Acknowledgments

The authors thank MIUR, INFN, INFM, and the ESF network COSLAB for partial financial support.

See also: Abelian Higgs Vortices; Algebraic Approach to Quantum Field Theory; Quantum Field Theory: A Brief Introduction; Quantum Field Theory in Curved Spacetime; Symmetries in Quantum Field Theory: Algebraic Aspects; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Topological Defects and their Homotopy Classification.

Further Reading

Anderson PW (1958) Coherent excited states in the theory of superconductivity: gauge invariance and the Meissner effect. *Physical Review* 110: 827–835.

Blasone M and Jizba P (2002) Topological defects as inhomogeneous condensates in quantum field theory: kinks in $(1+1)$ dimensional $\lambda\psi^4$ theory. *Annals of Physics* 295: 230–260.

Blasone M, Jizba P, and Vitiello G (2006) *Spontaneous Breakdown of Symmetry and Topological Defects*, London: Imperial College Press. (in preparation).

Goldstone J (1961) Field theories with "superconductor" solutions. *Nuovo Cimento* 19: 154–164.

Goldstone J, Salam A, and Weinberg S (1962) Broken symmetries. *Physical Review* 127: 965–970.

Higgs P (1960) Spontaneous symmetry breakdown without massless bosons. *Physical Review* 145: 1156–1163.

Kibble TWB (1967) Symmetry breaking in non-abelian gauge theories. *Physical Review* 155: 1554–1561.

Kibble TWB (1976) Topology of cosmic domains and strings. *Journal of Physics A* 9: 1387–1398.

Kibble TWD (1980) Some implications of a cosmological phase transition. *Physics Reports* 67: 183–199.

Kleinert H (1989) *Gauge Fields in Condensed Matter*, vols. I & II Singapore: World Scientific.

Manka R and Vitiello G (1990) Topological solitons and temperature effects in gauge field theory. *Annals of Physics* 199: 61–83.

Matsumoto H, Papastamatiou NJ, Umezawa H, and Vitiello G (1975a) Dynamical rearrangement in Anderson–Higgs–Kibble mechanism. *Nuclear Physics B* 97: 61–89.

Matsumoto H, Papastamatiou NJ, and Umezawa H (1975b) The boson transformation and the vortex solutions. *Nuclear Physics B* 97: 90–124.

Nambu Y and Jona-Lasinio G (1961) Dynamical model of elementary particles based on an analogy with superconductivity. I. *Physical Review* 122: 345–358.

Nambu Y and Jona-Lasinio G (1961) Dynamical model of elementary particles based on an analogy with superconductivity. II. *Physical Review* 124: 246–254.

Rajaraman R (1982) *Solitons and Instantons: An Introduction to Solitons and Instantons in Quantum Field Theory*. Amsterdam: North-Holland.

Umezawa H (1993) *Advanced Field Theory: Micro, Macro and Thermal Physics*. New York: American Institute of Physics.

Umezawa H, Matsumoto H, and Tachiki M (1982) *Thermo Field Dynamics and Condensed States*. Amsterdam: North-Holland.

Vitiello G (2001) *My Double Unveiled*. Amsterdam: John Benjamins.

Volovik GE (2003) *The Universe in a Helium Droplet*. Oxford: Clarendon.

von Neumann J (1955) *Mathematical Foundation of Quantum Mechanics*. Princeton: Princeton University Press.

Zurek WH (1997) Cosmological experiments in condensed matter systems. *Physics Reports* 276: 177–221.

Quantum Field Theory in Curved Spacetime

B S Kay, University of York, York, UK

© 2006 B S Kay. Published by Elsevier Ltd.
All rights reserved.

Introduction and Preliminaries

Quantum Field Theory (QFT) in curved spacetime is a hybrid approximate theory in which quantum matter fields are assumed to propagate in a fixed classical background gravitational field. Its basic physical prediction is that strong gravitational fields can polarize the vacuum and, when time dependent, lead to pair creation just as a strong and/or time-dependent electromagnetic field can polarize the vacuum and/or give rise to pair creation of charged particles. One expects it to be a good approximation to full quantum gravity provided the typical frequencies of the gravitational background are very much less than the Planck frequency $(c^5/G\hbar)^{1/2} \sim 10^{43}\,\mathrm{s}^{-1}$) and provided, with a suitable measure for energy, the energy of created particles is very much less than the energy of the background gravitational field or of its matter sources. Undoubtedly, the most important prediction of the theory is the Hawking effect, according to which a, say spherically symmetric, classical black hole of mass M will emit thermal radiation at the Hawking temperature $T=(8\pi M)^{-1}$ (here and from now on, we use Planck units where $G, c, \hbar$ and, k (Boltzmann's constant) are all taken to be 1).

On the mathematical side, the need to formulate the laws and derive the general properties of QFT on nonflat spacetimes forces one to state and prove results in local terms and, as a byproduct, thereby leads to an improved perspective on flat-spacetime QFT too. It is also interesting to formulate QFT on idealized spacetimes with particular global geometrical features. Thus, QFT on spacetimes with bifurcate Killing horizons is intimately related to the Hawking effect; QFT on spacetimes with closed timelike curves is intimately related to the question whether the laws of physics permit the manufacture of a time machine.

As is standard in general relativity, a curved spacetime is modeled mathematically as a (paracompact, Hausdorff) manifold $\mathcal{M}$ equipped with a pseudo-Riemannian metric g of signature $(-,+++)$ (we follow the conventions of the standard text by Misner *et al.* (1973)). We shall also assume, except where otherwise stated, our spacetime to be *globally hyperbolic*, that is, that $\mathcal{M}$ admits a *global time coordinate*, by which we mean a global coordinate t such that each constant-t surface is a smooth Cauchy surface, that is, a smooth spacelike 3-surface cut exactly once by each inextendible causal curve. (Without this default assumption, extra problems arise for QFT which we shall briefly mention in connection with the "time machine" question discussed later.) In view of this definition, globally hyperbolic spacetimes are clearly time-orientable and we shall assume a choice of time-orientation has been made so we can talk about the "future" and "past" directions. Modern formulations of the subject take, as the fundamental mathematical structure modeling the quantum field, a $*$-algebra $\mathcal{A}$ (with identity I) together with a family of local sub $*$-algebras $\mathcal{A}(\mathcal{O})$ labeled by bounded open regions $\mathcal{O}$ of the spacetime $(\mathcal{M}, g)$ and satisfying the *isotony* or *net* condition that $\mathcal{O}_1 \subset \mathcal{O}_2$ implies $\mathcal{A}(\mathcal{O}_1)$ is a subalgebra of $\mathcal{A}(\mathcal{O}_2)$ as well as the condition that whenever two bounded open regions $\mathcal{O}_1$ and $\mathcal{O}_2$ are spacelike separated, then $\mathcal{A}(\mathcal{O}_1)$ and $\mathcal{A}(\mathcal{O}_2)$ commute.

Standard concepts and techniques from algebraic quantum theory are then applicable: In particular, *states* are defined to be positive (this means $\omega(A^*A) \geq 0\ \forall A \in \mathcal{A}$) normalized (this means $\omega(I)=1$) linear functionals on $\mathcal{A}$. One distinguishes between *pure* states and *mixed* states, only the latter being writable as nontrivial convex combinations of other states. To each state, ω, the *GNS construction* associates a representation, ρ_ω, of $\mathcal{A}$ on a Hilbert space $\mathcal{H}_\omega$ together with a cyclic vector $\Omega \in \mathcal{H}_\omega$ such that

$$\omega(A) = \langle \Omega | \rho_\omega(A) \Omega \rangle$$

(and the *GNS triple* $(\rho_\omega, \mathcal{H}, \Omega)$ is unique up to equivalence). There are often technical advantages in formulating things so that the $*$-algebra is a C^*-algebra. Then the GNS representation is as everywhere-defined bounded operators and is irreducible if and only if the state is pure. A useful concept, due to Haag, is the *folium* of a given state ω which may be defined to be the set of all states ω_σ which arise in the form $\mathrm{tr}(\sigma\rho_\omega(\cdot))$, where σ ranges over the density operators (trace-class operators with unit trace) on $\mathcal{H}_\omega$.

Given a state, ω, and an automorphism, α, which preserves the state (i.e., $\omega \circ \alpha = \omega$) then there will be a unitary operator, U, on H_ω which *implements* α in the sense that $\rho_\omega(\alpha(A)) = U^{-1}\rho_\omega(A)U$ and U is chosen uniquely by the condition $U\Omega = \Omega$.

On a *stationary* spacetime, that is, one which admits a one-parameter group of isometries whose integral curves are everywhere timelike, the algebra will inherit a one-parameter group (i.e., satisfying $\alpha(t_1) \circ \alpha(t_2) = \alpha(t_1 + t_2)$) of time-translation

automorphisms, $\alpha(t)$, and, given any stationary state (i.e., one which satisfies $\omega \circ \alpha(t) = \omega \; \forall t \in \boldsymbol{R}$), these will be implemented by a one-parameter group of unitaries, $U(t)$, on its GNS Hilbert space satisfying $U(t)\Omega = \Omega$. If $U(t)$ is strongly continuous so that it takes the form e^{-iHt} and if the Hamiltonian, H, is positive, then ω is said to be a "ground state." Typically one expects *ground states* to exist and often be unique.

Another important class of stationary states for the algebra of a stationary spacetime is the class of *KMS states*, ω^β, at inverse temperature β; these have the physical interpretation of thermal equilibrium states. In the GNS representation of one of these, the automorphisms are also implemented by a strongly continuous unitary group, e^{-iHt}, which preserves Ω but (in place of H positive) there is a complex conjugation, J, on H_ω such that

$$e^{-\beta H/2}\rho_\omega(A)\Omega = J\rho_\omega(A^*)\Omega \qquad [1]$$

for all $A \in \mathcal{A}$. An attractive feature of the subject is that its main qualitative features are already present for linear field theories and, unusually in comparison with other questions in QFT, these are susceptible of a straightforward explicit and rigorous mathematical formulation. In fact, as our principal example, we give, in the next section a construction for the field algebra for the quantized real linear Klein–Gordon equation

$$(\Box_g - m^2 - V)\phi = 0 \qquad [2]$$

of mass m on a globally hyperbolic spacetime $(\mathcal{M}, g)$. Here, $\Box_g$ denotes the Laplace–Beltrami operator $g^{ab}\nabla_a\partial_b$ $(= (|\det(g)|)^{-1/2}\partial_a((|\det(g)|^{1/2}g^{ab}\partial_b))$. We include a scalar external background classical field, V, in addition to the external gravitational field represented by g. In case m is zero, taking V to equal $R/6$, where R denotes the Riemann scalar, makes the equation *conformally invariant*.

The main new feature of QFT in curved spacetime (present already for linear field theories) is that, in a general (neither flat nor stationary) spacetime there will not be any single preferred state but rather a family of preferred states, members of which are best regarded as on an equal footing with one another. It is this feature which makes the above algebraic framework particularly suitable, indeed essential, to a clear formulation of the subject. Conceptually, it is this feature which takes the most getting used to. In particular, one must realize that, as we shall explain later, the interpretation of a state as having a particular "particle content" is in general problematic because it can only be relative to a particular choice of "vacuum" state and, depending on the spacetime of interest, there may be one state or several states or, frequently, no states at all which deserve the name "vacuum" and even when there are states which deserve this name, they will often only be defined in some approximate or asymptotic or transient sense or only on some subregion of the spacetime.

Concomitantly, one does not expect global observables such as the "particle number" or the quantum Hamiltonian of flat-spacetime free-field theory to generalize to a curved spacetime context, and for this reason local observables play a central role in the theory. The quantized stress–energy tensor is a particularly natural and important such local observable and the theory of this is central to the whole subject. A brief introduction to it is given in a later section.

This is followed by a further section on the Hawking and Unruh effects and then a brief section on the problems of extending the theory beyond the "default" setting, to nonglobally hyperbolic spacetimes. Finally, we briefly mention a number of other interesting and active areas of the subject as well as issuing a few *warnings* to be borne in mind when reading the literature.

Construction of $*$-Algebra(s) for a Real Linear Scalar Field on Globally Hyperbolic Spacetimes and Some General Theorems

On a globally hyperbolic spacetime, the classical equation [2] admits well-defined *advanced* and *retarded Green* functions (strictly bidistributions) Δ^A and Δ^R and the standard covariant quantum free real (or "Hermitian") scalar field commutation relations familiar from Minkowski spacetime free-field theory naturally generalize to the (heuristic) equation

$$[\hat{\phi}(x), \hat{\phi}(y)] = i\Delta(x, y)I$$

where Δ is the *Lichnérowicz commutator function* $\Delta = \Delta^A - \Delta^R$. Here, the "$\hat{}$" on the quantum field $\hat{\phi}$ serves to distinguish it from a classical solution ϕ. In mathematical work, one does not assign a meaning to the field at a point itself, but rather aims to assign meaning to *smeared fields* $\hat{\phi}(F)$ for all real-valued test functions $F \in C_0^\infty(\mathcal{M})$ which are then to be interpreted as standing for $\int_{\mathcal{M}} \hat{\phi}(x)f(x)|\det(g)|^{1/2}\, d^4x$. In fact, it is straightforward to define a *minimal field algebra* (see below) $\mathcal{A}_{\min}$ generated by such $\hat{\phi}(F)$ which satisfy the suitably smeared version

$$[\hat{\phi}(F), \hat{\phi}(G)] = i\Delta(F, G)I$$

of the above commutation relations together with Hermiticity (i.e., $\hat{\phi}(F)^* = \hat{\phi}(F)$), the property of being a weak solution of eqn [2] (i.e., $\hat{\phi}((\Box_g - m^2 - V)F) = 0\ \forall F \in C_0^\infty(\mathcal{M})$) and linearity in test functions. There is a technically different alternative formulation of this minimal algebra, which is known as the *Weyl algebra*, which is constructed to be the C*-algebra generated by operators $W(F)$ (to be interpreted as standing for $\exp(i\int_{\mathcal{M}} \hat{\phi}(x)f(x)|\det(g)|^{1/2}\,d^4x)$ satisfying

$$W(F_1)W(F_2) = \exp(-i\Delta(F_1,F_2)/2)W(F_1+F_2)$$

together with $W(F)^* = W(-F)$ and $W((\Box_g - m^2 - V)F) = I$. With either the minimal algebra or the Weyl algebra one can define, for each bounded open region $\mathcal{O}$, subalgebras $A(\mathcal{O})$ as generated by the $\hat{\phi}(\cdot)$ (or the $W(\cdot)$) smeared with test functions supported in $\mathcal{O}$ and verify that they satisfy the above "net" condition and commutativity at spacelike separation.

Specifying a state, ω, on $\mathcal{A}_{min}$ is tantamount to specifying its collection of n-point distributions (i.e., smeared n-point functions) $\omega(\hat{\phi}(F_1)\ldots\hat{\phi}(F_n))$. (In the case of the Weyl algebra, one restricts attention to "regular" states for which the map $F \to \omega(W(F))$ is sufficiently often differentiable on finite-dimensional subspaces of $C_0^\infty(\mathcal{M})$ and defines the n-point distributions in terms of derivatives with respect to suitable parameters of expectation values of suitable Weyl algebra elements.) A particular role is played in the theory by the *quasifree* states for which all the *truncated* n-point distributions except for $n=2$ vanish. Thus, all the n-point distributions for odd n vanish while the four-point distribution is made out of the two-point distribution according to

$$\begin{aligned}\omega(\hat{\phi}(F_1)\hat{\phi}(F_2)\hat{\phi}(F_3)\hat{\phi}(F_4)) &= \omega(\hat{\phi}(F_1)\hat{\phi}(F_2))\omega(\hat{\phi}(F_3)\hat{\phi}(F_4)) \\ &\quad + \omega(\hat{\phi}(F_1)\hat{\phi}(F_3))\omega(\hat{\phi}(F_2)\hat{\phi}(F_4)) \\ &\quad + \omega(\hat{\phi}(F_1)\hat{\phi}(F_4))\omega(\hat{\phi}(F_2)\hat{\phi}(F_3))\end{aligned}$$

etc. The anticommutator distribution

$$G(F_1,F_2) = \omega(\hat{\phi}(F_1)\hat{\phi}(F_2)) + \omega(\hat{\phi}(F_2)\hat{\phi}(F_1)) \qquad [3]$$

of a quasifree state (or indeed of any state) will satisfy the following conditions (for all test functions F, F_1, F_2, etc.):

C1. *Symmetry*

$$G(F_1,F_2) = G(F_2,F_1)$$

C2. *Weak bisolution property*

$$\begin{aligned}G((\Box_g - m^2 - V)F_1, F_2) &= 0 \\ &= G(F_1, (\Box_g - m^2 - V)F_2)\end{aligned}$$

C3. *Positivity*

$$\begin{aligned}G(F,F) \geq 0 \text{ and } G(F_1,F_1)^{1/2}G(F_2,F_2)^{1/2} \\ \geq |\Delta(F_1,F_2)|\end{aligned}$$

and it can be shown that, to every bilinear functional G on $C_0^\infty(\mathcal{M})$ satisfying (C1)–(C3), there is a quasifree state with two-point distribution $(1/2)(G + i\Delta)$. One further declares a quasifree state to be *physically admissible* only if (for pairs of points in sufficiently small convex neighborhoods)

C4. *Hadamard condition*

$$\text{“}G(x_1,x_2) = \frac{1}{2\pi^2}\left(u(x_1,x_2)\mathrm{P}\frac{1}{\sigma} + v(x_1,x_2)\log|\sigma| + w(x_1,x_2)\right)\text{”}$$

This last condition expresses the requirement that (locally) the two-point distribution actually "is" (in the usual sense in which one says that a distribution "is" a function) a smooth function for pairs of non-null-separated points. At the same time, it requires that the two-point distribution be singular at pairs of null-separated points and locally specifies the nature of the singularity for such pairs of points with a leading "principal part of $1/\sigma$" type singularity and a subleading "$\log|\sigma|$" singularity, where σ denotes the square of the geodesic distance between x_1 and x_2. u (which satisfies $u(x_1,x_2) = 1$ when $x_1 = x_2$) and v are certain smooth two-point functions determined in terms of the local geometry and the local values of V by something called the *Hadamard procedure* while the smooth two-point function w depends on the state. We shall omit the details. The important point is that this Hadamard condition on the two-point distribution is believed to be the correct generalization to a curved spacetime of the well-known universal short-distance behavior shared by the truncated two-point distributions of all physically relevant states for the special case of our theory when the spacetime is flat (and V vanishes). In the latter case, u reduces to 1, and v to a simple power series $\sum_{n=0}^{\infty} v_n\sigma^n$ with $v_0 = m^2/4$, etc.

Actually, it is known (this is the content of "Kay's conjecture" which was proved by M Radzikowski in 1992) that (C1)–(C4) together imply that the two-point distribution is nonsingular at all pairs of (not necessarily close together) spacelike separated points. More important than this result itself is a reformulation of the Hadamard condition in terms of the concepts of *microlocal analysis* which Radzikowski originally introduced as a tool towards its proof.

C4′. *Wave front set (or microlocal) spectrum condition*

$$\begin{aligned}&\mathrm{WF}(G+\mathrm{i}\Delta)\\&\quad=\{(x_1,p_1;x_2,p_2)\in T^*(\mathcal{M}\times\mathcal{M})\setminus 0 | x_1 \text{ and } x_2 \text{ lie on a single null geodesic, } p_1 \text{ is tangent to that null geodesic and future pointing, and } p_2 \text{ when parallel transported along that null geodesic from } x_2 \text{ to } x_1 \text{ equals } -p_1\}\end{aligned}$$

For the gist of what this means, it suffices to know that to say that an element (x,p) of the cotangent bundle of a manifold (excluding the zero section 0) is in the *wave front set*, WF, of a given distribution on that manifold may be expressed informally by saying that that distribution is singular *at* the point x *in the direction* p. (And here the notion is applied to $G+\mathrm{i}\Delta$, thought of as a distribution on the manifold $\mathcal{M}\times\mathcal{M}$.)

We remark that generically (and, e.g., always if the spatial sections are compact and $m^2+V(x)$ is everywhere positive) the Weyl algebra for eqn [2] on a given stationary spacetime will have a unique ground state and unique KMS states at each temperature and these will be quasifree and Hadamard.

Quasifree states are important also because of a theorem of R Verch (1994, in verification of another conjecture of Kay) that (in the Weyl algebra framework) on the algebra of any bounded open region, the folia of the quasifree Hadamard states coincide. With this result one can extend the notion of physical admissibility to not-necessarily-quasifree states by demanding that, to be admissible, a state belong to the resulting common folium when restricted to the algebra of each bounded open region; equivalently, that it be a *locally normal* state on the resulting natural extension of the net of local Weyl algebras to a net of local W^*-algebras.

Particle Creation and the Limitations of the Particle Concept

Global hyperbolicity also entails that the Cauchy problem is well posed for the classical field equation [2] in the sense that for every Cauchy surface, $\mathcal{C}$, and every pair (f,p) of Cauchy data in $C_0^\infty(\mathcal{C})$, there exists a unique solution ϕ in $C_0^\infty(\mathcal{M})$ such that $f=\phi|_{\mathcal{C}}$ and $p=|\det(g)|^{1/2}g^{ab}\partial_b\phi|_{\mathcal{C}}$. Moreover, ϕ has compact support on all other Cauchy surfaces. Given a global time coordinate t, increasing towards the future, foliating $\mathcal{M}$ into a family of constant-t Cauchy surfaces, $\mathcal{C}_t$, and given a choice of global timelike vector field τ^a (e.g., $\tau^a=g^{ab}\partial_b t$) enabling one to identify all the $\mathcal{C}_t$, say with $\mathcal{C}_0$, by identifying points cut by the same integral curve of τ^a, a single such classical solution ϕ may be pictured as a family $\{(f_t,p_t): t\in R\}$ of time-evolving Cauchy data on $\mathcal{C}_0$. Moreover, since [2] implies, for each pair of classical solutions, ϕ_1,ϕ_2, the conservation (i.e., $\partial_a j^a=0$) of the current $j^a=|\det(g)|^{1/2}g^{ab}(\phi_1\partial_b\phi_2-\phi_2\partial_b\phi_1)$, the symplectic form (on $C_0^\infty(\mathcal{C})\times C_0^\infty(\mathcal{C})$)

$$\sigma((f_t^1,p_t^1);(f_t^2,p_t^2))=\int_{\mathcal{C}_0}(f_t^1p_t^2-p_t^1f_t^2)\mathrm{d}^3x$$

will be conserved in time.

Corresponding to this picture of classical dynamics, one expects there to be a description of quantum dynamics in terms of a family of sharp-time quantum fields (φ_t,π_t) on $\mathcal{C}_0$, satisfying heuristic canonical commutation relations

$$\begin{aligned}[\varphi_t(\boldsymbol{x}),\varphi_t(\boldsymbol{y})]&=0\\[\pi_t(\boldsymbol{x}),\pi_t(\boldsymbol{y})]&=0\\[\varphi_t(\boldsymbol{x}),\pi_t(\boldsymbol{y})]&=\mathrm{i}\delta^3(\boldsymbol{x},\boldsymbol{y})I\end{aligned}$$

and evolving in time according to the same dynamics as the Cauchy data of a classical solution. (Both these expectations are correct because the field equation is linear.) An elegant way to make rigorous mathematical sense of these expectations is in terms of a $*$-algebra with identity generated by Hermitian objects "$\sigma((\varphi_0,\pi_0);(f,p))$" ("symplectically smeared sharp-time fields at $t=0$") satisfying linearity in f and p together with the commutation relations

$$\begin{aligned}&[\sigma((\varphi_0,\pi_0);(f^1,p^1)),\sigma((\varphi_0,\pi_0);(f^2,p^2))]\\&\quad=\mathrm{i}\sigma((f^1,p^1);(f^2,p^2))I\end{aligned}$$

and to define (symplectically smeared) time-t sharp-time fields by demanding

$$\sigma((\varphi_t,\pi_t);(f_t,p_t))=\sigma((\varphi_0,\pi_0);(f_0,p_0))$$

where (f_t,p_t) is the classical time-evolute of (f_0,p_0). This $*$-algebra of sharp-time fields may be identified with the (minimal) field $*$-algebra of the previous section, the $\hat{\phi}(F)$ of the previous section being identified with $\sigma((\varphi_0,\pi_0);(f,p))$, where (f,p) are the Cauchy data at $t=0$ of $\Delta * F$. (This identification is of course many–one since $\hat{\phi}(F)=0$ whenever F arises as $(\Box_g-m^2-V)G$ for some test function $G\in C_0^\infty(\mathcal{M})$.)

Specializing momentarily to the case of the free scalar field $(\Box-m^2)\phi=0$ $(m\neq 0)$ in Minkowski space with a flat $t=0$ Cauchy surface, the "symplectically smeared" two-point function of the usual ground state ("Minkowski vacuum state"), ω_0, is given, in this formalism, by

$$\begin{aligned}&\omega_0(\sigma((\varphi,\pi);(f^1,p^1))\sigma((\varphi,\pi);(f^2,p^2)))\\&\quad=\tfrac{1}{2}(\langle f^1|\mu f^2\rangle+\langle p^1|\mu^{-1}p^2\rangle\\&\qquad+\mathrm{i}\sigma((f^1,p^1);(f^2,p^2)))\end{aligned}\qquad[4]$$

where the inner products are in the *one-particle Hilbert space* $\mathcal{H} = L^2_C(\boldsymbol{R}^3)$ and $\mu = (m^2 - \nabla^2)^{1/2}$. The GNS representation of this state may be concretely realized on the familiar *Fock space* $\mathcal{F}(\mathcal{H})$ over $\mathcal{H}$ by

$$\rho_0(\sigma((\varphi,\pi);(f,p))) = -\mathrm{i}(\hat{a}^\dagger(\mathrm{a}) - (\hat{a}^\dagger(\mathrm{a}))^*)$$

where a denotes the element of $\mathcal{H}$:

$$\mathrm{a} = \frac{(\mu^{1/2}f + \mathrm{i}\mu^{-1/2}p)}{\sqrt{2}}$$

(we note in passing that, if we equip $\mathcal{H}$ with the symplectic form $2\,\mathrm{Im}\langle\cdot|\cdot\rangle$, then $K:(f,p)\mapsto \mathrm{a}$ is a symplectic map) and $\hat{a}^\dagger(\mathrm{a})$ is the usual smeared *creation operator* (= "$\int \hat{a}^\dagger(\boldsymbol{x})\mathrm{a}(\boldsymbol{x})\mathrm{d}^3\boldsymbol{x}$") on $\mathcal{F}(\mathcal{H})$ satisfying

$$[(\hat{a}^\dagger(\mathrm{a}^1))^*, \hat{a}^\dagger(\mathrm{a}^2)] = \langle \mathrm{a}^1|\mathrm{a}^2\rangle_{\mathcal{H}} I$$

The usual (smeared) *annihilation operator*, $\hat{a}(\mathrm{a})$, is $(\hat{a}^\dagger(C\mathrm{a}))^*$, where C is the natural complex conjugation, $\mathrm{a}\mapsto \mathrm{a}^*$ on $\mathcal{H}$. Both of these operators annihilate the *Fock vacuum vector* $\Omega^{\mathcal{F}}$. In this representation, the one-parameter group of time-translation automorphisms

$$\alpha(t):\sigma((\varphi_0,\pi_0);(f,p))\mapsto\sigma((\varphi_t,\pi_t);(f,p)) \qquad [5]$$

is implemented by $\exp(-\mathrm{i}Ht)$ where H is the second quantization of μ (i.e., the operator otherwise known as $\int\mu(k)\hat{a}^\dagger(k)\hat{a}(k)\mathrm{d}^3k$) on $\mathcal{F}(\mathcal{H})$.

The most straightforward (albeit physically artificial) situation involving "particle creation" in a curved spacetime concerns a globally hyperbolic spacetime which, outside of a compact region, is isometric to Minkowski space with a compact region removed – that is, to a globally hyperbolic spacetime which is flat except inside a localized "bump" of curvature (see **Figure 1**). (One could also allow the function V in [2] to be nonzero inside the bump.) On the field algebra (defined as in the previous section) of such a spacetime, there will be an "in" vacuum state (which may be identified with the Minkowski vacuum to the past of the bump) and an "out" vacuum state (which may be identified with the Minkowski vacuum to the future of the bump) and one expects, for example, the "in vacuum" to arise as a many-particle state in the GNS representation of the "out vacuum" corresponding to the creation of particles out of the vacuum by the bump of curvature.

In the formalism of this section, if we choose our global time coordinate on such a spacetime so that, say, the $t=0$ surface is to the past of the bump and the $t=T$ surface to its future, then the single automorphism $\alpha(T)$ (defined as in [5]) encodes the overall effect of the bump of curvature on the quantum field and one can ask whether it is implemented by a unitary operator in the GNS representation of the Minkowski vacuum state [4].

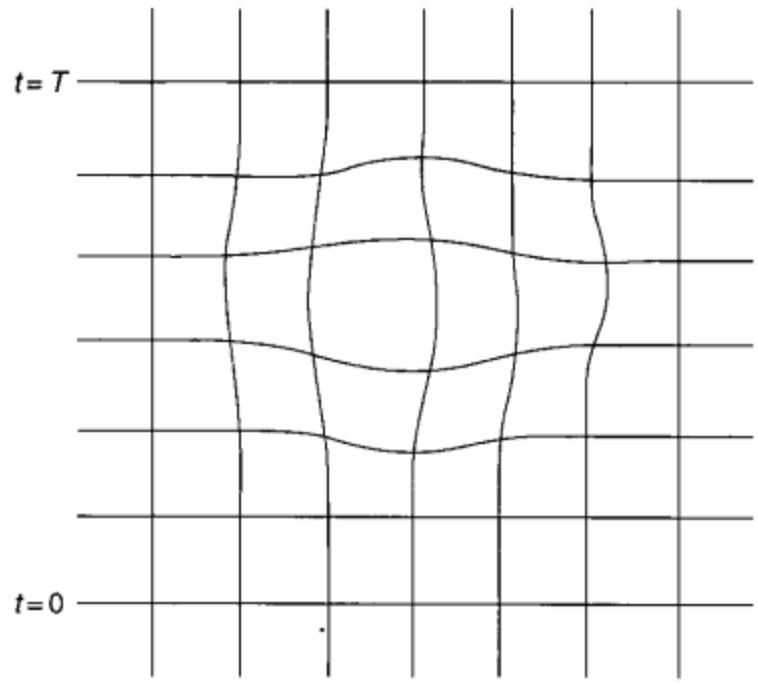


Figure 1 A spacetime which is flat outside of a compact bump of curvature.

This question may be answered by referring to the real linear map $\mathcal{T}:\mathcal{H}\to\mathcal{H}$ which sends $\mathrm{a}_T = 2^{-1/2}(\mu^{1/2}f_T + \mathrm{i}\mu^{-1/2}p_T)$ to $\mathrm{a}_0 = 2^{-1/2}(\mu^{1/2}f_0 + \mathrm{i}\mu^{-1/2}p_0)$. By the conservation in time of σ and the symplecticity, noted in passing above, of the map $K:(f,p)\mapsto\mathrm{a}$, this satisfies the defining relation

$$\mathrm{Im}\langle\mathcal{T}\mathrm{a}^1|\mathcal{T}\mathrm{a}^2\rangle = \mathrm{Im}\langle\mathrm{a}^1|\mathrm{a}^2\rangle$$

of a *classical Bogoliubov transformation*. Splitting $\mathcal{T}$ into its complex-linear and complex-antilinear parts by writing

$$\mathcal{T} = \alpha + \beta C$$

where α and β are complex-linear operators, this relation may alternatively be expressed in terms of the pair of relations

$$\alpha^*\alpha - \bar{\beta}^*\bar{\beta} = I, \qquad \bar{\alpha}^*\bar{\beta} = \beta^*\alpha$$

where $\bar{\alpha} = C\alpha C, \bar{\beta} = C\beta C$.

We remark that there is an easy-to-visualize equivalent way of defining α and β in terms of the analysis, to the past of the bump, into *positive- and negative-frequency parts* of complex solutions to [2] which are purely *positive frequency* to the future of the bump. In fact, if, for any element $\mathrm{a}\in\mathcal{H}$, we identify the positive-frequency solution to the Minkowski-space Klein–Gordon equation

$$\phi^{\mathrm{pos}}_{\mathrm{out}}(t,\boldsymbol{x}) = ((2\mu)^{-1/2}\exp(-\mathrm{i}\mu t)\mathrm{a})(\boldsymbol{x})$$

with a complex solution to [2] to the future of the bump, then (it may easily be seen) to the past of the bump, this same solution will be identifiable with

the (partly positive-frequency, partly negative-frequency) Minkowski-space Klein–Gordon solution

$$\phi_{\text{in}}(t,\boldsymbol{x}) = \left((2\mu)^{-1/2}\exp(-\mathrm{i}\mu t)\alpha\mathrm{a}\right)(\boldsymbol{x}) + \left((2\mu)^{-1/2}\exp(\mathrm{i}\mu t)\bar{\beta}\mathrm{a}\right)(\boldsymbol{x})$$

and this could be taken to be the defining equation for the operators α and β.

It is then known (by a 1962 theorem of Shale) that the automorphism [5] (strictly, its Weyl algebra counterpart) will be unitarily implemented if and only if β is a Hilbert–Schmidt operator on $\mathcal{H}$. Wald (1979, in case $m \geq 0$) and Dimock (1979, in case $m \neq 0$) have verified that this condition is satisfied in the case of our bump-of-curvature situation. In that case, if we denote the unitary implementor by U, we have the following results:

R1. The expectation value $\langle U\Omega|N(\mathrm{a})U\Omega\rangle_{\mathcal{F}(\mathcal{H})}$ of the number operator, $N(\mathrm{a}) = \hat{a}^{\dagger}(\mathrm{a})\hat{a}(\mathrm{a})$, where a is a normalized element of $\mathcal{H}$, is equal to $\langle\beta\mathrm{a}\,|\,\beta\mathrm{a}\rangle_{\mathcal{H}}$.

R2. First note that there exists an orthonormal basis of vectors, $e_i, (i = 1 \ldots \infty)$, in $\mathcal{H}$ such that the (Hilbert–Schmidt) operator $\bar{\beta}^{*}\bar{\alpha}^{*-1}$ has the canonical form $\sum_i \lambda_i\langle Ce_i|\cdot\rangle|e_i\rangle$. We then have (up to an undetermined phase)

$$U\Omega = N\exp\left(-\frac{1}{2}\sum_i \lambda_i\hat{a}^{\dagger}(e_i)\hat{a}^{\dagger}(e_i)\right)\Omega$$

where the normalization constant N is chosen so that $\|U\Omega\| = 1$. This formula makes manifest that the particles are created in pairs.

We remark that, identifying elements, a, of $\mathcal{H}$ with positive-frequency solutions (below, we shall call them "modes") as explained above, result (R1) may alternatively be expressed by saying that the expectation value, $\omega_{\text{in}}(N(\mathrm{a}))$, in the *in-vacuum state* of the occupation number, $N(\mathrm{a})$, of a *normalized mode*, a, to the future of the bump, is given by $\langle\beta\mathrm{a}|\beta\mathrm{a}\rangle_{\mathcal{H}}$.

This formalism and the results, (R1) and (R2) above, will generalize (at least heuristically, and sometimes rigorously – see especially the rigorous scattering-theoretic work in the 1980s by Dimock and Kay and more recently by A Bachelot and others) to more realistic spacetimes which are only asymptotically flat or asymptotically stationary. In favorable cases, one will still have notions of classical solutions which are positive frequency asymptotically towards the future/past, and, in consequence, one will have well-defined asymptotic notions of "vacuum" and "particles." Also, in, for example, cosmological, models where the background spacetime is slowly varying in time, one can define approximate *adiabatic* notions of classical positive-frequency solutions, and hence also of quantum "vacuum" and "particles" at each finite value of the cosmological time. But, at times where the gravitational field is rapidly varying, one does not expect there to be any sensible notion of "particles." And, in a rapidly time-varying background gravitational field which never settles down, one does not expect there to be any sensible particle interpretation of the theory at all. To understand these statements, it suffices to consider the $(1+0)$-dimensional Klein–Gordon equation with an external potential V:

$$\left(-\frac{\mathrm{d}^2}{\mathrm{d}t^2} - m^2 - V(t)\right)\phi = 0$$

which is of course a system of one degree of freedom, mathematically equivalent to the harmonic oscillator with a time-varying angular frequency $\varpi(t) = (m^2 + V(t))^{1/2}$. One could of course express its quantum theory in terms of a time-evolving Schrödinger wave function $\Psi(\varphi, t)$ and attempt to give this a particle interpretation at each time, s, by expanding $\Psi(\varphi, s)$ in terms of the harmonic oscillator wave functions for a harmonic oscillator with some particular choice of angular frequency. But the problem is, as is easy to convince oneself, that there is no such good choice. For example, one might think that a good choice would be to take, at time s, the set of harmonic oscillator wave functions with angular frequency $\varpi(s)$. (This is sometimes known as the method of "instantaneous diagonalization of the Hamiltonian.") But suppose we were to apply this prescription to the case of a smooth $V(\cdot)$ which is constant in time until time 0 and assume the initial state is the usual vacuum state. Then at some positive time s, the number of particles predicted to be present is the same as the number of particles predicted to be present on the same prescription at all times after s for a $\hat{V}(\cdot)$ which is equal to $V(\cdot)$ up to time s and then takes the constant value $V(s)$ for all later times (see **Figure 2**). But $\hat{V}(\cdot)$ will generically have a sharp corner in its graph (i.e., a

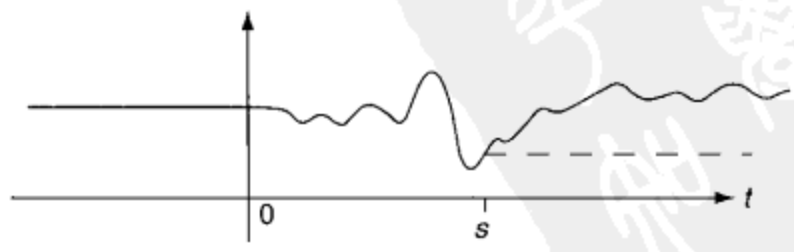


Figure 2 Plots of ϖ against t for the two potentials V (continuous line) and $\hat{V}$ (continuous line upto s and then dashed line) which play a role in our critique of "instantaneous diagonalization."

discontinuity in its time derivative) at time s, and one would expect a large part of the particle production in the latter situation to be accounted for by the presence of this sharp corner – and therefore a large part of the predicted particle production in the case of $V(\cdot)$ to be spurious.

Back in $1+3$ dimensions, even where a good notion of particles is possible, it depends on the choice of time evolution, as is dramatically illustrated by the Unruh effect discussed in the relevant section.

Theory of the Stress–Energy Tensor

To orient ideas, consider first the free (minimally coupled) scalar field, $(\Box - m^2)\phi = 0$, in Minkowski space. If one quantizes this system in the usual Minkowski-vacuum representation, then the expectation value of the *renormalized stress-energy tensor* (which in this case is the same thing as the *normal ordered stress–energy tensor*) in a vector state Ψ in the Fock space will be given by the formal *point-splitting* expression

$$\begin{aligned}&\langle\Psi|T_{ab}(x)\Psi\rangle\\&\quad=\lim_{(x_1,x_2)\to(x,x)}\left(\partial_a^1\partial_b^2-\tfrac{1}{2}\eta_{ab}(\eta^{cd}\partial_c^1\partial_d^2+m^2)\right)\\&\qquad\times(\langle\Psi|\rho_0(\phi(x_1)\phi(x_2))\Psi\rangle\\&\qquad-\langle\Omega^{\mathcal{F}}|\rho_0(\phi(x_1)\phi(x_2))\Omega^{\mathcal{F}}\rangle)\end{aligned}\qquad[6]$$

where η_{ab} is the usual Minkowski metric. A sufficient condition for the limit here to be finite and well defined would, for example, be for Ψ to consist of a (normalized) finite superposition of n-particle vectors of form $\hat{a}^\dagger(\mathrm{a}_1),\ldots,\hat{a}^\dagger(\mathrm{a}_n)\Omega^{\mathcal{F}}$ where the smearing functions $\mathrm{a}_1,\ldots,\mathrm{a}_n$ are all C^∞ elements of $\mathcal{H}$ (i.e., of $L^2_{\mathrm{C}}(\boldsymbol{R}^3)$. The reason this works is that the two-point function in such states shares the same short-distance singularity as the Minkowski-vacuum two-point function. For exactly the same reason, one obtains a well-defined finite limit if one defines the expectation value of the stress–energy tensor in any physically admissible quasifree state by the expression

$$\begin{aligned}&\omega(T_{ab}(x))\\&\quad=\lim_{(x_1,x_2)\to(x,x)}\left(\partial_a^1\partial_b^2-\tfrac{1}{2}\eta_{ab}(\eta^{cd}\partial_c^1\partial_d^2+m^2)\right)\\&\qquad\times(\omega(\phi(x_1)\phi(x_2))-\omega_0(\phi(x_1)\phi(x_2)))\end{aligned}\qquad[7]$$

This latter point-splitting formula generalizes to a definition for the *expectation value of the renormalized stress–energy tensor* for an arbitrary physically admissible quasifree state (or indeed for an arbitrary state whose two-point function has *Hadamard form* – i.e., whose anticommutator function satisfies condition (C4)) on the minimal field algebra and to other linear field theories (including the stress tensor for a conformally coupled linear scalar field) on a general globally hyperbolic spacetime (and the result obtained agrees with that obtained by other methods, including *dimensional regularization* and *zeta-function regularization*). However, the generalization to a curved spacetime involves a number of important new features which we now briefly list (see Wald (1978) for details).

First, the subtraction term which replaces $\omega_0(\phi(x_1)\phi(x_2))$ is, in general, not the expectation value of $\phi(x_1)\phi(x_2)$ in any particular state, but rather a particular *locally constructed Hadamard two-point function* whose physical interpretation is more subtle; the renormalization is thus in general not to be regarded as a normal ordering. Second, the immediate result of the resulting limiting process will not be covariantly conserved and, in order to obtain a covariantly conserved quantity, one needs to add a particular *local geometrical correction term*. The upshot of this is that the resulting expected stress–energy tensor is covariantly conserved but possesses a (state-independent) *anomalous trace*. In particular, for a massless conformally coupled linear scalar field, one has (for all physically admissible quasifree states, ω) the *trace anomaly formula*

$$\omega(T^a_a(x))=(2880\pi^2)^{-1}\left(C_{abcd}C^{abcd}+R_{ab}R^{ab}-\tfrac{1}{3}R^2\right)$$

plus an arbitrary multiple of $\Box R$. In fact, in general, the thus-defined renormalized stress–energy tensor operator (see below) is only defined up to a *finite renormalization ambiguity* which consists of the addition of arbitrary multiples of the functional derivatives with respect to g_{ab} of the quantities

$$I_n=\int_{\mathcal{M}}F_n(x)|\det(g)|^{1/2}\mathrm{d}^4x$$

where n ranges from 1 to 4 with $F_1=1$, $F_2=R$, $F_3=R^2$, and $F_4=R_{ab}R^{ab}$. In the Minkowski-space case, only the first of these ambiguities arises and it is implicitly resolved in the formulas [6], [7] inasmuch as these effectively incorporate the *renormalization condition* that $\omega_0(T_{ab})=0$. (For the same reason, the locally flat example we give below has no ambiguity.)

One expects, in both flat and curved cases, that, for test functions, $F\in C_0^\infty(\mathcal{M})$, there will exist operators $T_{ab}(F)$ which are *affiliated to* the net of

local W^*-algebras referred to earlier and that it is meaningful to write

$$\int_{\mathcal{M}} \omega(T_{ab}(x))F(x)|\det(g)|^{1/2}\mathrm{d}^4x = \omega(T_{ab}(F))$$

provided that, by ω on the right-hand side, we understand the extension of ω from the Weyl algebra to this net. ($T_{ab}(F)$ is however not expected to belong to the minimal algebra or be affiliated to the Weyl algebra.)

An interesting simple example of a renormalized stress–energy tensor calculation is the so-called *Casimir effect* calculation for a linear scalar field on a (for further simplicity, $(1+1)$-dimensional) timelike cylinder spacetime of radius R (see **Figure 3**). This spacetime is globally hyperbolic and stationary and, while locally flat, globally distinct from Minkowski space. As a result, while – provided the regions $\mathcal{O}$ are sufficiently small (such as the diamond region in **Figure 3**) – elements $\mathcal{A}(\mathcal{O})$ of the minimal net of local algebras on this spacetime will be identifiable, in an obvious way, with elements of the minimal net of local algebras on Minkowski space, the stationary ground state ω_{cylinder} will, when restricted to such thus-identified regions, be distinct from the Minkowski vacuum state ω_0. The resulting renormalized stress–energy tensor (as first pointed out in Kay (1979)), definable, once the above identification has been made, exactly as in [7]) turns out, in the massless case, to be nonzero and, interestingly, to have a (in the natural coordinates, constant) *negative energy-density* T_{00}. In fact, in this massless case,

$$\omega_{\text{cylinder}}(T_{ab}) = \frac{1}{24\pi R^2}\eta_{ab}$$

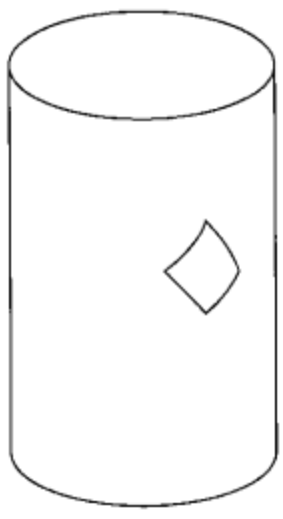


Figure 3 The timelike cylinder spacetime of radius R with a diamond region isometric to a piece of Minkowski space. See Kay (1979). Casimir effect in quantum field theory. (Original title: The Casimir effect without magic.) *Physical Review D* 20: 3052–3062. Reprinted with permission © 1979 by the American Physical Society.

Hawking and Unruh Effects

The original calculation by Hawking (1975) concerned a model spacetime for a star which collapses to a black hole. For simplicity, we shall only discuss the spherically symmetric case (see **Figure 4**). Adopting a similar "mode" viewpoint to that mentioned after results (R1) and (R2) discussed earlier, the result of the calculation may be stated as follows: For a real linear scalar field satisfying [2] with $m=0$ (and $V=0$) on this spacetime, the expectation value $\omega_{\text{in}}(N(\mathrm{a}_{\varpi,\ell}))$ of the occupation number of a one-particle outgoing mode $\mathrm{a}_{\varpi,\ell}$ localized (as far as a normalized mode can be) around ϖ in angular-frequency space and about retarded time v, and with angular momentum "quantum number" ℓ, in the in-vacuum state (i.e., on the minimal algebra for a real scalar field on this model spacetime) ω_{in} is, at late retarded times, given by the formula

$$\omega_{\text{in}}(N(\mathrm{a}_{\varpi,\ell})) = \frac{\Gamma(\varpi,\ell)}{\exp(8\pi M\varpi) - 1}$$

where M is the mass of the black hole and the *absorption factor* (alternatively known as *gray-body factor*) $\Gamma(\varpi,\ell)$ is equal to the norm-squared of that part of the one-particle mode $\mathrm{a}_{\varpi,\ell}$ which, viewed as a complex positive-frequency classical solution propagating backwards in time from late retarded times, would be absorbed by the black hole. (Note the independence of the right-hand side of this formula from the retarded time, v.) This calculation can be understood as an application of result (R1)

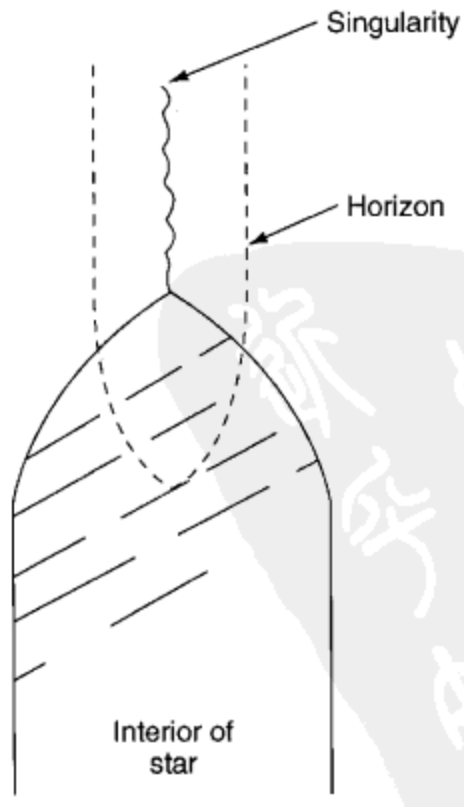


Figure 4 The spacetime of a star collapsing to a spherical black hole.

(even though the spacetime is more complicated than one with a localized "bump of curvature" and even though the relevant overall time evolution will not be unitarily implemented, the result still applies when suitably interpreted) and the heart of the calculation is an asymptotic estimate of the relevant "β" Bogoliubov coefficient which turns out to be dependent on the geometrical optics of rays which pass through the star just before the formation of the horizon. This result suggests that the in-vacuum state is indistinguishable at late retarded times from a state of blackbody radiation at the *Hawking temperature*, $T_{\text{Hawking}} = 1/8\pi M$, in Minkowski space from a blackbody (gray body) with the same absorption factor. This was confirmed by further work by many authors. Much of that work, as well as the original result of Hawking was partially heuristic but later work by Dimock and Kay (1987), by Fredenhagen and Haag (1990), and by Bachelot (1999) and others has put different aspects of it on a rigorous mathematical footing. The result generalizes to nonzero mass and higher spin fields to interacting fields as well as to other types of black hole and the formula for the Hawking temperature generalizes to

$$T_{\text{Hawking}} = \kappa/2\pi$$

where κ is the *surface gravity* of the black hole.

This result suggests that there is something fundamentally "thermal" about quantum fields on black-hole backgrounds and this is confirmed by a number of mathematical results. In particular, the theorems in the two papers Kay and Wald (1991) and Kay (1993), combined together, tell us that there is a unique state on the Weyl algebra for the *maximally extended Schwarzschild spacetime* (a.k.a. *Kruskal–Szekeres spacetime*) (see **Figure 5**) which is invariant under the *Schwarzschild isometry group* and whose two-point function has Hadamard form. Moreover, they tell us that this state, when restricted to a single wedge (i.e., the exterior Schwarzschild spacetime) is necessarily a KMS state at the Hawking temperature. This unique state is known as the *Hartle–Hawking–Israel state*. These results in fact apply more generally to a wide class of globally hyperbolic spacetimes with *bifurcate Killing horizons* including de Sitter space – where the unique state is sometimes called the *Euclidean* and sometimes the *Bunch–Davies* vacuum state – as well as to Minkowski space, in which case the unique state is the usual Minkowski vacuum state, the analog of the exterior Schwarzschild wedge is a so-called *Rindler wedge*, and the relevant isometry group is a one-parameter family of wedge-preserving Lorentz boosts. In the latter situation, the fact that the Minkowski vacuum state is a KMS state (at "temperature" $1/2\pi$)

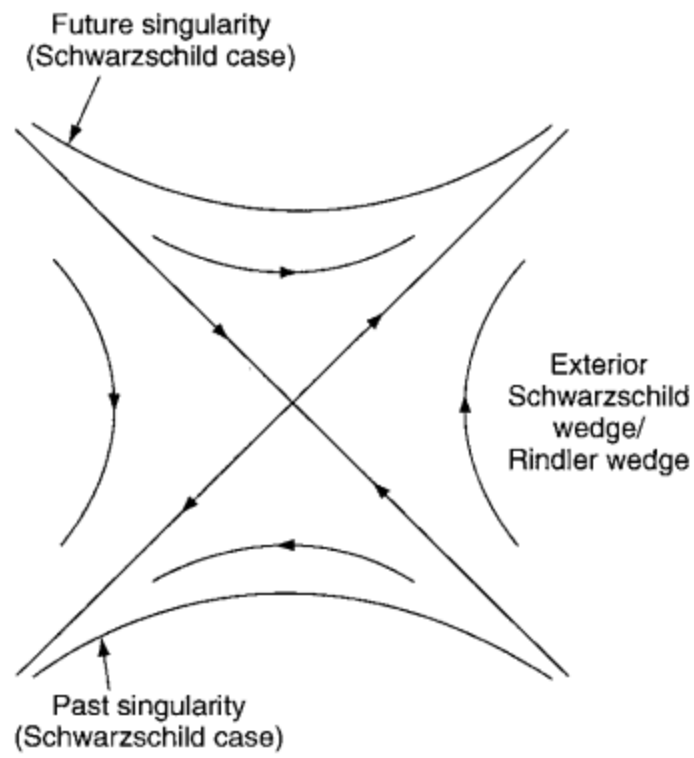


Figure 5 The geometry of maximally extended Schwarzschild (/or Minkowski) spacetime. In the Schwarzschild case, every point represents a 2-sphere (/in the Minkowski case, a 2-plane). The curves with arrows on them indicate the Schwarzschild time evolution (/one-parameter family of Lorentz boosts). These curves include the (straight lines at right angles) event horizons (/Killing horizons).

when restricted to a Rindler wedge and regarded with respect to the time evolution consisting of the wedge-preserving one-parameter family of Lorentz boosts is known as the *Unruh effect* (1975). This latter property of the Minkowski vacuum in fact generalizes to general *Wightman QFTs* and is in fact an immediate consequence of a combination of the *Reeh–Schlieder theorem* (applied to a Rindler wedge) and the *Bisognano–Wichmann theorem* (1975). The latter theorem says that the defining relation [1] of a KMS state holds if, in [1], we identify the operator J with the complex conjugation which implements wedge reflection and H with the self-adjoint generator of the unitary implementor of Lorentz boosts. We remark that the Unruh effect illustrates how the concept of "vacuum" (when meaningful at all) is dependent on the choice of time evolution under consideration. Thus, the usual Minkowski vacuum is a ground state with respect to the usual Minkowski time evolution but not (when restricted to a Rindler wedge) with respect to a one-parameter family of Lorentz boosts; with respect to these, it is, instead, a KMS state.

Nonglobally Hyperbolic Spacetimes and the "Time Machine" Question

Hawking (1992) argued that a spacetime in which a time machine gets manufactured should be modeled (see **Figure 6**) by a spacetime with an initial *globally*

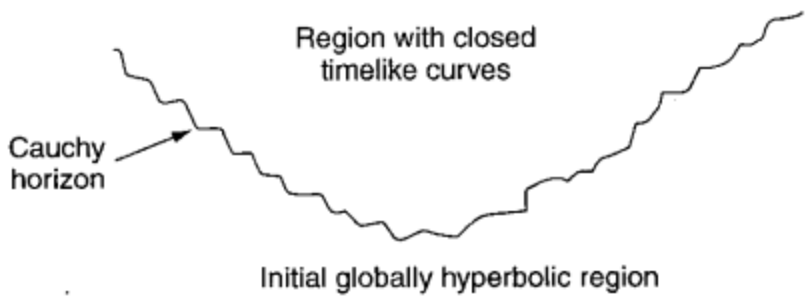


Figure 6 The schematic geometry of a spacetime in which a time machine gets manufactured.

hyperbolic region with a region containing closed timelike curves to its future and such that the future boundary of the globally hyperbolic region is a *compactly generated Cauchy horizon*. On such a spacetime, Kay *et al.* (1997) proved that it is impossible for any distributional bisolution which satisfies (even a certain weakened version of) the Hadamard condition on the initial globally hyperbolic region to continue to satisfy that condition on the full spacetime – the (weakened) Hadamard condition being necessarily violated at at least one point *on* the Cauchy horizon. This result implies that, however one extends a state, satisfying our conditions (C1)–(C4), on the minimal algebra for [2] on the initial globally hyperbolic region, the expectation value of its stress–energy tensor must necessarily become singular on the Cauchy horizon. This result, together with many heuristic results and specific examples considered by many other authors appears to support the validity of the (Hawking 1992) *chronology protection conjecture* to the effect that it is impossible in principle to manufacture a time machine. However, there are potential loopholes in the physical interpretation of this result as pointed out by Visser (1997), as well as other claims by various authors that one can nevertheless violate the chronology protection conjecture. For a recent discussion on this question, we refer to Visser (2003).

Other Related Topics and Some Warnings

There is a vast computational literature, calculating the expectation values of stress–energy tensors in states of interest for scalar and higher spin linear fields (and also some work for interacting fields) on interesting cosmological and black-hole backgrounds. QFT on de Sitter and anti-de Sitter space is a big subject area in its own right with recent renewed interest because of its relevance to *string theory* and *holography*. Also important on black-hole backgrounds is the calculation of gray-body factors, again with renewed interest because of relevance to string theory and to *brane-world* scenarios.

There are many further mathematically rigorous results on algebraic and axiomatic QFT in a curved spacetime setting, including versions of *PCT*, *spin-statistics* and *Reeh–Schlieder* theorems and also rigorous *energy inequalities* bounding the extent to which expected energy densities can be negative, etc.

There is much mathematical work controlling scattering theory on black holes, partly with a view to further elucidating the Hawking effect.

Perturbative renormalization theory of interacting quantum fields in curved spacetime is also now a highly developed subject.

Beyond QFT in a fixed curved spacetime is *semiclassical gravity* which takes into account the *back-reaction* of the expectation value of the stress–energy tensor on the classical gravitational background. There are also interesting condensed matter analogs of the Hawking effect such as *dumb holes*.

Readers exploring the wider literature, or doing further research on the subject should be aware that the word "vacuum" is sometimes used to mean "ground state" and sometimes just to mean "quasifree state." They should be cautious of attempts to define particles on Cauchy surfaces in *instantaneous diagonalization* schemes (cf. the remarks at the end of the section "Particle creation and the limitations of the particle concept"). When studying (or performing) calculations of the "expectation value of the stress–energy tensor" it is always important to ask oneself with respect to which state the expectation value is being taken. It is also important to remember to check that candidate two-point (anticommutator) functions satisfy the positivity condition (C3) discussed earlier. Typically, two-point distributions obtained via *mode sums* automatically satisfy condition (C3) (and condition (C4)), but those obtained via *image* methods do not always satisfy it. (When they do not, the presence of nonlocal spacelike singularities is often a tell-tale sign as can be inferred from Kay's conjecture/Radzikowski's theorem discussed earlier.) There are a number of apparent implicit assertions in the literature that some such two-point functions arise from "states" when of course they cannot. Some of these concern proposed analogs to the Hartle–Hawking–Israel state for the (appropriate maximal globally hyperbolic portion of the maximally extended) Kerr spacetime. That they cannot belong to states is clear from a theorem in Kay and Wald (1991) which states that there is no stationary Hadamard state on this spacetime at all. Others of them concern claimed "states" on spacetimes such as those discussed in the previous section which, if they really were states would seem to be in conflict with the chronology protection conjecture. Finally, beware states (such as the so-called α-*vacua* of de Sitter spacetime) whose two-point

distributions violate the "Hadamard" condition (C4) and which therefore do not have a well-defined finite expectation value for the renormalized stress–energy tensor.

See also: AdS/CFT Correspondence; Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Black Hole Mechanics; Bosons and Fermions in External Fields; Integrability and Quantum Field Theory; Quantum Fields with Indefinite Metric: Non-Trivial Models; Quantum Fields with Topological Defects; Quantum Geometry and Its Applications; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Thermal Quantum Field Theory.

Further Reading

Birrell ND and Davies PCW (1982) *Quantum Fields in Curved Space*. Cambridge: Cambridge University Press.

Brunetti R, Fredenhagen K, and Verch R (2003) The generally covariant locality principle – a new paradigm for local quantum physics. *Communications in Mathematical Physics* 237: 31–68.

DeWitt BS (1975) Quantum field theory in curved space-time. *Physics Reports* 19(6): 295–357.

Dimock J (1980) Algebras of local observables on a manifold. *Communications in Mathematical Physics* 77: 219–228.

Haag R (1996) *Local Quantum Physics*. Berlin: Springer.

Hartle JB and Hawking SW (1976) Path-integral derivation of black-hole radiance. *Physical Review D* 13: 2188–2203.

Hawking SW (1975) Particle creation by black holes. *Communications in Mathematical Physics* 43: 199–220.

Hawking SW (1992) The chronology protection conjecture. *Physical Review D* 46: 603–611.

Israel W (1976) Thermo-field dynamics of black holes. *Physics Letters A* 57: 107–110.

Kay BS (1979) Casimir effect in quantum field theory. (Original title: The Casimir effect without magic.) *Physical Review D* 20: 3052–3062.

Kay BS (1993) Sufficient conditions for quasifree states and an improved uniqueness theorem for quantum fields on space-times with horizons. *Journal of Mathematical Physics* 34: 4519–4539.

Kay BS (2000) Application of linear hyperbolic PDE to linear quantum fields in curved spacetimes: especially black holes, time machines and a new semi-local vacuum concept. *Journées Équations aux Dérivées Partielles*, Nantes, 5–9 juin 2000, GDR 1151 (CNRS): IX1–IX19. (Also available at http://www.math.sciences.univ-nantes.fr or as gr-qc/0103056.)

Kay BS, Radzikowski MJ, and Wald RM (1997) Quantum field theory on spacetimes with a compactly generated Cauchy horizon. *Communications in Mathematical Physics* 183: 533–556.

Kay BS and Wald RM (1991) Theorems on the uniqueness and thermal properties of stationary, nonsingular, quasifree states on spacetimes with a bifurcate Killing horizon. *Physics Reports* 207(2): 49–136.

Misner CW, Thorne KS, and Wheeler JA (1973) *Gravitation*. San Francisco: W.H. Freeman.

Unruh W (1976) Notes on black hole evaporation. *Physical Review D* 14: 870–892.

Visser M (2003) The quantum physics of chronology protection. In: Gibbons GW, Shellard EPS, and Rankin SJ (eds.) *The Future of Theoretical Physics and Cosmology*. Cambridge: Cambridge University Press.

Wald RM (1978) Trace anomaly of a conformally invariant quantum field in a curved spacetime. *Physical Review D* 17: 1477–1484.

Wald RM (1994) *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics*. Chicago: University of Chicago Press.

Quantum Phase Transitions

S Sachdev, Yale University, New Haven, CT, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The study of second-order phase transitions at nonzero temperatures has a long and distinguished history in statistical mechanics. Many key physical phenomena, such as the loss of ferromagnetism in iron at the Curie temperature or the critical endpoint of CO_2, are now understood in precise quantitative detail. This understanding began in the work of Onsager, and is based upon what may now be called the Landau–Ginzburg–Wilson theory. The content of this sophisticated theory may be summarized in a few basic principles: (1) The collective thermal fluctuations near second-order transitions can be accurately described by simple classical models, that is, quantum-mechanical effects can be entirely neglected. (2) The classical models identify an "order parameter," a collective variable which has to be treated on par with other thermodynamic variables, and whose correlations exhibit distinct behavior in the phases on either side of the transition. (3) The thermal fluctuations of the order parameter near the transition are controlled by a continuum field theory whose structure is usually completly dictated by simple symmetry considerations.

Notes:

This article will not consider such nonzero temperature phase transitions, but will instead describe second-order phase transitions at the absolute zero of temperature. Such transitions are driven by quantum fluctuations mandated by the Heisenberg uncertainty principle: one can imagine moving across the quantum critical point by effectively "tuning the value of Planck's constant, $\hbar$." Clearly, quantum mechanics plays a central role at such transitions, unlike the situation at nonzero temperatures. The reader may object that absolute zero is an idealization not realized by any experimental system; hence, the study of quantum phase transitions is a subject only of academic interest. As we will illustrate below, knowledge of the zero-temperature quantum critical points of a system is often the key to understanding its finite-temperature properties, and in some cases the influence of a zero-temperature critical point can be detected at temperatures as high as ambient room temperature.

We will begin in the following section by introducing some simple lattice models which exhibit quantum phase transitions. Next the theory of the critical point in these models is based upon a natural extension of the Landau–Ginzburg–Wilson (LGW) method, and this will be presented. This section will also describe the consequences of a zero-temperature critical point on the nonzero temperature properties. Finally, we will consider more complex models in which quantum interference effects play a more subtle role, and which cannot be described in the LGW framework: such quantum critical points are likely to play a central role in understanding many of the correlated electron systems of current interest.

Simple Models

Quantum Ising Chain

This is a simple model of N qubits, labeled by the index $j = 1, \ldots, N$. On each "site" j there are two qubit quantum states $|\uparrow\rangle_j$ and $|\downarrow\rangle_j$ (in practice, these could be two magnetic states of an ion at site j in a crystal). The Hilbert space therefore consists of 2^N states, each consisting of a tensor product of the states on each site. We introduce the Pauli spin operators, $\hat{\sigma}_j^\alpha$, on each site j, with $\alpha = x, y, z$:

$$\hat{\sigma}^x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \hat{\sigma}^y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$
$$\hat{\sigma}^z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad [1]$$

These operators clearly act on the two states of the qubit on site j, and the Pauli operators on different sites commute.

The quantum Ising chain is defined by the simple Hamiltonian

$$H_I = -J \sum_{j=1}^{N-1} \hat{\sigma}_j^z \hat{\sigma}_{j+1}^z - gJ \sum_{j=1}^{N} \hat{\sigma}_j^x \qquad [2]$$

where $J > 0$ sets the energy scale, and $g \geq 0$ is a dimensionless coupling constant. In the thermodynamic limit ($N \to \infty$), the ground state of H_I exhibits a second-order quantum phase transition as g is tuned across a critical value $g = g_c$ (for the specific case of H_I it is known that $g_c = 1$), as we will now illustrate.

First, consider the ground state of H_I for $g \ll 1$. At $g = 0$, there are two degenerate "ferromagnetically ordered" ground states

$$|\Uparrow\rangle = \prod_{j=1}^{N} |\uparrow\rangle_j, \qquad |\Downarrow\rangle = \prod_{j=1}^{N} |\downarrow\rangle_j \qquad [3]$$

Each of these states breaks a discrete "Ising" symmetry of the Hamiltonian rotations of all spins by 180° about the x-axis. These states are more succinctly characterized by defining the ferromagnetic moment, N_0, by

$$N_0 = \langle \Uparrow | \hat{\sigma}_j^z | \Uparrow \rangle = -\langle \Downarrow | \hat{\sigma}_j^z | \Downarrow \rangle \qquad [4]$$

At $g = 0$ we clearly have $N_0 = 1$. A key point is that in the thermodynamic limit, this simple picture of the ground state survives for a finite range of small g (indeed, for all $g < g_c$), but with $0 < N_0 < 1$. The quantum tunneling between the two ferromagnetic ground states is exponentially small in N (and so can be neglected in the thermodynamic limit), and so the ground state remains 2-fold degenerate and the discrete Ising symmetry remains broken. The change in the wave functions of these states from eqn [3] can be easily determined by perturbation theory in g: these small g quantum fluctuations reduce the value of N_0 from unity but do not cause the ferromagnetism to disappear.

Now consider the ground state of H_I for $g \gg 1$. At $g = \infty$ there is a single nondegenerate ground state which fully preserves all symmetries of H_I:

$$|\Rightarrow\rangle = 2^{-N/2} \prod_{j=1}^{N} \left(|\uparrow\rangle_j + |\downarrow\rangle_j \right) \qquad [5]$$

It is easy to verify that this state has no ferromagnetic moment $N_0 = \langle \Rightarrow | \hat{\sigma}_j^z | \Rightarrow \rangle = 0$. Further, perturbation theory in $1/g$ shows that these features of the ground state are preserved for a finite range of large g values

(indeed, for all $g > g_c$). One can visualize this ground state as one in which strong quantum fluctuations have destroyed the ferromagnetism, with the local magnetic moments quantum tunneling between "up" and "down" on a timescale of order $\hbar/J$.

Given the very distinct signatures of the small g and large g ground states, it is clear that the ground state cannot evolve smoothly as a function of g. These must be at least one point of nonanalyticity as a function of g: for H_I it is known that there is only a single nonanalytic point, and this is at the location of a second-order quantum phase transition at $g = g_c = 1$.

The character of the excitations above the ground state also undergoes a qualitative change across the quantum critical point. In both the $g < g_c$ and $g > g_c$ phases, these excitations can be described in the Landau quasiparticle scheme, that is, as superpositions of nearly independent particle-like excitations; a single well-isolated quasiparticle has an infinite lifetime at low excitation energies. However, the physical nature of the quasiparticles is very different in the two phases. In the ferromagnetic phase, with $g < g_c$, the quasiparticles are domain walls between regions of opposite magnetization:

$$|j, j+1\rangle = \prod_{k=1}^{j} |\uparrow\rangle_k \prod_{\ell=j+1}^{N} |\downarrow\rangle_\ell \qquad [6]$$

This is the exact wave function of a stationary quasiparticle excitation between sites j and $j+1$ at $g = 0$; for small nonzero g the quasiparticle acquires a "cloud" of further spin-flips and also becomes mobile. However its qualitative interpretation as a domain wall between the two degenerate ground states remains valid for all $g < g_c$. In contrast, for $g > g_c$, there is no ferromagnetism, and the nondegenerate paramagnetic state has a distinct quasiparticle excitation:

$$|j\rangle = 2^{-N/2}\left(|\uparrow\rangle_j - |\downarrow\rangle_j\right) \prod_{k \neq j} \left(|\uparrow\rangle_k + |\downarrow\rangle_k\right) \qquad [7]$$

This is a stationary "flipped spin" quasiparticle at site j, with its wave function exact at $g = \infty$. Again, this quasiparticle is mobile and applicable for all $g > g_c$, but there is no smooth connection between eqns [7] and [6].

Coupled Dimer Antiferromagnet

This model also involves qubits, but they are now placed on the sites, j, of a two-dimensional square lattice. Models in this class describe the magnetic excitations of many experimentally important spin gap compounds.

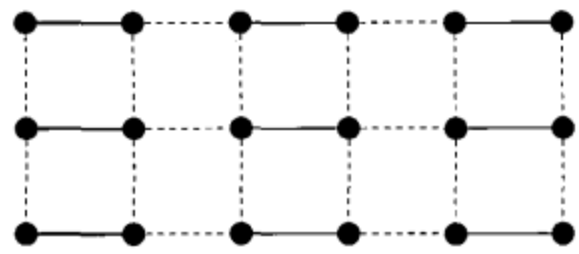


Figure 1 The coupled dimer antiferromagnet. Qubits (i.e., $S = 1/2$ spins) are placed on the sites, the $\mathcal{A}$ links are shown as full lines, and the $\mathcal{B}$ links as dashed lines.

The Hamiltonian of the dimer antiferromagnet is illustrated in **Figure 1** and is given by

$$H_d = J \sum_{\langle jk \rangle \in A} \left(\hat{\sigma}_j^x \hat{\sigma}_k^x + \hat{\sigma}_j^y \hat{\sigma}_k^y + \hat{\sigma}_j^z \hat{\sigma}_k^z\right) + \frac{J}{g} \sum_{\langle jk \rangle \in B} \left(\hat{\sigma}_j^x \hat{\sigma}_k^x + \hat{\sigma}_j^y \hat{\sigma}_k^y + \hat{\sigma}_j^z \hat{\sigma}_k^z\right) \qquad [8]$$

where $J > 0$ is the exchange constant, $g \geq 1$ is the dimensionless coupling, and the set of nearest-neighbor links $\mathcal{A}$ and $\mathcal{B}$ are defined in **Figure 1**. An important property of H_d is that it is now invariant under the full O(3) group of spin rotations under which the $\hat{\sigma}^\alpha$ transform as ordinary vectors (in contrast to the Z_2 symmetry group of H_I). In analogy with H_I, we will find that H_d undergoes a quantum phase transition from a paramagnetic phase which preserves all symmetries of the Hamiltonian at large g, to an antiferromagnetic phase which breaks the O(3) symmetry at small g. This transition occurs at a critical value $g = g_c$, and the best current numerical estimate is $1/g_c = 0.52337(3)$.

As in the previous section, we can establish the existence of such a quantum phase transition by contrasting the disparate physical properties at large g with those at $g \approx 1$. At $g = \infty$ the exact ground state of H_d is

$$|\text{spin gap}\rangle = \prod_{\langle jk \rangle \in A} \frac{1}{\sqrt{2}}\left(|\uparrow\rangle_j|\downarrow\rangle_k - |\downarrow\rangle_j|\uparrow\rangle_k\right) \qquad [9]$$

and is illustrated in **Figure 2**. This state is nondegenerate and invariant under spin rotations, and so is a paramagnet: the qubits are paired into spin singlet valence bonds across all the $\mathcal{A}$ links.

The excitations above the ground state are created by breaking a valence bond, so that the pair of spins form a spin triplet with total spin $S = 1$ – this is illustrated in **Figure 3**. It costs a large energy to create this excitation, and at finite g the

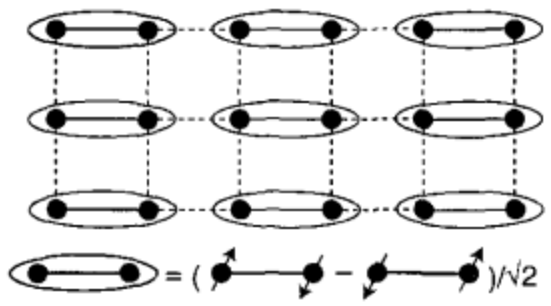


Figure 2 The paramagnetic state of H_d for $g > g_c$. The state illustrated is the exact ground state for $g = \infty$, and it is adiabatically connected to the ground state for all $g > g_c$.

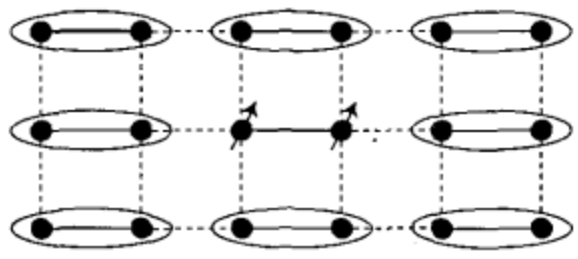


Figure 3 The triplon excitation of the $g > g_c$ paramagnet. The stationary triplon is an eigenstate only for $g = \infty$ but it becomes mobile for finite g.

triplet can hop from link to link, creating a gapped "triplon" quasiparticle excitation. This is similar to the large g paramagnet for H_I, with the important difference that each quasiparticle is now 3-fold degenerate.

At $g = 1$, the ground state of H_d is not known exactly. However, at this point H_d becomes equivalent to the nearest-neighbor square lattice antiferromagnet, and this is known to have antiferromagnetic order in the ground state, as illustrated in **Figure 4**. This state is similar to the ferromagnetic ground state of H_I, with the difference that the magnetic moment now acquires a staggered pattern on the two sublattices, rather than the uniform moment of the ferromagnet. Thus, in this ground state

$$\langle \mathrm{AF}|\hat{\sigma}_j^\alpha|\mathrm{AF}\rangle = N_0 \eta_j n_\alpha \qquad [10]$$

where $0 < N_0 < 1$ is the antiferromagnetic moment, $\eta_j = \pm 1$ identifies the two sublattices in **Figure 4**, and n_α is an arbitrary unit vector specifying the orientation of the spontaneous magnetic moment which breaks the O(3) spin rotation invariance of H_d. The excitations above this antiferromagnet are also distinct from those of the paramagnet: they are a doublet of spin waves consisting of a spatial variation in the local orientation, n_α, of the antiferromagnetic order: the energy of this excitation vanishes in the limit of long wavelengths, in contrast to the finite energy gap of the triplon excitation of the paramagnet.

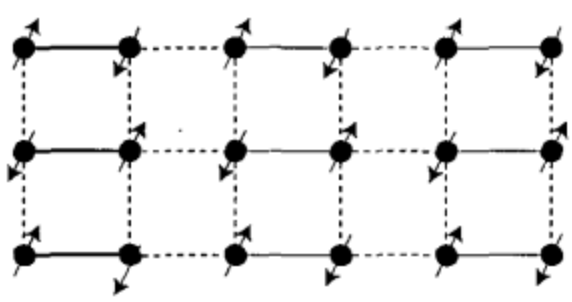


Figure 4 Schematic of the ground state with antiferromagnetic order with $g < g_c$.

As with H_I, we can conclude from the distinct characters of the ground states and excitations for $g \gg 1$ and $g \approx 1$ that there must be a quantum critical point at some intermediate $g = g_c$.

Quantum Criticality

The simple considerations of the previous section have given a rather complete description (based on the quasiparticle picture) of the physics for $g \ll g_c$ and $g \gg g_c$. We turn, finally, to the region $g \approx g_c$. For the specific models discussed in the previous section, a useful description is obtained by a method that is a generalization of the LGW method developed earlier for thermal phase transitions. However, some aspects of the critical behavior (e.g., the general forms of eqns [13]–[15]) will apply also to the quantum critical point of the section "Beyond LGW theory."

Following the canonical LGW strategy, we need to identify a collective order parameter which distinguishes the two phases. This is clearly given by the ferromagnetic moment in eqn [4] for the quantum Ising chain, and the antiferromagnetic moment in eqn [10] for the coupled dimer antiferromagnet. We coarse-grain these moments over some finite averaging region, and at long wavelengths this yields a real order parameter field ϕ_a, with the index $a = 1, \ldots, n$. For the Ising case we have $n = 1$ and ϕ_a is a measure of the local average of N_0 as defined in eqn [4]. For the antiferromagnet, a extends over the three values x, y, z (so $n = 3$), and three components of ϕ_a specify the magnitude and orientation of the local antiferromagnetic order in eqn [10]; note the average orientation of a specific spin at site j is η_j times the local value of ϕ_a.

The second step in the LGW approach is to write down a general field theory for the order parameter, consistent with all symmetries of the underlying model. As we are dealing with a quantum transition, the field theory has to extend over spacetime, with the temporal fluctuations representing the sum over histories in the Feynman path-integral approach. With this reasoning, the proposed partition function

for the vicinity of the critical point takes the following form:

$$\mathcal{Z}_\phi = \int \mathcal{D}\phi_a(x,\tau) \times \exp\left[-\int \mathrm{d}^d x\, \mathrm{d}\tau \left(\frac{1}{2}\left((\partial_\tau \phi_a)^2 + c^2(\nabla_x \phi_a)^2 + s\phi_a^2\right) + \frac{u}{4!}\left(\phi_a^2\right)^2\right)\right] \quad [11]$$

Here τ is imaginary time; there is an implied summation over the n values of the index a, c is a velocity, and s and $u > 0$ are coupling constants. This is a field theory in $d+1$ spacetime dimensions, in which the Ising chain corresponds to $d=1$ and the dimer antiferromagnet to $d=2$. The quantum phase transition is accessed by tuning the "mass" s: there is a quantum critical point at $s=s_c$ and the $s<s_c (s>s_c)$ regions correspond to the $g<g_c (g>g_c)$ regions of the lattice models. The $s<s_c$ phase has $\langle\phi_a\rangle \neq 0$ and this corresponds to the spontaneous breaking of spin rotation symmetry noted in eqns [4] and [10] for the lattice models. The $s>s_c$ phase is the paramagnet with $\langle\phi_a\rangle = 0$. The excitations in this phase can be understood as small harmonic oscillations of ϕ_a about the point (in field space) $\phi_a = 0$. A glance at eqn [11] shows that there are n such oscillators for each wave vector. These oscillators clearly constitute the $g>g_c$ quasiparticles found earlier in eqn [7] for the Ising chain (with $n=1$) and the triplon quasiparticle (with $n=3$) illustrated in **Figure 3** for the dimer antiferromagnet.

We have now seen that there is a perfect correspondence between the phases of the quantum field theory $\mathcal{Z}_\phi$ and those of the lattice models H_I and H_d. The power of the representation in eqn [11] is that it also allows us to get a simple description of the quantum critical point. In particular, readers may already have noticed that if we interpret the temporal direction τ in eqn [11] as another spatial direction, then $\mathcal{Z}_\phi$ is simply the classical partition function for a thermal phase transition in a ferromagnet in $d+1$ dimensions: this is the canonical model for which the LGW theory was originally developed. We can now take over standard results for this classical critical point, and obtain some useful predictions for the quantum critical point of $\mathcal{Z}_\phi$. It is useful to express these in terms of the dynamic susceptibility defined by

$$\chi(k,\omega) = \frac{\mathrm{i}}{\hbar}\int \mathrm{d}^d x \times \int_0^\infty \mathrm{d}t \left\langle \left[\hat{\phi}(x,t), \hat{\phi}(0,0)\right]\right\rangle_T \mathrm{e}^{-\mathrm{i}kx + \mathrm{i}\omega t} \quad [12]$$

Here $\hat{\phi}$ is the Heisenberg field operator corresponding to the path integral in eqn [11], the square brackets represent a commutator, and the angular brackets an average over the partition function at a temperature T. The structure of χ can be deduced from the knowledge that the quantum correlators of $\mathcal{Z}_\phi$ are related by analytic continuation in time to the corresponding correlators of the classical statistical mechanics problem in $d+1$ dimensions. The latter are known to diverge at the critical point as $\sim 1/p^{2-\eta}$ where p is the $(d+1)$-dimensional momentum, η is defined to be the anomalous dimension of the order parameter ($\eta = 1/4$ for the quantum Ising chain). Knowing this, we can deduce the form of the quantum correlator in eqn [12] at the zero-temperature quantum critical point

$$\chi(k,\omega) \sim \frac{1}{(c^2k^2 - \omega^2)^{1-\eta/2}}, \quad T=0, \quad g=g_c \quad [13]$$

The most important property of eqn [13] is the absence of a quasiparticle pole in the spectral density. Instead, $\mathrm{Im}(\chi(k,\omega))$ is nonzero for all $\omega > ck$, reflecting the presence of a continuum of critical excitations. Thus the stable quasiparticles found at low enough energies for all $g \neq g_c$ are absent at the quantum critical point.

We now briefly discuss the nature of the phase diagram for $T>0$ with g near g_c. In general, the interplay between quantum and thermal fluctuations near a quantum critical point can be quite complicated, and we cannot discuss it in any detail here. However, the physics of the quantum Ising chain is relatively simple, and also captures many key features found in more complex situations, and is summarized in **Figure 5**. For all $g \neq g_c$ there is a range of low temperatures ($T \lesssim |g - g_c|$) where the long time dynamics can be described using a dilute gas of thermally excited quasiparticles. Further, the

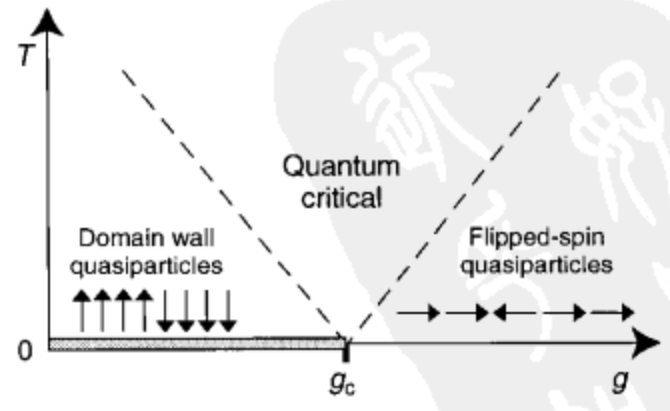


Figure 5 Nonzero temperature phase diagram of H_I. The ferromagnetic order is present only at $T=0$ on the shaded line with $g < g_c$. The dashed lines at finite T are crossovers out of the low-T quasiparticle regimes where a quasiclassical description applies. The state sketched on the paramagnetic side used the notation $|{\rightarrow}\rangle_j = 2^{-1/2}(|{\uparrow}\rangle_j + |{\downarrow}\rangle_j)$ and $|{\leftarrow}\rangle_j = 2^{-1/2}(|{\uparrow}\rangle_j - |{\downarrow}\rangle_j)$.

dynamics of these quasiparticles is quasiclassical, although we reiterate that the nature of the quasiparticles is entirely distinct on opposite sides of the quantum critical point. Most interesting, however, is the novel quantum critical region, $T \gtrsim |g - g_c|$, where neither quasiparticle picture nor a quasiclassical description are appropriate. Instead, we have to understand the influence of temperature on the critical continuum associated with eqn [13]. This is aided by scaling arguments which show that the only important frequency scale which characterizes the spectrum is $k_B T/\hbar$, and the crossovers near this scale are universal, that is, independent of specific microscopic details of the lattice Hamiltonian. Consequently, the zero-momentum dynamic susceptibility in the quantum critical region takes the following form at small frequencies:

$$\chi(k = 0, \omega) \sim \frac{1}{T^{2-\eta}} \frac{1}{(1 - i\omega/\Gamma_R)} \qquad [14]$$

This has the structure of the response of an overdamped oscillator, and the damping frequency, Γ_R, is given by the universal expression

$$\Gamma_R = \left(2 \tan \frac{\pi}{16}\right) \frac{k_B T}{\hbar} \qquad [15]$$

The numerical proportionality constant in eqn. [15] is specific to the quantum Ising chain; other models also obey eqn [15] but with a different numerical value for this constant.

Beyond LGW Theory

The quantum transitions discussed so far have turned to have a critical theory identical to that found for classical thermal transitions in $d+1$ dimensions. Over the last decade it has become clear that there are numerous models, of key physical importance, for which such a simple classical correspondence does not exist. In these models, quantum Berry phases are crucial in establishing the nature of the phases, and of the critical boundaries between them. In less technical terms, a signature of this subtlety is an important simplifying feature which was crucial in the analyses of the section "Simple models": both models had a straightforward $g \to \infty$ limit in which we were able to write down a simple, nondegenerate, ground-state wave function of the "disordered" paramagnet. In many other models, identification of the disordered phase is not as straightforward: specifying absence of a particular magnetic order is not enough to identify a quantum state, as we still need to write down a suitable wave function. Often, subtle quantum interference effects induce new types of order in the disordered state, and such effects are entirely absent in the LGW theory.

An important example of a system displaying such phenomena is the $S = 1/2$ square lattice antiferromagnet with additional frustrating interactions. The quantum degrees of freedom are identical to those of the coupled dimer antiferromagnet, but the Hamiltonian preserves the full point-group symmetry of the square lattice:

$$H_s = \sum_{j<k} J_{jk} \left(\hat{\sigma}_j^x \hat{\sigma}_k^x + \hat{\sigma}_j^y \hat{\sigma}_k^y + \hat{\sigma}_j^z \hat{\sigma}_k^z \right) + \cdots \qquad [16]$$

Here the $J_{jk} > 0$ are short-range exchange interactions which preserve the square lattice symmetry, and the ellipses represent possible further multiple spin terms. Now imagine tuning all the non-nearest-neighbor terms as a function of some generic coupling constant g. For small g, when H_s is nearly the square lattice antiferromagnet, the ground state has antiferromagnetic order as in **Figure 4** and eqn [10]. What is now the disordered ground state for large g? One natural candidate is the spin-singlet paramagnet in **Figure 2**. However, because all nearest neighbor bonds of the square lattice are now equivalent, the state in **Figure 2** is degenerate with three other states obtained by successive 90° rotations about a lattice site. In other words, the state in **Figure 2**, when transferred to the square lattice, breaks the symmetry of lattice rotations by 90°. Consequently it has a new type of order, often called valence-bond-solid (VBS) order. It is now believed that a large class of models like H_s do indeed exhibit a second-order quantum phase transition between the antiferromagnetic state and a VBS state – see **Figure 6**. Both the existence of VBS order in the paramagnet, and of a second-order quantum transition, are features that are not predicted by LGW theory: these can only be

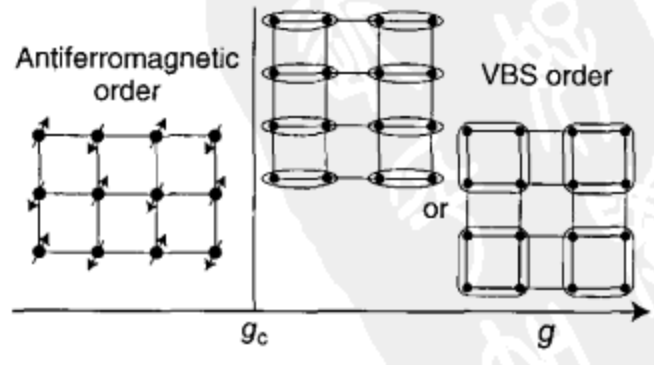


Figure 6 Phase diagram of H_s. Two possible VBS states are shown: one which is the analog of **Figure 2**, and the other in which spins form singlets in a plaquette pattern. Both VBS states have a 4-fold degeneracy due to breaking of square lattice symmetry. So the novel critical point at $g = g_c$ (described by Z_z) has the antiferromagnetic and VBS orders vanishing as it is approached from either side: this coincident vanishing of orders is generically forbidden in LGW theories.

understood by a careful study of quantum interference effects associated with Berry phases of spin fluctuations about the antiferromagnetic state. We will not enter into details of this analysis here, but will conclude our discussion by writing down the theory so obtained for the quantum critical point in **Figure 6**:

$$\mathcal{Z}_z = \int \mathcal{D}z_\alpha(x,\tau)\mathcal{D}A_\mu(x,\tau) \times \exp\left(-\int \mathrm{d}^2x\,\mathrm{d}\tau\left[|(\partial_\mu - \mathrm{i}A_\mu)z_\alpha|^2 + s|z_\alpha|^2 + \frac{u}{2}(|z_\alpha|^2)^2 + \frac{1}{2e^2}(\epsilon_{\mu\nu\lambda}\partial_\nu A_\lambda)^2\right]\right) \quad [17]$$

Here μ, ν, λ are spacetime indices which extend over the two spatial directions and τ, α is a spinor index which extends over $\uparrow$, $\downarrow$, and z_α is complex spinor field. In comparing $\mathcal{Z}_z$ to $\mathcal{Z}_\phi$, note that the vector order parameter ϕ_a has been replaced by a spinor z_α, and these are related by $\phi_a = z^*_\alpha \sigma^a_{\alpha\beta} z_\beta$, where σ^a are the Pauli matrices. So the order parameter has fractionalized into the z_α. A second novel property of $\mathcal{Z}_z$ is the presence of a U(1) gauge field A_μ: this gauge force emerges near the critical point, even though the underlying model in eqn [16] only has simple two spin interactions. Studies of fractionalized critical theories like $\mathcal{Z}_c$ in other models with spin and/or charge excitations is an exciting avenue for further theoretical research.

See also: Bose–Einstein Condensates; Boundary Conformal Field Theory; Fractional Quantum Hall Effect; Ginzburg–Landau Equation; High *Tc* Superconductor Theory; Quantum Central-Limit Theorems; Quantum Spin Systems; Quantum Statistical Mechanics: Overview.

Further Reading

Matsumoto M, Yasuda C, Todo S, and Takayama H (2002) *Physical Review* B 65: 014407.

Sachdev S (1999) *Quantum Phase Transitions.* Cambridge: Cambridge University Press.

Senthil T, Balents L, Sachdev S, Vishwanath A, and Fisher MPA, http://arxiv.org/abs/cond-mat/0312617.

Renormalization: Statistical Mechanics and Condensed Matter

M Salmhofer, Universität Leipzig, Leipzig, Germany

© 2006 Elsevier Ltd. All rights reserved.

Renormalization Group and Condensed Matter

Statistical mechanical systems at critical points exhibit scaling laws of order parameters, susceptibilities, and other observables. The exponents of these laws are universal, that is, independent of most details of the system. For example, the liquid–gas transition for real gases has the same exponents as the magnetization transition in the three-dimensional Ising model.

The renormalization group (RG) was developed by Kadanoff, Wilson, and Wegner, to understand these critical phenomena (Domb and Green 1976). The central idea is that the system becomes scale invariant at the critical point, which makes it natural to average over degrees of freedom on increasing length scales successively in the

Notes:

calculation of the partition function. This leads to a map between effective interactions associated to different length scales. Thus, the focus shifts from the analysis of a single interaction to that of a flow on a space of interactions. This space is in general much larger than the original formulation of the model would suggest: the description of long-distance or low-energy properties may be in terms of variables that were not even present in the original formulation of the system. Phenomenologically, this corresponds to the emergence of collective degrees of freedom.

Condensed matter theory is itself already an effective theory, and its "microscopic" formulation gets inputs from the underlying theories, which determine in particular the statistics of the particles and their interactions at the scale of atomic energies. At much lower-energy scales, which are relevant for low-temperature phenomena in condensed matter, collective excitations of different, sometimes exotic, statistics may emerge, but the starting point is given naturally in terms of fermionic and bosonic particles. For this reason, the discussion given below will be split in these two cases.

A major difference between high-energy and condensed matter systems is that the latter have a well-defined Hamiltonian which can be used to define the finite-volume ensembles of quantum statistical mechanics and which determines the time evolution, as well as various analyticity properties.

The relevant spatial dimensions in condensed matter are $d \leq 3$, but some results in higher dimensions relevant for the development of the method will also be discussed below. The cases $d=1$ and $d=2$ have always been of mathematical interest but in recent years have become important for the theory of new materials.

Some interesting topics cannot be covered here due to space restrictions, notably the application of renormalization methods to membrane theory (see Wiese (2001)) and renormalization methods for operators (see Bach *et al.* (1998)).

The Renormalization Group

In this section we briefly describe the setup of two important versions of the RG, namely the block spin RG and the RG based on scale decompositions of singular covariances.

Block spin RG

Let Λ be a finite lattice, for example, a finite subset of $\mathbb{Z}^d$. For the following, it is convenient to take Λ to be a cube of side-length L^K for $L > 1$ and some large K. Let T be a set and $\Phi_\Lambda = \{\phi : \Lambda \to T\}$ be the set of spin configurations. Common examples for the target space T are $T=\{-1,1\}$ for the Ising model, $T=S^{N-1}$ for the O(N) model, and $T=\mathbb{R}^n$ for unbounded spins. Let $S_\Lambda : \Phi_\Lambda \to \mathbb{R}, \phi \mapsto S_\Lambda(\phi)$ be an interaction and

$$Z(\Lambda, S_\Lambda) = \int \prod_{x\in\Lambda} \mathrm{d}\phi(x) \mathrm{e}^{-S_\Lambda(\phi)} \qquad [1]$$

In the unbounded case, S_Λ is assumed to grow sufficiently fast for $|\phi| \to \infty$, so that Z exists; for the case of a finite set T, the integral is replaced by a sum. Denote the corresponding Boltzmann factor by $\rho(\Lambda, S_\Lambda)$,

$$\rho(\Lambda, S_\Lambda)(\phi) = \frac{1}{Z(\Lambda, S_\Lambda)} \mathrm{e}^{-S_\Lambda(\phi)} \qquad [2]$$

The block spin transformation consists of an integration step and a rescaling step. Divide the lattice into cubic blocks of side-length L and define a new lattice Λ' by associating one lattice site of the new lattice to each L-block of the old lattice. For any $\phi' : \Lambda' \to T$, let

$$\rho'(\phi') = \int \prod_{x\in\Lambda} \mathrm{d}\phi(x) P(\phi', \phi) \mathrm{e}^{-S_\Lambda(\phi)} \qquad [3]$$

where $P(\phi', \phi) \geq 0$ and $\int \prod_{x'\in\Lambda'} \mathrm{d}\phi'(x') P(\phi', \phi) = 1$ for all ϕ, so that ρ' remains a probability distribution. Since ρ' is positive, one defines

$$S'_{\Lambda'}(\phi') = -\log \rho'(\phi') \qquad [4]$$

By construction, the partition function is invariant: $Z(\Lambda', S'_{\Lambda'}) = Z(\Lambda, S_\Lambda)$. The new lattice Λ' has spacing L; now rescale to make it a unit lattice. This completes the RG step in finite volume.

In an algorithmic sense, the "blocking rule" $P(\phi', \phi)$ can be viewed as a transition probability of a configuration ϕ to a configuration ϕ'. P may be deterministic, that is, simply fix ϕ' as a function of ϕ. From the intuition of averaging over local fluctuations, ϕ' is often taken to be some average of $\phi(x)$ at x in a block around x', hence the name.

Obviously, the thus defined RG transformation often cannot be iterated arbitrarily, since in every application, the number of points of the lattice shrinks by a factor L^d, so that after K iterations, a lattice with only a single point is left over. It is necessary to take the infinite-volume limit $L \to \infty$ to obtain a map that operates from a space to itself. However, [4] can become problematic in that limit: Gibbs measures ρ can map to measures ρ' whose large-deviation properties differ from those of Gibbs measures. The discussion of this problem and its solution is reviewed in Bricmont and Kupiainen (2001). The problem can be

solved in different ways, relaxing conditions on Gibbs measures or, in the Ising model, changing the description from the spins to the contours. The crucial point is that the difficulties arise only because [4] is applied globally, that is, to every ϕ'. The set of bad ϕ' has very small probability.

Block spin methods have been used in mathematical construction of quantum field theories, for example, in the work of Gawedzki and Kupiainen (1985) and Balaban (1988) (see the subsection "Field theory and statistical mechanics"). The above-mentioned problem was avoided there by not taking a logarithm in the so-called large-field region (which has very small probability).

Scale Decomposition RG

The generating functionals of quantum field theory and quantum statistical mechanics can be cast into the form

$$Z(C, V, \phi) = \int \mathrm{d}\mu_C(\phi')\, \mathrm{e}^{-V(\phi'+\phi)} \qquad [5]$$

Here $\mathrm{d}\mu_C$ denotes the Gaussian measure with covariance C, and V is the two-body interaction between the particles. The field variables are real or complex for bosons and Grassmann-valued for fermions. Differentiating $\log Z$ with respect to the external field ϕ generates the connected amputated correlation functions. The covariance determines the free propagation of particles; the interaction their collisions.

In most cases, such functional integrals are *a priori* ill-defined, even if V is small (and bounded from below) because the covariance C is singular. That is, the integral kernel $C(X, X')$ of the operator C either diverges as $|x - x'| \to 0$ (ultraviolet (UV) problem) or $C(X, X')$ has a slow decay as $|x - x'| \to \infty$ (infrared (IR) problem). In our notational convention, X may, in addition to the configuration variable x, also contain discrete indices of the fields, such as a spin or color index. The dependence of C on x and x' is assumed to be of the form $x - x'$. A typical example is the massless Gaussian field in d dimensions, where C is the inverse Fourier transform of $\hat{C}(k) = 1/k^2$, $k \in \mathbb{R}^d$, which has both a UV and an IR problem, or its lattice analog,

$$\hat{D}(k) = \left(\frac{2}{a^2} \sum_{i=1}^{d} (1 - \cos(ak_i)) \right)^{-1}$$

with a the lattice constant, which has only an IR problem. A typical interaction is of the type

$$V(\phi) = \int \mathrm{d}X\, \mathrm{d}Y \bar{\phi}(X) \phi(X) v(X, Y) \bar{\phi}(Y) \phi(Y) \qquad [6]$$

Again, we assume that the potential v depends on x and y only via $x - y$, so that translation invariance holds. In both UV and IR cases, naive perturbation theory fails even as a formal power series. That is, writing $V = \lambda V_0$, with a coupling constant λ which is treated as a formal expansion parameter, the singularity of C leads to termwise divergences in the series. The theory is called perturbatively renormalizable if all divergences can be removed by posing counter-terms of certain types, which are fixed by physically sensible renormalization conditions. Identifying the UV renormalizable theories was a breakthrough in high-energy physics. The IR renormalization problem is different, and in some respects harder, because there is almost no freedom to put counter-terms: the microscopic model is given from the start. This will be discussed in more detail below for an example.

A much more ambitious, and largely open, project is to do this renormalization nonperturbatively, that is, to treat λ as a real (typically, small) parameter. Some results will be discussed below.

The RG is set up by a scale decomposition $C = \sum_j C_j$. In the example of the massless Gaussian field, one would take each $\hat{C}_j$ to be a C^∞ function supported in the region $\{k \in \mathbb{R}^d : M^j \le k^2 \le M^{j+1}\}$, where $M > 1$ is a fixed constant, and the summation over j runs over $\mathbb{Z}$.

The scale decomposition of C leads to a representation of [5] by an iteration of Gaussian convolution integrals with covariances C_j, hence a sequence of effective interactions V_j, defined recursively by

$$\mathrm{e}^{-V_j(\phi)} = \int \mathrm{d}\mu_{C_{j+1}}(\phi')\, \mathrm{e}^{-V_{j+1}(\phi'+\phi)}, \quad V_0 = V \qquad [7]$$

For a singular covariance, the scale decomposition is an infinite sum. A formal object like [5] is now regularized by starting with a finite sum, that is, imposing a UV and IR cutoff, which is mathematically well defined, and then taking limits of the thus defined objects. Again, in condensed matter applications, imposing an IR cutoff is an operation that needs to be justified, for example, by showing that taking the limit as the cutoff is removed commutes with the infinite-volume limit.

Note that the RG map, which is the iteration $V_j \mapsto V_{j-1}$, goes to lower and lower j, corresponding to longer and longer length scales. The convention that the iteration starts at some fixed j, for example, $j = 0$, is appropriate for IR problems. In UV problems, the iteration would start at some large J_{UV}, which defines a UV cutoff and is taken to infinity, to remove the cutoff, at the end.

A variant using a continuous scale decomposition, $C = \int \mathrm{d}s \dot{C}_s$, originally due to Wegner and Houghton, became very popular after Polchinski (1984) used it

to give a short argument for perturbative renormalizability. Polchinski's equation, the analog of the recursion [7], reads

$$\frac{\partial V}{\partial s} = -\frac{1}{2} e^{V} \Delta_{\dot{C}_s} e^{-V} = \frac{1}{2} \Delta_{\dot{C}_s} V - \frac{1}{2} \left(\frac{\delta V}{\delta \phi}, \dot{C}_s \frac{\delta V}{\delta \phi} \right) \quad [8]$$

Here

$$\Delta_C = \left(\frac{\delta}{\delta \phi}, C \frac{\delta}{\delta \phi} \right)$$

denotes the Laplacian in field space associated to the covariance C. Polchinski's argument has been developed into a mathematical tool that applies to many models. For an introduction to perturbative renormalization using this method, see Salmhofer (1998). Equations of the type [8] have also been very useful beyond perturbation theory: much work has been done based on the beautiful representation of Mayer expansions found in Brydges and Kennedy (1987) using RG equations.

Mathematical Structure and Difficulties

The RG flow is thus, depending on the implementation, either a sequence or a continuous flow of interactions. Setting up this flow in mathematical terms is not easy and indeed part of the mathematical RG analysis is to find a suitable space of interactions that is left invariant by the successive convolutions, and then to control the RG iteration. A serious problem is the proliferation of interactions: already a single application of the RG transformation [7] maps a simple interaction, such as [6], to a nonlocal functional of the fields,

$$V_j(\phi) = \sum_{m \geq 0} \int dX_1 \cdots dX_m \times v_m^{(j)}(X_1, \ldots, X_m)\, \phi(X_1) \cdots \phi(X_m) \quad [9]$$

Already for perturbative renormalization, one needs to extract local terms, calculate their flow more explicitly, and control the power counting of the remainder. The convergence of the series is not an issue in formal perturbation theory because in every finite order r in λ, the sum over m is finite.

For nonperturbative renormalization, however, the problem is much more serious. For bosonic systems, the expansion in powers of the fields in [9] is divergent, and one needs a split into small-field and large-field regions and cluster expansions to obtain a well-defined sequence of effective actions (Gawedzki and Kupiainen 1985, Feldman *et al.* 1987, Rivassean 1993). That is, the local parts are extracted and treated explicitly only in the small-field region, and this is combined with estimates on the rareness of large-field regions using cluster expansions. For fermions, the expansion in powers of the fields can be proved to converge for regular, summable covariances, which leads to substantial technical simplifications.

The spatial proliferation of interactions is absent only in certain one-dimensional and in specially constructed higher-dimensional models, the so-called "hierarchical models." In these models, the search for an RG fixed point is still a nonlinear fixed-point problem, whose treatment leads to interesting mathematical results.

This article will be restricted to the mathematical use of the RG both in perturbative and nonperturbative quantum field theory of condensed matter systems. Many nonrigorous but very interesting applications have also come out of this method, showing that it also works well in practice, but they will not be reviewed here. Before discussing condensed matter systems, the pioneering works done on the mathematical RG, which were largely motivated by high-energy physics, will be reviewed briefly, as they laid the foundation of much of the technique used later in the condensed matter case.

Field Theory and Statistical Mechanics

Because of the close connection between quantum field theory and statistical mechanics given by formulas of the Feynman–Kac type, a significant amount of work on the mathematical RG focused on models of classical statistical mechanics in connection with field theories and gauge theories. Here we mention some of the pioneering results in that field.

The scale decomposition method was developed in a mathematical form and applied to perturbative UV renormalization of scalar field theories, as well as nonperturbative analysis of some models, by Gallavotti and Nicolò (Gallavotti 1985).

Infrared ϕ^4 theory in four dimensions was constructed using block spin methods (Gawedzki and Kupiainen 1985) and scale decomposition RG (Feldman *et al.* 1987). An essential feature of the ϕ^4_4 model is its IR asymptotic freedom, meaning that the local part of the effective quartic interaction tends to zero in the IR limit.

Block spin methods were used by Balaban (1988) to construct gauge theories in three and four dimensions. For gauge theories, the block spin RG has the major advantage that it allows to define a gauge-invariant RG flow. The scale decomposition violates gauge invariance, which creates substantial technical problems (Rivasseau 1993).

Condensed Matter: Fermions

Starting with the seminal work of Feldman and Trubowitz (1990, 1991) and Benfatto and Gallavotti (1995), this field has become one of the most successful applications of the mathematical RG. We use this example to discuss the scale decomposition method in a bit more detail.

We shall mainly focus on models in $d \geq 2$ dimensions (the case $d=1$ is described in detail in Benfatto and Gallavotti (1995)). The system is put into a finite (very large) box Λ of side-length L. For simplicity we take periodic boundary conditions. The Hilbert space for spin-1/2 electrons is the fermionic Fock space $\mathcal{F} = \bigoplus_{n\geq 0} \bigwedge^n L^2(\Lambda, \mathbb{C}^2)$. The grand canonical ensemble in finite volume is given by the density operator $\rho = Z^{-1} e^{-\beta(H-\mu N)}$, with the Hamiltonian H and the number operator N, in the usual second quantized form. The parameter $\beta = T^{-1}$ is the inverse temperature and the chemical potential μ is an auxiliary parameter used to fix the average particle number.

The grand canonical trace defining the ensemble can be rewritten in functional-integral form. It takes the form [5], but now $d\mu_C$ stands for a Grassmann Gaussian "measure," which is really only a linear functional (for definitions, see, e.g., Salmhofer (1998, chapter 4 and appendix B)). A two-body interaction corresponds to a quartic interaction polynomial V, as in [6]. The covariance is (in the infinite-volume limit $L \to \infty$)

$$C(\tau, x) = \frac{1}{\beta} \sum_{\omega \in \mathbb{M}_F} \int \frac{dk}{(2\pi)^d} e^{i(k\cdot x - \omega\tau)} \hat{C}(\omega, k)$$
$$\hat{C}(\omega, k) = \frac{1}{i\omega - e(k)} \qquad [10]$$

where $\tau \in (0, \beta]$ is a Euclidian time variable and k is the spatial momentum. The summation over ω runs over the set of fermionic Matsubara frequencies $\mathbb{M}_F = \pi T(2\mathbb{Z}+1)$. The function $e(k) = \varepsilon(k) - \mu$, where $\varepsilon(k)$ is the band function given by the single-particle term in the Hamiltonian. For a lattice system, $k \in \mathcal{B}_d$, the momentum space torus (e.g., for the lattice $\mathbb{Z}^d$, $\mathcal{B}_d = \mathbb{R}^d/2\pi\mathbb{Z}^d$); for a continuous system, $k \in \mathbb{R}^d$, hence there is a spatial UV problem. Electrons in a crystal have a natural spatial UV cutoff (see Salmhofer (1998, chapter 4) for a discussion) so we assume in the following that there is either a UV cutoff or that the system is on a lattice. A nonperturbative definition of the functional integral involves a limit from discrete times (by the Trotter product formula); see, for example, Salmhofer (1998) or Feldman *et al.* (2003, 2004).

Perturbative Renormalization

Renormalization of the Fermi surface at zero temperature In the limit $T \to 0$, the Matsubara frequency ω becomes a real variable, hence the propagator has a singularity at $\omega = 0$ and $k \in S$, where $S = \{k : e(k) = 0\}$, a codimension-1 subset of $\mathcal{B}_d$, is the Fermi surface. The existence of a Fermi surface which does not degenerate to a point is a characteristic feature of systems showing metallic behavior.

The singularity implies that $\hat{C} \notin L^p(\mathbb{R} \times \mathcal{B}_d)$ for any $p \geq 2$. Because terms of the type

$$\int d\omega \int dk F(\omega, k) \hat{C}(\omega, k) \times \prod_{i=1}^{p-1} \left(T_i(\omega, k) \hat{C}(\omega, k) \right) \qquad [11]$$

appear for all $p \geq 1$ in the formal perturbation expansion, with functions T_i and F that do not vanish on the singularity set of C, the perturbation expansion for observables is termwise divergent. The deeper reason for these problems is that the interaction shifts the Fermi surface so that the true propagator has a singularity of the form $G(\omega, k) = (i\omega - e(k) - \sigma(\omega, k))^{-1}$. If the self-energy σ is a sufficiently regular function, G has the same integrability properties as C, but the singularity of G is on the set $\tilde{S} = \{k : e(k) + \sigma(0, k) = 0\}$ (the singularity in ω remains at $\omega = 0$).

Let $1 = \sum_{j \leq 0} \chi_j(\omega, k)$ be a C^∞ partition of unity such that

$$\text{for } j < 0 \quad \text{supp } \chi_j \subset \{(\omega, k) : \epsilon_0 M^{j-2} \leq |i\omega - e(k)| \leq \epsilon_0 M^j\} \qquad [12]$$

where $M > 1$ and ϵ_0 is a fixed constant (an energy scale determined by the global properties of the function e; see Salmhofer (1998, chapter 4)). The corresponding covariances $\hat{C}_j = \hat{C}_{\chi_j}$ have the properties that for $j < 0$, $\|\hat{C}_j\|_1 \leq \text{const.} M^j$ and $\|\hat{C}_j\|_\infty \leq \text{const.} M^{-j}$. Using these bounds and expanding $v_m^{(j)} = \sum_{r \geq 1} v_{m,r}^{(j)} \lambda^r$, one can derive estimates for the coefficient functions $v_{m,r}^{(j)}$.

Of course, the scale decomposition by itself does not solve the problem of the moving singularity. It only allows us to pinpoint the problematic terms in the expansion. To construct the self-energy σ, as well as all higher Green functions, a two-step method is used (Feldman and Trubowitz 1990, 1991, Feldman *et al.* 1996, 2000). First, a counterterm function K which modifies e is introduced, so that all two-point insertions T_i get subtracted on the Fermi surface, hence replaced by $\tilde{T}_i(\omega, k) = T_i(\omega, k) - T_i(0, k')$, with k' obtained from k by a

projection to the Fermi surface (Feldman and Trubouitz 1990, 1991). Consequently, the $\tilde{T}_i$ vanish linearly on the Fermi surface, so that the integral over k in [11] converges. The effect of the counter-term function K can be described less technically: it fixes the Fermi surface to be S, the zero set of e. Thus, K forces S to be the Fermi surface of the interacting system. To achieve this, K must be chosen a function of e, k, and λV. In contrast to the situation for covariances with point singularities, the function K will, for a nontrivial Fermi surface, be very different from the original e. It can, however, be constructed to all orders in perturbation theory for a large class of Fermi surfaces. More precisely, one can prove: if $e \in C^2(\mathcal{B}_d, \mathbb{R})$, $\hat{v} \in \tilde{C}^2(\mathcal{B}_d, \mathbb{R})$, and the Fermi surface S contains no points k with $\nabla e(k) = 0$ and no flat sides, then $K = \sum_r \lambda^r K_r$ exists as a formal power series in λ and the map $e \mapsto e + K$ is locally injective on this set of e's (Feldman *et al.* 1996, 2000). With this counter-term, the order-r m-point functions on scale j satisfy the bounds

$$\left\| \hat{v}^{(j)}_{m,r} \right\|_\infty \leq w_{m,r} M^{(4-m)j/2} |j|^r$$

and

$$\left\| \hat{v}^{(j)}_{m,r} \right\|_1 \leq \tilde{w}_{m,r} \qquad [13]$$

with constants $w_{m,r}$ and $\tilde{w}_{m,r}$. Here $\hat{v}^{(j)}_{m,r}$ is the Fourier transform of $v^{(j)}_{m,r}$ (see [9], with the momentum conservation delta function from translation invariance removed.

Equation [13] implies that in the RG sense, the two-point function is relevant, the four-point function is marginal, and all higher m-point functions are irrelevant.

In one dimension, the Fermi "surface" reduces to two points which are related by a symmetry, so the counter-term function K is just a constant, that is, an adjustment of the chemical potential μ, which is justified because μ is only an auxiliary parameter used to fix the average value of the particle number. The counter-term function is a constant also in higher dimensions in the special case $e(k) = k^2 - \mu$: there, rotational symmetry implies that K can be chosen independent of k (if v is also rotationally symmetric). However, in the generic case of non-spherical Fermi surfaces, K depends nontrivially on k, and an inversion problem arises: adding the counter-term changes the model. To obtain the Green functions of a model with a given dispersion relation and interaction (E, V), one needs to show that given E in a suitable set, the equation

$$e(k) + K(\lambda, e, V)(k) = E(k) \qquad [14]$$

has a unique solution. If this is done, the procedure for renormalization is as follows. For a model given by dispersion relation and interaction (E, V), solve [14], then add and subtract e in the kinetic term. This automatically puts $K = E - e$ as a counter-term, and the expansion is now set up automatically with the right counter-term. The function K describes the shift from the Fermi surface of the free system (the zero set of E) to that of the interacting system (the zero set of e). Proving that K is sufficiently regular and solving [14] is nontrivial. Uniqueness of the solution follows from the above stated properties of K as a function of e. Existence was shown for a class of Fermi surfaces with strictly positive curvature in Feldman *et al.* (1996, 2000), to every order in perturbation theory. This implies a bijective relation between the Fermi surfaces of the free and the interacting model.

Positive temperature and the zero-limit temperature

One advantage of the functional-integral approach is that the setup at positive temperatures is identical to that at zero temperature, save for the discreteness of the set $\mathbb{M}_F$ at $T > 0$. Because $0 \notin \mathbb{M}_F$, the temperature effectively provides an IR cutoff, so that all term-by-term divergences are regularized in a natural way. However, renormalization is still necessary because the temperature is a physical parameter and unrenormalized expansions give disastrous bounds for the behavior of observables as functions of the temperature. Renormalization carries over essentially unchanged (the counter-term function is constructed slightly differently).

Because $|\omega| \geq \pi/\beta$ for all $\omega \in \mathbb{M}_F$, [12] implies supp $\chi_j = \emptyset$ for $j < -J_\beta$, where

$$J_\beta = \log_M \frac{\beta \epsilon_0}{\pi} \qquad [15]$$

Thus, the scale decomposition is now a finite sum over $0 \geq j \geq -J_\beta$. This restriction is inessential for the problem of renormalizing the Fermi surface, but it puts a cutoff on the marginal growth of the four-point function: [15] and [13] imply that

$$\| \hat{v}^{(j)}_{m,r} \|_1 \leq \tilde{w}_{m,r} \left(\log \frac{\beta \epsilon_0}{\pi} \right)^r \qquad [16]$$

If one can show that $\tilde{w}_{m,r} \leq AB^r$ with constants A and B, this implies that perturbation theory converges for $|\lambda| \log (\beta \epsilon_0 / \pi) < B^{-1}$. Such a bound has been shown using constructive methods (Disertori and Rivasseau 2000, Feldman *et al.* 2003, 2004) (see below). The logarithm of β is due to the Cooper instability (see Feldman and Trubowitz (1990, 1991) and Salmhofer (1998, section 4.5)).

The application of renormalization at positive temperature also led to the solution of a longstanding puzzle in solid-state physics, namely the (seeming) discontinuity of the results of perturbation theory as a function of the temperature claimed in the early literature. When renormalization is done correctly, there is no discontinuity in the temperature.

Nonperturbative Renormalization for Fermions

It is a remarkable feature of fermionic field theories that for a covariance for which $\|\hat{C}\|_1$ and $\|C\|_1$ are both finite, the effective action defined in [7] exists and is analytic in the fields and in the original interaction V, thanks to determinant bounds. For a V as in [6], with λv weak and of short range, the skeleton functions (where all relevant m-point functions are projected back to their initial values in the RG iteration) satisfy

$$\|\hat{v}_m^{(j)}\|_\infty \leq \text{const.} \|\hat{C}_j\|_1^{-(m/2)+1} \|C_j\|_1^{-1} \qquad [17]$$

For the many-electron covariance [10], with a positively curved C^d Fermi surface and with the scale decomposition [12], $\|\hat{C}_j\|_1$ is of order M^j and $\|C_j\|_1$ is of order $M^{-j(d+1)/2}$. The right-hand side of [17] then contains $M^{(d+3-m)j/2}$, which agrees (up to logarithms) with the perturbative power-counting bounds [13] only for $d=1$. In dimension $d=2$, the method has been refined by dividing the Fermi surface into angular sectors. The corresponding sectorized propagators have a better decay bound $\|C_j\|_1$, but the trade-off is sector sums at every vertex. Momentum conservation restricts these sector sums sufficiently in two dimensions to allow for good power-counting bounds. This has allowed for the construction of an interesting class of interacting fermionic models.

The major results obtained with the RG method are as follows.

Luttinger liquid behavior at zero temperature was proved for one-dimensional models with a repulsive interaction (Benfatto and Gallavotti 1995).

Fermi liquid behavior in the region where $|\lambda| \log(\beta\epsilon_0) \ll 1$ was proved for the two-dimensional model with $e(k)=k^2-1$, a local potential V, and a UV cutoff both on k and the Matsubara frequencies ω in Disertori and Rivasseau (2000).

A two-dimensional model with a band function $e(k)$ that is nonsymmetric under $k \to -k$ and a general short-range interaction was proved to be a Fermi liquid at zero temperature (Feldman *et al.* 2003, 2004). Due to the asymmetry under $k \to -k$, the Cooper instability can be proved to be absent. In Feldman *et al.* (2003, 2004), a counter-term function as in Feldman *et al.* (1996, 2000) was used. The nonperturbative proof of the corresponding inversion theorem remains open.

In $d=3$, the proof of Fermi liquid behavior remains an open problem, despite some partial results.

Condensed Matter: Bosons

Recent advances in quantum optice, in particular the trapping of ultracold atoms, have led to the experimental realization of Bose–Einstein condensation (BEC), which caused a surge of theoretical and mathematical works. For bosons, the definition of the ensembles is similar to, but more involved than in, the fermionic case. On a formal level, the functional-integral representation is analogous to fermions, except that the fields are not Grassmann fields but complex fields, and the covariance is given by a sum as in [10], but now the summation over ω runs over the bosonic Matsubara frequencies $M_B = 2\pi TZ$. The existence of even the free partition function in finite volume restricts the chemical potential (for free particles, $\mu < \inf_k \varepsilon(k)$ must hold). Note that C is complex and Gaussian measures with complex covariances exist in infinite dimensions only under rather restricted conditions, which are not satisfied by [10]. This is inessential for perturbative studies, where everything can be reduced to finite-dimensional integrals involving the covariance, but a nonperturbative definition of functional integrals for such systems requires again a carefully regularized (e.g., discrete-time) definition of the functional integral.

Bose–Einstein Condensation

The problem was treated to all orders in perturbation theory at positive particle density $\rho > 0$ by Benfatto (Benfatto and Gallavotti 1995). The initial interaction is again quartic, $\varepsilon(k)=k^2$, and one considers the problem at zero temperature, in the limit $\mu \to 0^-$, which is the limit in which BEC occurs for free particles. The interaction is expected to change the value of μ, given the density, so a chemical potential term is included in the action, to give the interaction

$$V(\phi) = \int d\tau dx dy |\phi(\tau,x)|^2 v(x-y) |\phi(\tau,y)|^2 + \nu \int d\tau dx |\phi(\tau,x)|^2 \qquad [18]$$

After writing $\phi(\tau,x) = \xi + \varphi(\tau,x)$, where ξ is independent of τ and x, the density condition becomes $\rho = |\xi|^2$. ν now needs to be chosen such that the free energy has a minimum at $\xi = \sqrt{\rho}$. This can be reformulated in terms of the self-energy of the boson.

Benfatto uses the RG to prove that the propagator of the interacting system no longer has the singularity structure $(i\omega - k^2)^{-1}$ but instead $(\omega^2 + c^2k^2)^{-1}$, where c is a constant. This requires a nontrivial analysis of Ward identities in the RG flow.

BEC has been proved in the Gross–Pitaevskii limit (Lieb *et al.* 2002). In the present formulation, this limit corresponds to an infinite-volume limit $L \to \infty$ where the density ρ is taken to zero as an inverse power of L. A nonperturbative proof of BEC at fixed positive particle density remains an open problem.

Superconductivity

Superconductivity (SC) occurs in fermionic systems, but it happens at energy scales where the relevant excitations have bosonic character: the Cooper pairs are bosons. In the RG framework, they arise naturally when the fermionic RG flow discussed above is stopped before it leaves the weak-coupling region and the dominant Cooper pairing term is rewritten by a Hubbard–Stratonovich transformation. The fermions can then be integrated over, resulting in the typical Mexican hat potential of an O(2) nonlinear sigma model. Effectively, one now has to deal with a problem similar to the one for BEC, but the action is considerably more complicated.

The Nonlinear Sigma Models

The prototypical model, into whose universality class both examples mentioned above fall, is that of O(N) nonlinear sigma models: both BEC and SC can be reformulated as spontaneous symmetry breaking (SSB) in the O(2) model in dimensions $d \geq 3$. For $d = 2$, long-range order is possible only at zero temperature because only then does the time direction truly represent a third dimension, preventing the Mermin–Wagner theorem from applying.

SSB has been proved for lattice O(N) models by reflection positivity and Gaussian domination methods (Fröhlich *et al.* 1976). The elegance and simplicity of this method is unsurpassed, but only very special actions satisfy reflection positivity, so that the method cannot be used for the effective actions obtained in condensed matter models. Results in the direction of proving SSB in O(N) models for $d \geq 3$ by RG methods, which apply to much more general actions, have been obtained by Balaban (1995).

See also: Bose–Einstein Condensates; Fermionic Systems; High T_c Superconductor Theory; Holomorphic Dynamics; Operator Product Expansion in Quantum Field Theory; Perturbative Renormalization Theory and BRST; Phase Transition Dynamics; Reflection Positivity and Phase Transitions.

Further Reading

Bach V, Fröhlich J, and Sigal IM (1998) Renormalization group analysis of spectral problems in quantum field theory. *Advances in Mathematics* 137: 205–298.

Balaban T (1988) Convergent renormalization expansions for lattice gauge theories. *Communications in Mathematical Physics* 119: 243–285.

Balaban T (1995) A low-temperature expansion for classical N-vector models I. A renormalization group flow. *Communications in Mathematical Physics* 167: 103–154.

Benfatto G and Gallavotti G (1995) *Renormalization Group*. Princeton: Princeton University Press.

Bricmont J and Kupiainen A (2001) Renormalizing the renormalization group pathologies. *Physics Reports* 348: 5–31.

Brydges DC and Kennedy T (1987) Mayer expansions and the Hamilton–Jacobi equation. *Journal of Statistical Physics* 48: 19–49.

Disertori M and Rivasseau V (2000) Interacting Fermi liquid in two dimensions at finite temperature I. Convergent contributions. *Communications in Mathematical Physics* 215: 251–290.

Domb C and Green M (eds.) (1976) *Phase Transitions and Critical Phenomena*, vol. 6. London: Academic Press.

Feldman J, Magnen J, Rivasseau V, and Sénéor R (1987) Construction of infrared ϕ_4^4 by a phase space expansion. *Communications in Mathematical Physics* 109: 437.

Feldman J and Trubowitz E (1990) Perturbation theory for many-fermion systems. *Helvetica Physica Acta* 63: 157.

Feldman J and Trubowitz E (1991) The flow of an electron-phonon system to the superconducting state. *Helvetica Physica Acta* 64: 213.

Feldman J, Salmhofer M, and Trubowitz E (1996) Perturbation theory around non-nested Fermi surfaces I. Keeping the Fermi surface fixed. *Journal of Statistical Physics* 84: 1209–1336.

Feldman J, Salmhofer M, and Trubowitz E (2000) An inversion theorem in Fermi surface theory. *Communications on Pure and Applied Mathematics* 53: 1350–1384.

Feldman J, Knörrer H, and Trubowitz E (2003) A class of Fermi liquids. *Reviews in Mathematical Physics* 15: 949–1169.

Feldman J, Knörrer H, and Trubowitz E (2004) *Communications in Mathematical Physics* 247: 1–319.

Fröhlich J, Simon B, and Spencer T (1976) Infrared bounds, phase transitions, and continuous symmetry breaking. *Communications in Mathematical Physics* 50: 79.

Gallavotti G (1985) Renormalization theory and ultraviolet stability via renormalization group methods. *Reviews of Modern Physics* 57: 471–569.

Gawedzki K and Kupiainen A (1985) Massless lattice ϕ_4^4 theory: Rigorous control of a renormalizable asymptotically free model. *Communications in Mathematical Physics* 99: 197–252.

Lieb E, Seiringer R, Solovej JP, and Yngvason J (2002) The ground state of the Bose gas. In: *Current Developments in Mathematics, 2001*, pp. 131–178. Cambridge: International Press.

Polchinski J (1984) Renormalization and effective Lagrangians. *Nuclear Physics B* 231: 269.

Rivasseau V (1993) *From Perturbative to Constructive Renormalization*. Princeton, NJ: Princeton University Press.

Salmhofer M (1998) *Renormalization: An Introduction*, Springer Texts and Monographs in Physics. Heidelberg: Springer.

Wiese KJ (2001) Polymerized membranes, a review. In: Domb C and Lebowitz J (eds.) *Phase Transitions and Critical Phenomena*, vol. 19. Academic Press.

Scattering, Asymptotic Completeness and Bound States

D Iagolnitzer, CEA/DSM/SPhT, CEA/Saclay, Gif-sur-Yvette, France
J Magnen, Ecole Polytechnique, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Relativistic quantum field theory (QFT) has been mainly developed since the 1950s in the perturbative framework. Quantities of interest then appear as infinite sums of Feynman integrals, corresponding to infinite series expansions with respect to couplings. This approach has led to basic successes for practical purposes, but suffered due to crucial defects from conceptual and mathematical viewpoints. First, individual terms were *a priori* infinite: this was solved by perturbative renormalization. However, even so, the series remain divergent. Two rigorous approaches have been developed since the 1960s. The axiomatic approach aims to establish a general framework independent of any particular model (Lagrangian interaction) and to analyze general properties that can be derived in that framework from basic principles. The "constructive" approach aims to rigorously establish the existence of nontrivial QFT models (theories) and to directly analyze their properties. Some of the fundamental bases are described in this encyclopedia in the articles by J Bros, D Buchholz and J Summers, and by G Gallavotti, respectively. This article aims to a deeper study of particle analysis and scattering of theories. In contrast to the articles by Buchholz and Summers and G Gallavotti, it is restricted to massive theories, a rather strong restriction, but for the latter goes much beyond in particle analysis.

From a purely physical viewpoint, results remain limited: the models rigorously defined so far are weakly coupled models in spacetime dimensions 2 or 3, results on bound states depend on specific kinematical factors in these dimensions, proofs of asymptotic completeness (AC) are not yet complete, ... On the positive side, we might say that the analysis and results are of interest from both conceptual and physical viewpoints; on the other hand, these works have also largely been related and have contributed to important, purely mathematical developments, for example, in the domain of analytic functions of several complex variables, microlocal analysis, ...

The general framework of QFT based on Wightman axioms is introduced in the next section. Massive theories are characterized in that framework by a condition on the mass spectrum. Haag–Ruelle asymptotic theory then allows one to define, in the Hilbert space $\mathcal{H}$ of states, two subspaces $\mathcal{H}_{\text{in}}$ and $\mathcal{H}_{\text{out}}$ corresponding to states that are asymptotically tangent, before and after interactions, respectively, to free-particle states. The AC condition $\mathcal{H}=\mathcal{H}_{\text{in}}=\mathcal{H}_{\text{out}}$ introduces a further important implicit particle content in the theory. Collision amplitudes or scattering functions are then well defined in the space of on-mass-shell initial and final energy–momenta (satisfying energy–momentum conservation). The LSZ "reduction formulas" give their link with chronological functions of the fields.

Basic properties of scattering amplitudes that follow from the Wightman axioms are then outlined. In particular, these axioms allow one to define the "N-point functions," which are analytic in a domain of complex energy–momentum space containing the Euclidean region (imaginary energy components), and from which chronological and scattering functions can be recovered. Other results at that stage include the on-shell physical sheet analyticity properties of four-point functions, as also general asymptotic causality and local analyticity properties for $N \geq 4$.

Next, we describe results derived from AC and regularity conditions on analyticity and asymptotic causality in terms of particles. In particular, the analysis of the links between analyticity properties of irreducible kernels (satisfying Bethe–Salpeter type equations) and AC in low-energy regions are included, following ideas of K Symanzik.

The final three sections are devoted to the analysis of models.

Models of QFT have been rigorously defined in Euclidean spacetime, through cluster and, more generally, phase-space expansions which are shown

Notes:

to be convergent at small coupling (and replace the nonconvergent expansions, of perturbative QFT). Examples of such models are the super-renormalizable massive φ^4 models in dimensions 2 or 3 (in the 1970s) and the "just renormalizable" massive (fermionic) Gross–Neveu model – in dimension 2 – in the 1980s. The N-point functions of these models can be shown to have exponential fall-off in Euclidean spacetime. By the usual Fourier–Laplace transform theorem, one obtains in turn analyticity properties in corresponding regions away from the Euclidean energy–momentum space.

On the other hand, *à la* Osterwalder–Schrader properties can be established in Euclidean spacetime. By analytic continuation from imaginary to real times, it is in turn shown that a corresponding nontrivial theory satisfying the Wightman axioms is recovered on the Minkowskian side. This analysis is omitted here. However, no information is obtained in that way on the mass spectrum, AC, energy–momentum space analyticity,.... Such results can be obtained through the use of irreducible kernels. This was initiated by T Spencer in the 1970s and then developed along the same line (Spencer and Zirilli, Dimock and Eckmann, Koch, Combescure, and Dunlop). We outline here the more general approach of the present authors. In the latter, irreducible kernels are directly defined through "higher-order" cluster expansions which are again convergent at sufficiently small coupling. They are shown to satisfy exponential fall-off in Euclidean spacetime with rates better than those of the N-point functions, and hence corresponding analyticity in larger regions around (and away from) the Euclidean energy–momentum space. Results will then be established by analytic continuation, from the Euclidean up to the Minkowskian energy–momentum space, of structure equations that express the N-point functions in terms of irreducible kernels. These structure equations are infinite series expansions, with again convergence properties at small coupling. In the cases $N=2$ and $N=4$ (even theories), the re-summation of these structure equations give, respectively, the Lippmann–Schwinger and Bethe–Salpeter (BS) integral equations (up to some regularization).

The one-particle irreducible (1PI) two-point kernel G_1 is analytic up to $s=(2m)^2-\varepsilon$, where ε is small at small coupling (s is the squared center of mass energy of the channel). A simple argument then allows one to show analyticity of the actual two-point function in the same region up to a pole at $k^2=m_{\mathrm{ph}}^2$: this shows the existence of a first basic physical mass m_{ph} (close at small coupling to the bare mass m). In a free theory (zero coupling) with one mass m, there is only one corresponding particle. At small coupling, the existence of other (stable) particles is not *a priori* expected; nevertheless, we will see that such particles (two-particle bound states) will occur in some models in view of kinematical threshold effects.

The 2PI four-point kernel G_2 is shown to be analytic up to $s=(4m)^2-\varepsilon$ in an even theory. On the other hand, it satisfies a (regularized) BS equation. In a way analogous to the section "AC and analyticity," starting here from the analyticity of G_2, the actual four-point function F is in turn analytic or meromorphic in that region up to the cut at $s\geq 4m^2$, and the discontinuity formula associated with AC in the low-energy region is obtained.

For some models (depending on the signs of some couplings), it will be shown that F has a pole in the physical sheet, below the two-particle threshold (at a distance from it which tends to zero as the coupling itself tends to zero). This pole then corresponds to a further stable particle.

More generally, and up to some technical problems, the structure equations should allow one to derive various discontinuity formulas of N-point functions including those associated with AC in increasingly higher-energy regions. Asymptotic causality in terms of particles and related analyticity properties (Landau singularities...) should also follow. However, in this approach, results should be obtained only for very small couplings as the energy region considered increases.

Note: Notations used are different in the next two sections on the one hand, and the final three sections on the other. These notations follow the use of, respectively, axiomatic and constructive field theory; for instance, x and p are real on the Minkowskian side in the next two sections whereas they are real on the Euclidean side in the last three sections. The mass m in the next two sections is a physical mass, whereas it is a bare mass in the last three sections (where a physical mass is noted m_{ph}).

The General Framework of Massive Field Theories

We denote by $x=(x_0,\boldsymbol{x})$ a (real) point in Minkowski spacetime with respective time and space components x_0 and $\boldsymbol{x}$ (in a given Lorentz frame); $x^2=x_0^2-\boldsymbol{x}^2$. Besides the usual spacetime dimension $d=4$, possible values 2 or 3 will also be considered. In all that follows, the unit system is such that the velocity c of light is equal to 1. Energy–momentum variables, dual (by Fourier transformation) to time and space

variables, respectively, are denoted by $p=(p_0,\mathbf{p})$; $p^2=p_0^2-\mathbf{p}^2$.

We describe below the Wightman axiomatic framework, though alternative ones such as "local quantum physics" based on the Araki–Haag–Kastler axioms may be used similarly for present purposes. For simplicity, unless otherwise stated, we consider a theory with only one basic (neutral, scalar) field A; A is defined on spacetime as an operator-valued distribution: for each test function f, $A(f)$ (formally $\int A(x)f(x)\mathrm{d}x$) is an operator in a Hilbert space $\mathcal{H}$ of states. A physical state is represented by a (normalized) vector in $\mathcal{H}$ modulo scalar multiples. It has to be physically understood as "sub specie aeternitatis" (i.e., "with all its evolution," the Heisenberg picture of quantum mechanics being always adopted). It is assumed that there exists in $\mathcal{H}$ a representation of the Poincaré group (semidirect product of pure Lorentz transformations and spacetime translations).

The Wightman axioms include:

1. local commutativity: $A(x)$ and $A(y)$ commute if $x-y$ is spacelike: $(x-y)^2<0$.
2. the spectral condition (= positivity of the energy in relativistic form): the spectrum of the energy–momentum operators (infinitesimal generators of spacetime translations) is contained in the cone $V_+(p^2\geq 0, p_0\geq 0)$. In a massive theory, the spectrum is more precisely assumed to be contained in the union of the origin (that will correspond to the vacuum vector introduced next), of one or more discrete mass-shell hyperboloids $H_+(m_i)(p^2=m_i^2, p_0>0)$ with strictly positive masses m_i, and of a continuum. For simplicity, and unless otherwise stated, we consider in this section a theory with only one mass m and a continuum starting at $2m$ (but this will not be so in a theory with "two-particle bound states"). This condition introduces a first (partial) particle content of the theory. In models, physical masses will not be introduced at the outset but will have to be determined.
3. existence in $\mathcal{H}$ of a vacuum vector Ω, which is the only invariant vector under Poincaré transformations up to scalar multiples; it is moreover assumed that the vector space generated by the action of field operators on the vacuum is dense in $\mathcal{H}$.
4. Poincaré covariance of the theory.

Subspaces $\mathcal{H}_{\text{in}}$ and $\mathcal{H}_{\text{out}}$ of $\mathcal{H}$ can be defined by limiting procedures. To that purpose, one considers test functions $f_{j,t}(\mathbf{x})$ with Fourier transforms of the form $\tilde{f}_j(\mathbf{p})\mathrm{e}^{\mathrm{i}(p_0-[\mathbf{p}^2+m^2]^{1/2})t}$, where the functions $\tilde{f}_j$ have their supports in a neighborhood of the mass-shell $H_+(m)$. It can then be shown that vectors of the form $\Psi_t=A(f_{1,t})A(f_{2,t})\cdots A(f_{n,t})\Omega$ converge to limits in $\mathcal{H}$ when $t\to\pm\infty$, respectively, and that these limits depend only on the mass-shell restrictions of the test functions $\tilde{f}_{j|H_+(m)}$.

$\mathcal{H}_{\text{in}}$ and $\mathcal{H}_{\text{out}}$ are interpreted physically as subspaces of states that are "asymptotically tangent" before, respectively, after the interactions, to free-particle states with particles of mass m. They are in fact both isomorphic to the free-particle Fock space $\mathcal{F}$, namely the direct sum of n-particle spaces of "wave functions" depending on n on mass-shell energy–momenta $p_1,p_2,\ldots,p_n$.

AC is the assertion that $\mathcal{H}=\mathcal{H}_{\text{in}}=\mathcal{H}_{\text{out}}$, that is, that each state in $\mathcal{H}$ is asymptotically tangent to a free-particle state, with particles of mass m, both before and after interactions (the two free-particle states are different if there are interactions). This condition cannot be expected to always hold in the general framework introduced above, even if we restrict our attention to "physically reasonable" theories in which states of $\mathcal{H}$ are asymptotically tangent to free-particle states before and after interactions: the absence of other stable particles with different masses is not guaranteed. For instance, even if A is "neutral," the action of field operators on the vacuum might generate pairs of "charged" particles with opposite charges, whatever "charge" one might imagine. Individual charged particles cannot occur in the neutral space $\mathcal{H}$ and their mass thus does not appear in the spectral condition. Hence, such states of pairs of charged particles will not belong to $\mathcal{H}_{\text{in}}$ or $\mathcal{H}_{\text{out}}$ although they belong to $\mathcal{H}$. However, if the set of charged particles is known, it can be shown that the above framework might be enlarged by defining charged fields, in such a way that AC might still be valid in the enlarged framework (see the article of Buchholz and Summers). For simplicity, we restrict below our attention to the simplest theories in which AC holds in the way stated above.

If AC holds, it is shown that there exists a linear operator S from $\mathcal{H}$ to $\mathcal{H}$, called "collision operator" or "S-matrix," that relates the "initial" and "final" free-particle states to which a state in $\mathcal{H}$ is tangent before and after interactions, respectively; if AC does not hold, S can also be defined as in operator in $\mathcal{F}$. Collision amplitudes or scattering functions are the energy–momentum kernels of S for given numbers m and n of initial and final particles. As easily seen, they are well-defined distributions on the space of all initial and final on-shell energy–momenta. For convenience, we will denote by p_k the physical energy–momentum of a final particle with index $k(p_k\in H_+(m))$, and by $-p_k$ the physical energy–momentum of an initial particle $(-p_k\in H_+(m))$.

Wightman Functions, Chronological Functions, and LSZ Reduction Formulas

The N-point Wightman "functions" W_N are defined as the vacuum expectation values (VEVs) of the products of N field operators, namely:

$$W_N(x_1, x_2, \ldots, x_N) = < \Omega, A(x_1)A(x_2)\cdots A(x_N)\Omega >$$

The chronological functions T_N are the VEVs of the chronological products of the fields $A(x_1), \ldots, A(x_N)$: in the latter, fields are ordered according to decreasing values of the time components of the points x_k. T_N is essentially well defined due to local commutativity with, however, problems not treated here at coinciding points.

$\tilde{T}_N(p_1, \ldots, p_N)$ will denote the Fourier transform of T_N. In view of the invariance of the theory under spacetime translations, functions above are invariant under global spacetime translation of all points x_k together. Hence, their Fourier transforms contain an energy–momentum conservation (e.m.c.) delta function $\delta(p_1 + p_2 + \cdots + p_N)$. Connected N-point functions are defined by induction (over N) via a formula expressing each (nonconnected) function as the sum of the corresponding connected function and of products of connected functions depending on subsets of points. In contrast to nonconnected functions, the analysis shows that connected functions in energy–momentum space do not contain in general e.m.c. delta functions involving subsets of energy–momenta.

It can be shown that the two-point function $\tilde{T}_2(p_1, p_2) = \delta(p_1 + p_2)\tilde{T}_2(p_1)$ has a pole of the form $1/(p_1^2 - m^2)$ and that $\tilde{T}_N$ has similar poles for each energy–momentum variable p_k on the mass-shell. The connected, amputated chronological function $\tilde{T}_N^{\text{amp, c}}$ is defined by multiplying $(\tilde{T}_N)_{\text{connected}} = \tilde{T}_N^{\text{c}}$ (for $N \geq 2$) by the product of all factors $p_k^2 - m^2$ that cancel these poles. It is then shown that it can be restricted as a distribution to the mass-shell of any physical process with m initial and n final particles, with $m + n = N$, and that this restriction coincides with the collision amplitude of the process. A process is here characterized by fixing the initial and final indices.

The analyticity properties of interest (described below) will apply to the connected functions after factoring out their global e.m.c. delta functions.

The Analytic N-point Functions

The Wightman axioms (without so far AC) yield general analyticity, as also asymptotic causality, properties that we now describe. The analysis is essentially based on the interplay of support properties in x-space arising from local commutativity and the definition of chronological operators, and support properties in p-space due to the spectral condition. Support properties in x-space apply to cell and more general "paracell" functions which are VEVs of adequate combinations of products of "partial" chronological operators. It is shown that each such function has support in x-space in a closed cone C_S (with apex at the origin). Moreover, for cell functions, the cone C_S is convex and salient. Hence, in view of the usual Laplace transform theorem, the cell function in p-space (after Fourier transformation) is the boundary value of a function analytic in complex space in the tube $\text{Re}\, p$ arbitrary, $\text{Im}\, p$ in the open dual cone $\tilde{C}_S$ of C_S. It is also shown that, near any real point $P = (P_1, \ldots, P_N)$, the chronological function in p-space coincides with one or more cell functions.

Together with support properties in p-space arising from the spectral condition and the use of coincidence relations between some cell functions (in adequate real regions in p-space), one then shows the existence, for each N, of a well-defined, unique analytic function F_N, called the "analytic N-point function," whose domain of analyticity, the "primitive domain of analyticity," in complex p-space contains all the tubes $\mathcal{T}_S$ associated with the cell functions. It also contains in particular a complex neighborhood of the Euclidean energy–momentum space which consists of energy momenta P_k with real $\boldsymbol{P}_k$ and imaginary energies $(P_k)_0$. Moreover, the chronological function $\tilde{T}_N^{\text{amp, c}}$ is the boundary value of F_N at all real points P, from imaginary directions which include those of the convex envelope of the cones $\tilde{C}_S$ associated with cell functions that coincide locally with $\tilde{T}_N^{\text{amp, c}}$.

However, the primitive domain has an empty intersection with the complex mass-shell, and thus gives no result on analyticity properties of collision amplitudes on the (real or complex) mass-shell. For $N = 4$, it has been possible to largely extend the primitive domain (which is not a "natural domain of holomorphy") by computing (parts of) its holomorphy envelope, which now has a nonempty intersection with the complex mass shell. It is shown in turn that the four-point function F_4 can be restricted to the complex mass-shell in a one-sheeted domain, called the "physical sheet," that admits each (real) physical region on its boundary (there is here one physical region for each choice of the two initial and the two final indices, the corresponding physical regions being disconnected from each other). In each physical region, the collision amplitude is the boundary value of the mass-shell restriction of F_4, from the corresponding half-space of "$+i\varepsilon$" directions $\text{Im}\ s > 0$, where s is the (squared) energy of the process.

The analyticity domain on the complex mass-shell contains paths of analytic continuation between the various physical regions ("crossing property") and admits cuts s_{ij} real $\geq (2m)^2$ covering the various physical regions. From these analyticity properties in the physical sheet, one can also derive "dispersion relations" (*see* Dispersion Relations).

Asymptotic causality and analyticity properties for $N \geq 4$

No similar result has been achieved at $N > 4$, and as a matter of fact, no similar result is expected if the AC condition is not assumed. The best results achieved so far are decompositions of the collision amplitude, in various parts of its physical region, as a sum of boundary values of functions analytic in domains of the complex mass-shell. In contrast to the case $N = 4$, the sum reduces to one term only in a certain subset of the physical region. Near other points, the N-point analytic function cannot be restricted locally to the complex mass-shell, though it can be decomposed as a sum of terms which, individually, are locally analytic in a larger domain that intersects the complex mass-shell.

These analyticity properties for $N \geq 4$ are a direct consequence of (and equivalent to) an asymptotic causality property that we now outline. Let $f_{k,\tau}(p)$ be, for each index k, a test function of the form

$$f_{k,\tau}(p) = e^{ip.\tau u_k} e^{-\gamma\tau|p_k - P_k|^2}$$

where each u_k is a point in spacetime, P_k is a given on-shell energy–momentum, and τ will be a spacetime dilatation parameter ($\gamma > 0$). It is well localized in p-space around the point P_k and its Fourier transform is well localized in x-space around the point τu_k up to an exponential fall-off of width $\sqrt{\gamma\tau}$ which is small compared to τ as $\tau \to \infty$.

We now consider the action of the (connected, amputated) chronological function on such test functions. A configuration $u = (u_1, \ldots, u_N)$ will be called "noncausal" at $P = (P_1, \ldots, P_N)$ if this action decays exponentially as $\tau \to \infty$. In mathematical terms, u is then outside the "essential support" or "microsupport" at P. The asymptotic causality property established, has roughly the following content: the only possible causal configurations u at P are those for which energy–momentum can be transferred from the initial to the final points in future cones. Moreover, at least two initial "extremal" points must coincide, as also two extremal final points. The simplest example is the case $N = 4$; if, for example, indices 1,2 are initial and 3,4 final, then the only *a priori* possible causal situations are such that $u_3 = u_4$ is in the future cone of $u_1 = u_2$ (in this particular case Lorentz invariance implies that $u_3 - u_1$ must be proportional to $P_3 + P_4$). In more general cases, the possible causal configurations u depend on P.

AC and Analyticity

Asymptotic Causality in Terms of Particles and Landau Singularities

As a matter of fact, a better causality property "in terms of particles" – which is the best possible one – is expected for "physically reasonable" theories if the (stable) particles of the theory are known. (By physically reasonable, we mean the absence of "à la Martin" pathologies such as the occurrence of an infinite number of unstable particles with arbitrary long lifetime). That property expresses the idea that the only causal configurations u at P are those for which the energy–momentum can be transferred from the initial to the final points via intermediate stable particles in accordance with classical laws: there should exist a classical connected multiple scattering diagram in spacetime joining the initial and final points u_k, with physical on-shell energy–momenta for each intermediate particle and energy–momentum conservation at each (pointwise) interaction vertex.

This property, if it holds, yields in turn (and is equivalent to) improved analyticity of the analytic N-point function near real physical regions: the (on-shell) collision amplitude is the boundary value of a unique analytic function in its physical region, at least away from some "exceptional points." The boundary value (namely the collision amplitude) is moreover analytic outside Landau surfaces $L_+(\Gamma)$ of connected multiple scattering graphs Γ; and along these surfaces (which are in general smooth codimension-1 surfaces), it is in general obtained from well-specified "$+i\varepsilon$" directions (that depend in general on the real point P of L_+).

Exceptional points are those that lie at the intersection of two (or several) surfaces $L_+(\Gamma_1)$, $L_+(\Gamma_2)\ldots$, with opposite causal directions, and hence having no $+i\varepsilon$ directions in common (in the on-shell framework). Such points do not occur at $N = 4$ for two-body processes, in which case the surfaces L_+ are the n-particle thresholds $s = (nm)^2$, with $n \geq 2, s = (p_1 + p_2)^2$. They do occur more generally: in a $3 \to 3$ process, 1,2,3 initial, 4,5,6 final, this is the case of all points P such that $-P_1 = P_4$, $-P_2 = P_5$, $-P_3 = P_6$ which all belong to the Landau surfaces of the two graphs Γ_1, Γ_2, with only one internal line joining two interaction

vertices: in the case of Γ_1, (resp., Γ_2), the first vertex involves the external particles 1, 2, 4 (resp., 1, 3, 5), while the second one involves 3, 5, 6 (resp., 2, 4, 6). If moreover P_1, P_2, P_3 lie in a common plane, previous points P also lie on surfaces L_+ of "triangle" graphs with again opposite causal directions at P. The fact that $+i\epsilon$ directions are opposite can equally be checked for the corresponding Feynman integrals of perturbative field theory.

Remark The above points are no longer exceptional in spacetime dimension 2. In fact, all surfaces L_+ mentioned then coincide with the (on-shell) codimension-1 surface $-p_1 = p_4, -p_2 = p_5, -p_3 = p_6$, with two opposite causal directions. The previous asymptotic causality property, together with a further "causal factorization" property for causal configurations, then yields along that surface an actual factorization of the three-body (nonconnected) S-matrix into a product of two-body scattering functions modulo an analytic background. The latter vanishes outside the surface, hence is identically zero, for some special two-dimensional models.

In the absence of the AC condition, one clearly sees why the above causality in terms of particles cannot be established: as we have seen, there is *a priori* no control on the stable particles of the theory and on their masses, and pathologies such as those mentioned above cannot be excluded. Hopefully, the first problem should be solved if AC is assumed, and the second one should be removed by adequate regularity assumptions. This is the purpose of the so-called axiomatic nonlinear program, in which one also wishes to examine further problems, for example, analytic continuation into unphysical sheets, with the occurrence of possible unstable particle poles and other singularities, nature of singularities, possible multiparticle dispersion relations, . . . , to cite only a few. Results so far remain limited but provide a first insight into such problems.

The Nonlinear Axiomatic Program

Results described below are based on discontinuity formulas arising from – and essentially equivalent in adequate energy regions to – AC, together with some regularity conditions. They can be established either with or without the introduction of adequate "irreducible" kernels. The methods rely on some general preliminary results on Fredholm theory in complex space (and with complex parameters). Irreducible kernels are defined through integral (Fredholm type) equations, first in the Euclidean region (imaginary energies) and then by local distortions of integration contours allowing one to reach the Minkowskian region. From discontinuity formulas and algebraic arguments, these irreducible kernels are shown to have analyticity (or meromorphy) properties associated with the physical idea of irreducibility (see examples below).

Results obtained so far with or without irreducible kernels are comparable in the simplest cases. However, the method based on irreducible kernels gives more refined results and seems best adapted to "extricate" the analytic structure of N-point functions for $N > 4$.

$N = 4$, Two-Body Processes in the Low-Energy Region

By even theory, we mean theories in which N-point function vanishes identically for N odd.

Standard results on two-body processes with initial (resp., final) energy–momenta p_1, p_2 (resp., p'_1, p'_2) in the low-energy region $(2m)^2 \leq s < (3m)^2$ $(s = (p_1 + p_2)^2 = (p'_1 + p'_2)^2)$ are based on the "off-shell unitarity equation"

$$F_+ - F_- = F_+ \star F_- \qquad [1]$$

where $F_+(p_1, p_2; p'_1, p'_2)$ and $F_-(p_1, p_2; p'_1, p'_2)$ denote, respectively, the $+i\varepsilon$ and $-i\varepsilon$ boundary values of the four-point function F_4 from above or below the cut $s \geq (2m)^2$ in the physical sheet, and $\star$ denotes on-shell convolution over two intermediate energy–momenta. This relation is a direct consequence of AC for s less than $(3m)^2$, or less than $(4m)^2$ in an even theory. When the four external energy–momentum vectors p_1, p_2, p'_1, p'_2 are put on the mass shell (on both sides of that relation), one recovers the usual elastic unitarity relation for the collision amplitude T_+ and its complex conjugate T_-:

$$T_+ - T_- = T_+ \star T_-$$

In the exploitation of these relations outlined below, a regularity condition is moreover needed, for example, the continuity of F_+ in the low-energy region.

By considering the unitarity equation as a Fredholm equation for T_+ at fixed s (in the complex mass shell), one obtains the following result: T_+ can be analytically continued as a meromorphic function of s through the cut (in the low-energy region) in a two-sheeted (d even) or multisheeted (d odd) domain around the two-particle threshold. Possible poles in the second sheet (generated by Fredholm theory) will correspond physically to unstable particles. The singularity at the two-particle threshold is of the square-root type in s for d even, or in

$1/\log s$ for d odd. The difference between the two cases is due to the power $(d-1)/2$ of s, integer or half-integer, in the kinematical factor arising from on-shell convolution. This result can also be extended to the off-shell function F_4 by applying a further argument of analytic continuation making use of the off-shell unitarity equation.

Restricting now our attention to an even theory (for simplicity), a similar result also follows from the introduction of a two 2PI BS type kernel G satisfying (and here defined from F through) a regularized BS equation of the form

$$F = G + F \circ_M G \qquad [2]$$

where $\circ_M$ denotes convolution over two intermediate energy–momenta with two-point functions on the internal lines and a regularization factor in order to avoid convergence problems at infinity (G then depends on the choice of this factor but its properties and the subsequent analysis do not). Alternatively, one may also introduce a kernel satisfying a renormalized BS equation, but this is not useful for present purposes.

Starting from the above discontinuity formula [1], one shows in turn that G is indeed "2PI" in the analytic sense:

$$G_+ = G_- \qquad [3]$$

in the low-energy region. More precisely, G is analytic or meromorphic (with poles that may arise from Fredholm theory) in a domain that includes the two-particle threshold $s=(2m)^2$, in contrast to F itself.

The proof of [3] is based on the relation independent of M (and thus leaving the M dependence implicit).

$$\circ_+ - \circ_- = \star \qquad [4]$$

(which is a nontrivial adaptation of the decomposition of a mass-shell delta function as a sum of plus and minus $i\varepsilon$ poles). A simple algebraic argument then shows essentially the equivalence between the discontinuity formulas [1] and [3].

In turn, assuming that G has no poles, this analyticity allows one to recover the two-sheetedness (d even) or multisheetedness (d odd, singularity in 1/log) of F, in view of the BS type equation.

$N=6$, 3–3 Process in the Low-Energy Region (Even Theory)

The result, in the neighborhood of the 3–3 physical region, is here a "structure equation" expressing the 3–3 function F in the low-energy region as a sum of "à la Feynman contributions" associated with graphs with one internal line and with triangle graphs, with two-point functions on internal lines and four-point functions at each vertex, plus a remainder R. The latter is shown to be a boundary value from $+i\varepsilon$ directions Im s positive, where $s=(p_1+p_2+p_3)^2, p_1, p_2, p_3$ denoting the energy–momentum vectors of the initial particles. Further regularity conditions are needed to recover its local physical region analyticity. The various explicit contributions that we have just mentioned yield the actual physical region Landau singularities expected in the low-energy 3–3 physical region.

A more refined result, in the approach based on irreducible kernels outlined below, applies in a larger region and then includes further à la Feynman contributions associated with 2-loop and 3-loop diagrams (the latter do not contribute to "effective" singularities in the neighborhood of the physical region).

The first result can be established from discontinuity formulas for the three-point function around two-particle thresholds, arising from AC, and "microsupport" analysis of all terms involved. In the approach based on irreducible kernels, it is useful to introduce in particular a 3PI kernel G_3 that, in contrast to the 3–3 function, will be analytic or meromorphic in a domain including the three-particle threshold. To that purpose, an adequate set of integral equations is introduced and the three-particle irreducibility of G_3 in "the analytic sense" is then established. In turn it provides the complete structure equation mentioned above.

More General Analysis

There are so far only preliminary steps in more general situations, in view of (difficult) technical problems involved and the need of *ad hoc* regularity assumption at each stage. As already mentioned, the approach based on irreducible kernels seems best adapted. The analysis should clearly involve more general irreducible kernels with various irreducibility properties with respect to various channels (and not only with respect to the basic channel considered such as the 3–3 channel in the case above). From a heuristic viewpoint, one may first consider to that purpose adequate formal expansions into (infinite) sums of "à la Feynman contributions" adapted to the energy regions under investigation. These à la Feynman contributions will involve adequate irreducible kernels in the graphical sense at each vertex, and the above expansions correspond formally to the best possible regroupings of Feynman integrals with respect to the energy region considered. From such expansions, one might

determine adequate sets of integral equations allowing one, together with regularity assumptions, to carry out an analysis similar to above.

The Models

A Euclidean field-theoretical model can be defined by a probability measure $d\mu(\varphi)$ on the space of tempered distributions φ in Euclidean spacetime, whose moments verify the Osterwalder–Schrader (or similar) axioms. The moments of $d\mu$ are, for each N, the Euclidean (Schwinger) N-point functions:

$$S(x_1, \ldots, x_N) = \int \varphi(x_1) \cdots \varphi(x_N)\, d\mu(\varphi) \qquad [5]$$

In what follows, the measure $d\mu$ will be a perturbed Gaussian measure which, for the massive φ^4 model with a volume cutoff Λ and an ultraviolet cutoff ρ, is given in d dimensions by

$$d\mu_{\Lambda,\rho} = e^{-\lambda(\rho)\int_\Lambda \varphi^4(z)\,dz + a(\rho)\int_\Lambda \varphi^2(z)dz}\, d\nu_\rho(\varphi)/Z_{\Lambda,\rho} \qquad [6]$$

where $Z_{\Lambda,\rho}$ is the normalization factor and where $d\nu_\rho(\varphi)$ is the Gaussian measure of mean zero ($\int \varphi\, d\nu = 0$) and covariance

$$C(x-y;\rho) = \int d^d p\, e^{ip(x-y)}\, e^{-p^2/\rho^2}/(\zeta(\rho)p^2 + m^2)$$

where by convention m is called the bare mass.

For $d=2$ or 3 one can show that, for $\lambda(\rho)=\lambda$ small enough (depending on m) and $\zeta(\rho)=1$, there exists a function $a(\rho)$ ($a(\rho)=O(\lambda)$ as $\lambda \to 0$) such that, for any set of N distinct points, the function $S(x_1, \ldots, x_N) = \lim_{\Lambda,\rho\to\infty} S_{\Lambda,\rho}(x_1, \ldots, x_N)$ exists, is not Gaussian (hence does not correspond to a trivial, free theory), and satisfies the Osterwalder–Schrader axioms. The connected part $S(x_1, \ldots, x_N)_{\text{connected}}$ has the following perturbative series:

$$\lim_{\Lambda,\rho\to\infty} \sum_n \frac{(-1)^n}{n!} \int \varphi(x_1) \ldots \varphi(x_N) \times \left[\int_\Lambda [\lambda\varphi^4 - a(\rho)\varphi^2](z)\, dz\right]^n d\nu_{\Lambda,\rho}(\varphi)\Big|_{\text{connected}} \qquad [7]$$

which is the (divergent) sum of the connected renormalized (Euclidean) Feynman graphs.

The study of the perturbative series leads to the distinction of:

1. the super-renormalizable theories, where it is possible to take $\lambda(\rho), \zeta(\rho)$ not depending on ρ. In dimension 2, all the models where $\lambda\varphi^4$ is replaced by
$$c_{2p}\varphi^{2p} + c_{2p-1}\varphi^{2p-1} + \cdots + c_5\varphi^5 + \lambda\varphi^4 + c_3\varphi^3 \qquad [8]$$
also exist provided that $c_{2p} > 0$ is small enough depending on m and on the other coefficient c's and λ, and
2. the just renormalizable theories where $\lambda(\rho)$ (and possibly $\zeta(\rho)$) depend in general on ρ. In models mentioned below $\lambda(\rho) \to 0$ as $\rho \to \infty$; this characterizes "asymptotic freedom."

The proof of the existence of the N-point functions makes use of Taylor type expansions with remainder. The first orders are used to compute $\lambda(\rho), \zeta(\rho), a(\rho)$. The idea is to consider the functional integral [5] – at Λ, ρ finite – as an integral over roughly $\Lambda\rho^d$ "degrees of freedom" which are weakly coupled. This corresponds to a decomposition of the phase space (with cutoff both in x-space (the box Λ) and in p-space (roughly $|p| < \rho$)). The coupling between different regions in x-space comes from the propagators C_ρ; the coupling between different frequencies in p-space comes from the φ^4 term (the interaction vertex). The expansion is then, for each degree of freedom, a finite expansion in the coupling between this degree and the others so that, even if the expansion is perturbative up to the order $\Lambda\rho^d$, the bound on each term is qualitatively the one on a product of $\Lambda\rho^d$ finite order-independent expansions, the order of which can be fixed uniformly in ρ (and depending only on λ). To achieve this program, the propagator linking two points of distance of order L must have a decrease of order $e^{-L^{-1}|x-y|}$, that is, have momentum larger than L^{-1}, so that one must localize both in x-space and p-space ; for example, the smallest cells of phase space correspond to fields φ localized in x, p-spaces, the x-boxes being of side ρ^{-1} and the p-localization consisting of values such that roughly $(\rho/2) \le |p| \le \rho$. More generally, a generic cell (of index i) corresponds to fields φ at point x and momentum p, with x in a box of side $2^i\rho^{-1}$ and $2^{-i-1}\rho < |p| < 2^{-i}\rho$.

These expansions are mimicking the à la Wilson renormalization group. For just renormalizable theories (where $\lambda(\rho)$ depends on ρ), one is led to introduce the effective coupling constant $\lambda(2^{-i}\rho)$ whose perturbative expansion is the value at momentum zero of the sum of all the (connected, amputated) four-point functions containing only propagators of momentum (roughly) bigger than $2^{-i}\rho$ (plus $\lambda(\rho)$ which in fact tends to zero as $\rho \to \infty$).

Then by small coupling we mean a theory where $\lambda(2^{-i}\rho)/\zeta(2^{-i}\rho)^2$ is small for all i.

By convention we write $\lambda_{\text{ren}}, \zeta_{\text{ren}}, a_{\text{ren}}$ for the effective parameters of the theory at zero momentum.

The expansion obtained expresses $S_{\text{connected}}$ as a sum of terms each of them being associated to a

given set of phase-space cells which are "connected" together by "links" that are either propagators or vertices. Each term decreases exponentially with the difference $i_{\max} - i_{\min}$ of the upper and lower indices of the phase-space cells involved. Moreover, each set must contain the cells associated to the fields $\varphi(x_1) \cdots \varphi(x_N)$ whose indices are fixed by the order of magnitude of the distances between the points. On the other hand, the difference between the theory of cutoff ρ and the one of cutoff 2ρ are terms containing at least one cell of momentum of order ρ; these terms are thus small like $\mathrm{cst}(x_1, \ldots, x_N)\mathrm{e}^{-(\mathrm{cst})\rho}$, so that the limit as $\rho \to \infty$ exists.

So far, the "construction" of models is possible only at small coupling, apart from special cases. The φ^4 theory in dimension 4 is just renormalizable (from the perturbative viewpoint) but the above condition of small coupling cannot be achieved (and it is generally believed that this model cannot be defined as a nontrivial theory). A just renormalizable model has been shown to exist, namely the Gross–Neveu model which is a fermionic theory in dimension 2. The elementary particle physics models are just renormalizable but their construction has not been completed so far (in particular in view of the confinement problem). *See* Constructive Quantum Field Theory for details.

To state the result in a form convenient for our purposes here, we introduce a splitting of the covariance in two parts:

$$C(x-y;\rho) = C_M(x-y;\rho) + C_{>M}(x-y;\rho), \quad M > m$$

$$\tilde{C}_M(p;\rho) = (\mathrm{e}^{-p^2/\rho^2}/p^2 + m^2) - (\mathrm{e}^{-p^2/\rho^2}/p^2 + M^2)$$

so that $C_M(x-y)$ behaves like C at large distances but has an ultraviolet cutoff of size M, and $|C_{>M}(x-y)| \leq \mathrm{e}^{-M|x-y|}$ decreases exponentially depending on the (technical) choice of M. Let $\mathrm{d}\nu_M(\varphi)$ be the Gaussian measure of covariance C_M.

One divides also Λ in unit cubes and obtains for the connected N-point function an expansion as a sum over connected trees; a tree T is composed of lines ℓ and vertices v; each line joins two vertices or one of the external points $x_1, \ldots, x_N$ and a vertex; moreover, there are no loops.

To each line ℓ is associated a propagator $C_M(z_\ell, z'_\ell) = C_M(\ell)$.

To each vertex v are associated:

1. two subsets I_v, I'_v of $\{\ell\}$,
2. a connected set X_v of unit cubes such that all the $z_\ell, \ell \in I_v$ and all the $z'_\ell, \ell \in I'_v$ are contained in X_v; $|X_v|$ is the volume of X_v, and
3. a kernel $K_{X_v}(\{z, z'\}_v; \varphi)$

Finally, the external points are by convention z_ℓ points; then:

$$\begin{aligned} S_{\Lambda,\rho}(x_1, \ldots, x_N)_{\mathrm{connected}} &= \int \mathrm{d}\nu_M(\varphi) \sum_T \frac{1}{|T|!} \sum_{\substack{\{X_v\} \\ \text{nonoverlapping}}} \\ &\times \int \Big[\prod_{\substack{\ell \\ z_\ell \text{ not external}}} \mathrm{d}z_\ell \Big] \Big[\prod_{\ell \in T} \mathrm{d}z'_\ell C_M(\ell) \Big] \\ &\times \prod_{v \in T} K_{X_v}(\{z, z'\}_v; \varphi) \end{aligned} \quad [9]$$

where for coupling small enough:

$$\int \mathrm{d}\nu_M(\varphi) \prod_{v \in T} |K_{X_v}(\{z, z'\}_v; \varphi)| \leq \prod_{v \in T} \mathrm{e}^{-M(1-\epsilon)|X_v|} \quad [10]$$

The X's are 2×2 nonoverlapping; however, it will suffice to sum over all X's (without restriction) to get a bound showing the convergence of the expansion as $\Lambda \to \infty$. In this formula the $K(., \varphi)$'s are still coupled by the measure $\mathrm{d}\nu_M(\varphi)$; all the nonperturbativity is hidden in the K's (in particular the contribution of momentum bigger than M).

As a consequence of [9] and if $a(\rho, \lambda)$ has been chosen such that $a_{\mathrm{ren}} = 0$, for M large enough and at small coupling (depending on M, m):

$$\begin{aligned} &|S(x,y)_{\mathrm{connected}}| \\ &\quad \leq |C_M(x-y)| + \int \mathrm{d}z'_1 \mathrm{d}z'_2 |C_M(x-z'_1) \\ &\qquad \times \mathrm{e}^{-M(1-\epsilon)|z'_1 - z'_2|} C_M(z'_2 - y)| + \cdots \\ &\quad \leq (\mathrm{cst})\mathrm{e}^{-m(1-\epsilon)|x-y|} \end{aligned} \quad [11]$$

More generally, the connected N-point function satisfies

$$|S(x_1, \ldots, x_N)_{\mathrm{connected}}| \leq \mathrm{cst}\ \mathrm{e}^{-m(1-\epsilon)d(x_1, \ldots, x_N)} \quad [12]$$

where $d(x_1, \ldots, x_N)$ is the length of the smallest tree joining $x_1, \ldots, x_N$, with possibly intermediate points.

The Irreducible Kernels

The 1PI Kernel and a Lippmann–Schwinger Equation

To then show that a theory – if the perturbation series heuristically shows it – contains only one particle of mass smaller than $2m(1-\epsilon)$, it is necessary to expand further the coupling between the K's in [9]. Each perturbative step relatively to this coupling will generate a sum of terms such that in each one there is a "new" propagator C_M between two K's.

The fact that in [9] the X's are nonoverlapping has the consequence that an expansion where for each pair of K_X the number of propagators C_M

remains bounded (say by $n+1$) is convergent (for small enough couplings depending on m, n); this is because, for a given X, the others must be farther and farther as their number increases, and in view of the exponential decrease (in x-space) of C_M.

We then consider the expansion where we have further expanded the two-point function $S(x,y)$ such that each term can be decomposed in the channel $x \to y$ in C_M propagators and 1PI contributions (in the sense that any line cutting such a 1PI contribution (and outside the $X's$) cuts at least two propagators); that means that these 1PI contributions are no longer coupled by the $d\nu_M(\varphi)$ measure. They are made of propagators and of K_X which still have nonoverlapping restrictions; the latter are straightforwardly expanded using a kind of (convergent) Mayer expansion; the result is finally a Lippmann–Schwinger type equation:

$$S(x,y)_{\text{connected}} = C_M(x-y) + \int dz_1\, dz_2\, C_M(x-z_1) \times G_1(z_1,z_2)\, C_M(z_2-y) + \cdots \quad [13]$$

or

$$S(x,y)_{\text{connected}} = \left[C_M \sum_{p\geq 0} [G_1 C_M]^p \right](x,y)$$

which is equivalent to

$$S_{\text{connected}} = C_M + C_M G_1 C_M + C_M G_1 S_{\text{connected}} \quad [14]$$

where G_1 is a 1PI kernel that satisfies the bound

$$|G_1(t,u)| \leq \lambda_{\text{ren}} e^{-2m(1-\epsilon)|t-u|} \quad [15]$$

In Fourier transform, eqn [14] becomes

$$F(p) = \tilde{C}_M(p) + \tilde{C}_M(p)\tilde{G}_1(p)\tilde{C}_M(p) + \tilde{C}_M(p)G_1(p)F(p) \quad [16]$$

Denoting by $\delta(p+q)\,F(p,q)$ the Fourier transform of $S(x,y)_{\text{connected}}$, we can then compute $F(p)$:

$$F(p) = \frac{(p^2+m^2)[\tilde{C}_M + \tilde{C}_M\tilde{G}_1](p)}{(p^2+m^2) - (p^2+m^2)\tilde{C}_M\tilde{G}_1(p)} \quad [17]$$

where $(p^2+m^2)\tilde{C}_M(p) \to (1-m^2/M^2)$ as $p \to 0$ and $|\tilde{G}_1(p)| \leq \lambda_{\text{ren}}\,\text{cst}(m)$ so that (as expected) F has no pole in the Euclidean region at small coupling; but, as will be seen in the next section, it has a pole outside the Euclidean region.

The 2PI Kernel and a BS Equation

From the previous discussion, it is clear that one can extract from [9] as many propagators as we want between kernels K_X. If one considers a splitting of the external points in incoming $x_1,\ldots,x_p$ and outgoing $x_{p+1},\ldots,x_N$ points, this defines a channel. One then obtains nPI kernels (in the given channel). In the same way as above, one obtains a relevant structure equation; this equation makes sense only if the kernels K_X have a decrease corresponding to n-particle irreducibility; to that purpose we take $M > nm$. The expansion converges for couplings small enough depending on m and n.

In the case $n=2$ this gives a kind of BS equation (the Lippmann–Schwinger equation corresponding to the case $n=1$); if we restrict, for simplicity, the analysis to even theories one is led to jump directly to the case $n=3$:

$$S(x_1,x_2;x_3,x_4)_{\text{connected}} = \int dz_1\, dt_1\, dz_2\, dt_2 (\circ_M)(x_1,x_2;z_1,t_1) \times G_2(z_1,t_1;z_2,t_2)(\circ_M)(z_2,t_2;x_3,x_4) + \cdots \quad [18]$$

$$S = \circ_M \sum_{p\geq 1} [G_2 \circ_M]^p$$

or

$$S = \circ_M G_2 \circ_M + \circ_M G_2 S \quad [19]$$

where

$$(\circ_M)(x_1,x_2;x_3,x_4) = S(x_1,x_3)S(x_2,x_4) + S(x_1,x_4)S(x_2,x_3)$$

and where

$$|G_2(t_1,t_2;u_1,u_2)| \leq \lambda_{\text{ren}} \exp\{-4m(1-\epsilon)\max_{i,j}(|t_i-u_j|)\} \quad [20]$$

Equation [19] once amputated, and after Fourier transformation, is eqn [2].

More General Irreducible Kernels and Structure Equations

Irreducible kernels with various degrees of irreducibility in various channels can be defined in a similar way. Corresponding expansions of N-point functions follow, in terms of integrals involving these kernels and two-point functions. These kernels are again convergent at small coupling ($\to 0$ as their irreducibility $\to \infty$) as well as the corresponding structure equations (which generalize eqn [18]).

Analyticity, AC, and Bound States

As explained in the introduction, we now proceed by analytic continuation away from the Euclidean region in complex energy–momentum space.

First, it is easily seen that the two-point function is analytic in the region $s < (2m)^2 - \epsilon$ apart from a pole at $s = m_{ph}^2$ which defines the physical mass m_{ph} (m_{ph}^2 is the zero in p^2 of the denominator in formula [17]). In view of the bounds of the previous two sections, m_{ph} is close to the "bare" mass m.

The 2PI kernel, for even theories, is shown, again by Laplace transform theorem, to be analytic and bounded in domains around and away from the Euclidean region up to $s = (4m)^2 - \epsilon$, and is of the order of λ_{ren}.

As we have seen in the section "AC and analyticity," the analyticity of G_2 entails the analytic structure of F (two-sheeted or multisheeted at the threshold). On the other hand, further poles of F can be generated by the BS integral equation [2] in the physical or unphysical sheets. If a pole in the physical sheet occurs at $s < (2m_{ph})^2$ real, it will correspond to a new particle in the theory, namely a two-particle bound state.

AC in the Low-Energy Region

The analysis of possible bound states, which will be presented in the following, will show that there might be at most one two-particle bound state of mass $m_B < 2m_{ph}$ which tends to $2m_{ph}$ as the couplings tends to zero.

On the other hand, for even theories, in view of the analyticity properties of the two-point function and of the 2PI kernel G_2, equation [1] holds in the region $(2m_{ph}^2) < s < (4m_{ph})^2 - \epsilon$, where $*$ is on-shell convolution with particles of mass m_{ph}.

If there is no two-particle bound state, this characterizes the AC of the theory for $s < (4m_{ph})^2 - \epsilon$.

If there is a bound state of mass m_B, AC is established only in the region $s < (3m_{ph})^2 - \epsilon$.

For non-even theories, the analysis is similar but requires the introduction of new irreducible kernels in view of the fact that the non-evenness opens new channels. AC in all cases can be established, for small couplings, up to $s < (3m_{ph})^2 - \epsilon$.

Analysis of Possible Two-Particle Bound States for Even Theories at Small Coupling

It can be checked that such poles of F, if there are, either lie far away in the unphysical sheet(s) or are close to the two-particle threshold ($s = (2m_{ph})^2$). This is due to the convergence, at small coupling, of the Neumann series $F = G_2 + G_2 \circ_M G_2 + \cdots$. Individual terms $G_2 \circ_M \cdots \circ_M G_2$ are, in fact, defined away from the Euclidean region by analytic continuation in a two-sheeted (d even) or multisheeted (d odd) domain around the threshold: to that purpose locally distorted integration contours (initially the Euclidean region) are introduced as in the section "AC and analyticity," so as to avoid the pole singularities of the two-point functions involved in $\circ_M$, the threshold singularities being due to the pinching of this contour between the two poles as $s \to (2m_{ph})^2$. If a fixed neighborhood of the threshold is excluded, one does obtain uniform bounds of the form $(\text{cst}\ \lambda_{ren})^q$ (for a term with q factors G_2) in any bounded domain, which ensures the convergence of the Neumann series.

It remains to study the neighborhood of the threshold. To that purpose, the following method is convenient. One shows that the convolution operator $\circ_M$ can be written in the form

$$\circ_M = g(s) * + \nabla \qquad [21]$$

where $*$ is, as in the section "AC and analyticity," on-shell convolution for $s > (2m_{ph})^2$ or is obtained by analytic continuation for complex value of s around the threshold; $g(s) = 1/2$ for d even and, if d is odd, $g(s) = (i/2\pi) \log \sigma$, where $\sigma = 4m_{ph}^2 - s$. In view of this definition of $g(s)$, the operator ∇ is regular: it is an analytic one-sheeted operation around the threshold (this is equivalent to [4]), and it has no pole singularities. This property of ∇ can be established by geometric methods or by an explicit evaluation.

It is then useful to introduce a new kernel U linked to G_2 by the integral equation

$$U = G_2 + U \nabla G_2 \qquad [22]$$

In view of the regularity and bounds of ∇ and G_2, one sees (e.g., by a series expansion) that U, like G_2, is analytic in a neighborhood of the threshold and behaves in the same way at small λ_{ren}.

By a simple algebraic argument F and U are related by the integral equations

$$F = U + g(s) U * F = U + g(s) F * U \qquad [23]$$

Two-dimensional models We start the analysis with the case $d = 2$. The mass shell is trivial in this case; let f be the restriction of F to the mass shell; it depends only on $s = (p_3 + p_4)^2$ due to the mass shell and e.m.c. constraints (as also Lorentz invariance). On the mass shell, the operation $*$ becomes a mere multiplication and the integral equation [23] becomes

$$f(s) = u(s) + \frac{1}{a(s)} f(s) u(s) \qquad [24]$$

where u is the mass shell restriction of U and the factor $a(s)$ arising from $*$ is of the form $a(s) = \text{cst}\ s^{1/2} \sigma^{1/2}$, $\sigma = (2m_{ph})^2 - s$, which gives

$$f(s) = \frac{a(s) u(s)}{a(s) - u(s)} \qquad [25]$$

In turn one obtains

$$F = U + \frac{U_{|\ |}U}{a(s) - u(s)} \qquad [26]$$

where $U_{|}$ (resp., $_{|}U$) is U with p_3, p_4 (resp., p_1, p_2) restricted to the mass shell. Equation [26] completely characterizes the local structure of F in view of the local analyticity of U.

The analysis of the possible poles follows from the fact that U is equal to G_2 up to higher order in λ_{ren}; on the other hand, G_2 is equal to a first known term plus higher-order corrections in λ_{ren} (if we expand in λ_{ren} the expression for G_2 obtained in the previous section), so that the leading contribution of $u(s)$ is known and the results follow.

For a theory (see [8]) containing a $\lambda_{\text{ren}}\varphi^4$ term there is exactly one pole, which corresponds to the zero of $a(s) - u(s)$, lying in the region $(2m_{\text{ph}})^2 - \epsilon < s < (2m_{\text{ph}})^2$. This pole is either in the physical sheet for $\lambda_{\text{ren}} < 0$ or in the second sheet if $\lambda_{\text{ren}} > 0$. In the case $\lambda_{\text{ren}} < 0$, this pole corresponds to a two-particle bound state of physical mass m_{B} which tends to $2m_{\text{ph}}$ as $\lambda_{\text{ren}} \to 0$.

In a model without φ^4 term ($\lambda_{\text{ren}} = 0$) the lowest-order contribution to G_2, hence to U, is in general of the order of the square of the leading coupling, in which case there is always one bound state.

The treatment of the fermionic Gross–Neveu model, which involves spin and color indices, is analogous, with minor modifications. Equations now involve, in the two-particle region, 4×4 matrices; poles of F are now the zeros of $\det(a(s)I - m(s)u(s))$, where $m(s)$ is the 4×4 matrix obtained from 2×2 residue matrices (whose leading matrix elements are explicitly computable). The detailed analysis, which requires the consideration of different channels (various color and spin indices) is omitted.

Three-dimensional models The results are similar: F is decomposed as $F' + F''$, where F' is the $\ell = 0$ "partial wave component" of F, namely $F' = (1/2\pi) \int F \, d\theta$, where θ is the "scattering angle" of the channel; its complement F'' is shown to be locally bounded in view of a further factor σ. The analysis is then analogous to the case $d = 2$ with $a(s)$ now behaving like $\text{cst}/\log\sigma$ as $\sigma \to 0$. There is, *a priori*, either no pole, or one pole in the physical sheet at $s = m_{\text{B}}^2 < (2m_{\text{ph}})^2$ with $m_{\text{B}} = 2m_{\text{ph}} + O(e^{-\text{cst}/\lambda_{\text{ren}}})$, depending again on the signs of the couplings. For the existing even models such as the φ^4 model, there is no pole, hence no two-particle bound state.

Four-dimensional models The existence of the φ^4 model in dimension 4 is doubtful. If a four-dimensional model were defined, and if the 2PI kernel G_2 of a massive channel could be defined and shown to satisfy analyticity properties analogous to above, there would be no two-particle bound state at small coupling. In fact, the kinematical factor $\sigma^{(d-3)/2}$ (for d even) generated by the mass shell convolution is no longer equal to $\sigma^{-1/2}$ as in the $d = 2$ case but now to $\sigma^{1/2}$. As a consequence, the Neumann series giving F in terms of G_2 is convergent also in the neighborhood of the two-particle threshold.

Non-even theories The analysis for the non-even theories follows similar lines. As already mentioned, the analysis requires the introduction of new irreducible kernels. For the models $\lambda\varphi^4 + c_3\varphi^3$, which do exist at small couplings in dimensions 2 and 3, there will be either exactly one or no two-particle bound state, depending on the respective values of λ, c_3.

Structure Equations and AC in Higher-Energy Regions

The structure equations of the previous section provide, after analytical continuation away from the Euclidean region, a rigorous version of the analysis presented at the end of the section "AC and analyticity." The irreducible kernels can here be defined in a direct way following the previous section, together with their analyticity properties. One has then to derive the discontinuity formulas that in turn characterize AC. This program has been carried out in the $3 \to 3$ particle region, and partly in the general case. It seems possible to complete general proofs up to some technical (difficult) problems. As already mentioned, in this approach, the coupling should be taken smaller and smaller as the energy region considered increases.

See also: Axiomatic Quantum Field Theory; Constructive Quantum Field Theory; Dispersion Relations; Dynamical Systems in Mathematical Physics: An Illustration from Water Waves; Perturbation Theory and its Techniques; Quantum Chromodynamics; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Scattering in Relativistic Quantum Field Theory: the Analytic Program; Schrödinger operators.

Further Reading

Bros J (1984) *r*-particle irreducible kernels, asymptotic completeness and analyticity properties of several particle collision amplitudes. *Physica A* 124: 145.

Bros J, Epstein H, and Glaser V (1965) A proof of the crossing property for two-particle amplitudes in general quantum field theory. *Communications in Mathematical Physics* 1: 240.

Bros J, Epstein H, and Glaser V (1972) Local analyticity properties of the *n*-particle scattering amplitude. *Helvetica Physica Acta* 43: 149.

Epstein H, Glaser V, and Iagolnitzer D (1981) Some analyticity properties arising from asymptotic completeness in quantum field theory. *Communications in Mathematical Physics* 80: 99.

Glimm J and Jaffe A (1981, 1987) *Quantum Physics: A Functionnal Integral Point of View*. Heidelberg: Springer.

Iagolnitzer D (1992) *Scattering in Quantum Field Theories: The Axiomatic and Constructive Approaches*, Princeton Series in Physics. Princeton University Press.

Iagolnitzer D and Magnen J (1987a) Asymptotic completeness and multiparticle structure in field theories. *Communications in Mathematical Physics* 110: 51.

Iagolnitzer D and Magnen J (1987b) Asymptotic completeness and multiparticle structure in field theories. II. Theories with renormalization: the Gross–Neveu model. *Communications in Mathematical Physics* 111: 89.

Martin A (1970) *Scattering Theory: Unitarity, Analyticity and Crossing*. Heidelberg: Springer.

Rivasseau V (1991) *From Perturbative to Constructive Renormalization*. Princeton: Princeton University Press.

Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools

D Buchholz, Universität Göttingen, Göttingen, Germany
S J Summers, University of Florida, Gainesville, FL, USA

© 2006 Elsevier Ltd. All rights reserved.

Physical Motivation and Mathematical Setting

The primary connection of relativistic quantum field theory to experimental physics is through scattering theory, that is, the theory of the collision of elementary (or compound) particles. It is therefore a central topic in quantum field theory and has attracted the attention of leading mathematical physicists. Although a great deal of progress has been made in the mathematically rigorous understanding of the subject, there are important matters which are still unclear, some of which will be indicated below.

In the paradigmatic scattering experiment, several particles, which are initially sufficiently distant from each other that the idealization that they are not mutually interacting is physically reasonable, approach each other and interact (collide) in a region of microscopic extent. The products of this collision then fly apart until they are sufficiently well separated that the approximation of noninteraction is again reasonable. The initial and final states of the objects in the scattering experiment are therefore to be modeled by states of noninteracting, that is, free, fields, which are mathematically represented on Fock space. Typically, what is measured in such experiments is the probability distribution (cross section) for the transitions from a specified state of the incoming particles to a specified state of the outgoing particles.

It should be mentioned that until the late 1950s, the scattering theory of relativistic quantum particles relied upon ideas from nonrelativistic quantum-mechanical scattering theory (interaction representation, adiabatic limit, etc.), which were invalid in the relativistic context. Only with the advent of axiomatic quantum field theory did it become possible to properly formulate the concepts and mathematical techniques which will be outlined here.

Scattering theory can be rigorously formulated either in the context of quantum fields satisfying the Wightman axioms (Streater and Wightman 1964) or in terms of local algebras satisfying the Haag–Kastler–Araki axioms (Haag 1992). In brief, the relation between these two settings may be described as follows: in the Wightman setting, the theory is formulated in terms of operator-valued distributions ϕ on Minkowski space, the quantum fields, which act on the physical state space. These fields, integrated with test functions f having support in a given region $\mathcal{O}$ of spacetime (only four-dimensional Minkowski space R^4 will be treated here), $\phi(f) = \int \mathrm{d}^4x\, f(x)\phi(x)$, form under the operations of addition, multiplication, and Hermitian conjugation a polynomial *-algebra $\mathcal{P}(\mathcal{O})$ of unbounded operators. In the Haag–Kastler–Araki setting, one proceeds from these algebras to algebras $\mathcal{A}(\mathcal{O})$ of bounded operators which, roughly speaking, are formed by the bounded functions A of the operators $\phi(f)$. This step requires some mathematical care, but these subtleties will not be discussed here. As the statements and proofs of the results in these two frameworks differ only in technical details, the theory is presented here in the more convenient setting of algebras of bounded operators (C*-algebras).

Central to the theory is the notion of a particle, which, in fact, is a quite complex concept, the full nature of which is not completely understood, cf.

Notes:

below. In order to maintain the focus on the essential points, we consider in the subsequent sections primarily a single massive particle of integer spin s, that is, a boson. In standard scattering theory based upon Wigner's characterization, this particle is simply identified with an irreducible unitary representation U_1 of the identity component $\mathcal{P}_+^\uparrow$ of the Poincaré group with spin s and mass $m > 0$. The Hilbert space $\mathcal{H}_1$ upon which $U_1(\mathcal{P}_+^\uparrow)$ acts is called the one-particle space and determines the possible states of a single particle, alone in the universe. Assuming that configurations of several such particles do not interact, one can proceed by a standard construction to a Fock space describing freely propagating multiple particle states,

$$\mathcal{H}_F = \bigoplus_{n \in \mathbb{N}_0} \mathcal{H}_n$$

where $\mathcal{H}_0 = \mathbb{C}$ and $\mathcal{H}_n$ is the n-fold symmetrized direct product of $\mathcal{H}_1$ with itself. This space is spanned by vectors $\Phi_1 \otimes \cdots \otimes \Phi_n$, where $\otimes$ denotes the symmetrized tensor product, representing an n-particle state wherein the kth particle is in the state $\Phi_k \in \mathcal{H}_1$, $k = 1, \ldots, n$. The representation $U_1(\mathcal{P}_+^\uparrow)$ induces a unitary representation $U_F(\mathcal{P}_+^\uparrow)$ on $\mathcal{H}_F$ by

$$U_F(\lambda)(\Phi_1 \otimes \cdots \otimes \Phi_n) \doteq U_1(\lambda)\Phi_1 \otimes \cdots \otimes U_1(\lambda)\Phi_n \quad [1]$$

In interacting theories, the states in the corresponding physical Hilbert space $\mathcal{H}$ do not have such an *a priori* interpretation in physical terms, however. It is the primary goal of scattering theory to identify in $\mathcal{H}$ those vectors which describe, at asymptotic times, incoming, respectively, outgoing, configurations of freely moving particles. Mathematically, this amounts to the construction of certain specific isometries (generalized Møller operators), Ω^{in} and Ω^{out}, mapping $\mathcal{H}_F$ onto subspaces $\mathcal{H}^{\text{in}} \subset \mathcal{H}$ and $\mathcal{H}^{\text{out}} \subset \mathcal{H}$, respectively, and intertwining the unitary actions of the Poincaré group on $\mathcal{H}_F$ and $\mathcal{H}$. The resulting vectors

$$(\Phi_1 \otimes \cdots \otimes \Phi_n)^{\text{in/out}} \doteq \Omega^{\text{in/out}}(\Phi_1 \otimes \cdots \otimes \Phi_n) \in \mathcal{H} \quad [2]$$

are interpreted as incoming and outgoing particle configurations in scattering processes wherein the kth particle is in the state $\Phi_k \in \mathcal{H}_1$.

If, in a theory, the equality $\mathcal{H}^{\text{in}} = \mathcal{H}^{\text{out}}$ holds, then every incoming scattering state evolves, after the collision processes at finite times, into an outgoing scattering state. It is then physically meaningful to define on this space of states the scattering matrix, setting $S = \Omega^{\text{in}}\Omega^{\text{out}*}$. Physical data such as collision cross sections can be derived from S and the corresponding transition amplitudes $\langle (\Phi_1 \otimes \cdots \otimes \Phi_m)^{\text{in}}, (\Phi'_1 \otimes \cdots \otimes \Phi'_n)^{\text{out}} \rangle$, respectively, by a standard procedure. It should be noted, however, that neither the above physically mandatory equality of state spaces nor the more stringent requirement that every state has an interpretation in terms of incoming and outgoing scattering states, that is, $\mathcal{H} = \mathcal{H}^{\text{in}} = \mathcal{H}^{\text{out}}$ (asymptotic completeness), has been fully established in any interacting relativistic field theoretic model so far. This intriguing problem will be touched upon in the last section of this article.

Before going into details, let us state the few physically motivated postulates entering into the analysis. As discussed, the point of departure is a family of algebras $\mathcal{A}(\mathcal{O})$, more precisely a net, associated with the open subregions $\mathcal{O}$ of Minkowski space and acting on $\mathcal{H}$. Restricting attention to the case of bosons, we may assume that this net is local in the sense that if $\mathcal{O}_1$ is spacelike separated from $\mathcal{O}_2$, then all elements of $\mathcal{A}(\mathcal{O}_1)$ commute with all elements of $\mathcal{A}(\mathcal{O}_2)$. (In the presence of fermions, these algebras contain also fermionic operators which anticommute.) This is the mathematical expression of the principle of Einstein causality. The unitary representation U of $\mathcal{P}_+^\uparrow$ acting on $\mathcal{H}$ is assumed to satisfy the relativistic spectrum condition (positivity of energy in all Lorentz frames) and, in the sense of equality of sets, $U(\lambda)\mathcal{A}(\mathcal{O})U(\lambda)^{-1} = \mathcal{A}(\lambda\mathcal{O})$ for all $\lambda \in \mathcal{P}_+^\uparrow$ and regions $\mathcal{O}$, where $\lambda\mathcal{O}$ denotes the Poincaré transformed region. It is also assumed that the subspace of $U(\mathcal{P}_+^\uparrow)$-invariant vectors is spanned by a single unit vector Ω, representing the vacuum, which has the Reeh–Schlieder property, that is, each set of vectors $\mathcal{A}(\mathcal{O})\Omega$ is dense in $\mathcal{H}$. These standing assumptions will subsequently be amended by further conditions concerning the particle content of the theory.

Haag–Ruelle Theory

Haag and Ruelle were the first to establish the existence of scattering states within this general framework (Jost 1965); further substantial improvements are due to Araki and Hepp (Araki 1999). In all of these investigations, the arguments were given for quantum field theories with associated particles (in the Wigner sense) which have strictly positive mass $m > 0$ and for which m is an isolated eigenvalue of the mass operator (upper and lower mass gap). Moreover, it was assumed that states of a single particle can be created from the vacuum by local operations. In physical terms, these assumptions allow only for theories with short-range interactions and particles carrying strictly localizable charges.

In view of these limitations, Haag–Ruelle theory has been developed in a number of different directions. By now, the scattering theory of massive particles is under complete control, including also

particles carrying nonlocalizable (gauge or topological) charges and particles having exotic statistics (anyons, plektons) which can appear in theories in low spacetime dimensions. Due to constraints of space, these results must go without further mention; we refer the interested reader to the articles Buchholz and Fredenhagen (1982) and Fredenhagen *et al.* (1996). Theories of massless particles and of particles carrying charges of electric or magnetic type (infraparticles) will be discussed in subsequent sections.

We outline here a recent generalization of Haag–Ruelle scattering theory presented in Dybalski (2005), which covers massive particles with localizable charges without relying on any further constraints on the mass spectrum. In particular, the scattering of electrically neutral, stable particles fulfilling a sharp dispersion law in the presence of massless particles is included (e.g., neutral atoms in their ground states). Mathematically, this assumption can be expressed by the requirement that there exists a subspace $\mathcal{H}_1 \subset \mathcal{H}$ such that the restriction of $U(\mathcal{P}_+^\uparrow)$ to $\mathcal{H}_1$ is a representation of mass $m > 0$. We denote by P_1 the projection in $\mathcal{H}$ onto $\mathcal{H}_1$.

To establish notation, let $\mathcal{O}$ be a bounded spacetime region and let $A \in \mathcal{A}(\mathcal{O})$ be any operator such that $P_1 A\Omega \neq 0$. The existence of such localized (in brief, local) operators amounts to the assumption that the particle carries a localizable charge. That the particle is stable, that is, completely decouples from the underlying continuum states, can be cast into a condition first stated by Herbst: for all sufficiently small $\mu > 0$

$$\|E_\mu(1 - P_1)A\Omega\| \leq c\mu^\eta \qquad [3]$$

for some constants $c, \eta > 0$, where E_μ is the projection onto the spectral subspace of the mass operator corresponding to spectrum in the interval $(m - \mu, m + \mu)$. In the case originally considered by Haag and Ruelle, where m is isolated from the rest of the mass spectrum, this condition is certainly satisfied.

Setting $A(x) \doteq U(x)AU(x)^{-1}$, where $U(x)$ is the unitary implementing the spacetime translation $x = (x_0, \boldsymbol{x})$ (the velocity of light and Planck's constant are set equal to 1 in what follows), one puts, for $t \neq 0$,

$$A_t(f) = \int \mathrm{d}^4x\, g_t(x_0) f_{x_0}(\boldsymbol{x}) A(x) \qquad [4]$$

Here $x_0 \mapsto g_t(x_0) \doteq g((x_0 - t)/|t|^\kappa)/|t|^\kappa$ induces a time averaging about t, g being any test function which integrates to 1 and whose Fourier transform has compact support, and $1/(1 + \eta) < \kappa < 1$ with η as above. The Fourier transform of f_{x_0} is given by $\widetilde{f_{x_0}}(\boldsymbol{p}) \doteq \tilde{f}(\boldsymbol{p})\, \mathrm{e}^{-\mathrm{i}x_0\omega(\boldsymbol{p})}$, where f is some test function on $\mathbb{R}^3$ with $\tilde{f}(\boldsymbol{p})$ having compact support, and $\omega(\boldsymbol{p}) = (\boldsymbol{p}^2 + m^2)^{1/2}$. Note that $(x_0, \boldsymbol{x}) \mapsto f_{x_0}(\boldsymbol{x})$ is a solution of the Klein–Gordon equation of mass m.

With these assumptions, it follows by a straightforward application of the harmonic analysis of unitary groups that in the sense of strong convergence $A_t(f)\Omega \to P_1 A(f)\Omega$ and $A_t(f)^*\Omega \to 0$ as $t \to \pm\infty$, where $A(f) = \int \mathrm{d}^3x\, f(\boldsymbol{x}) A(0, \boldsymbol{x})$. Hence, the operators $A_t(f)$ may be thought of as creation operators and their adjoints as annihilation operators. These operators are the basic ingredients in the construction of scattering states. Choosing local operators A_k as above and test functions $f^{(k)}$ with disjoint compact supports in momentum space, $k = 1, \ldots, n$, the scattering states are obtained as limits of the Haag–Ruelle approximants

$$A_{1t}(f^{(1)}) \cdots A_{nt}(f^{(n)})\Omega \qquad [5]$$

Roughly speaking, the operators $A_{kt}(f^{(k)})$ are localized in spacelike separated regions at asymptotic times t, due to the support properties of the Fourier transforms of the functions $f^{(k)}$. Hence they commute asymptotically because of locality and, by the clustering properties of the vacuum state, the above vector becomes a product state of single-particle states. In order to prove convergence, one proceeds, in analogy to Cook's method in quantum-mechanical scattering theory, to the time derivatives,

$$\begin{aligned} &\partial_t A_{1t}(f^{(1)}) \cdots A_{nt}(f^{(n)})\Omega \\ &\quad = \sum_{k \neq l} A_{1t}(f^{(1)}) \cdots [\partial_t A_{kt}(f^{(k)}), A_{lt}(f^{(l)})] \cdots A_{nt}(f^{(n)})\Omega \\ &\qquad + \sum_k A_{1t}(f^{(1)}) \cdots \overset{k}{\vee} \cdots A_{nt}(f^{(n)}) \partial_t A_{kt}(f^{(k)})\Omega \end{aligned} \qquad [6]$$

where $\overset{k}{\vee}$ denotes omission of $A_{kt}(f^{(k)})$. Employing techniques of Araki and Hepp, one can prove that the terms in the first summation on the right-hand side (RHS) of [6], involving commutators, decay rapidly in norm as t approaches infinity because of locality, as indicated above. By applying condition [3] and the fact that the vectors $\partial_t A_{kt}(f^{(k)})\Omega$ do not have a component in the single-particle space $\mathcal{H}_1$, the terms in the second summation on the RHS of [6] can be shown to decay in norm like $|t|^{-\kappa(1+\eta)}$. Thus, the norm of the vector [6] is integrable in t, implying the existence of the strong limits

$$\begin{aligned} &\left(P_1 A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1 A_n(f^{(n)})\Omega\right)^{\text{in/out}} \\ &\quad \doteq \lim_{t \to \mp\infty} A_{1t}(f^{(1)}) \cdots A_{nt}(f^{(n)})\Omega \end{aligned} \qquad [7]$$

As indicated by the notation, these limits depend only on the single-particle vectors $P_1A_k(f^{(k)})\Omega \in \mathcal{H}_1$, $k=1,\ldots,n$, but not on the specific choice of operators and test functions. In order to establish their Fock structure, one employs results on clustering properties of vacuum correlation functions in theories without strictly positive minimal mass. Using this, one can compute inner products of arbitrary asymptotic states and verify that the maps

$$\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big) \mapsto \Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \quad [8]$$

extend by linearity to isomorphisms $\Omega^{\text{in/out}}$ from the Fock space $\mathcal{H}_F$ onto the subspaces $\mathcal{H}^{\text{in/out}} \subset \mathcal{H}$ generated by the collision states. Moreover, the asymptotic states transform under the Poincaré transformations $U(\mathcal{P}_+^\uparrow)$ as

$$U(\lambda)\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} = \Big(U_1(\lambda)P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes U_1(\lambda)P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \quad [9]$$

Thus, the isomorphisms $\Omega^{\text{in/out}}$ intertwine the action of the Poincaré group on $\mathcal{H}_F$ and $\mathcal{H}^{\text{in/out}}$. We summarize these results, which are vital for the physical interpretation of the underlying theory, in the following theorem.

Theorem 1 *Consider a theory of a particle of mass $m>0$ which satisfies the standing assumptions and the stability condition* [3]. *Then there exist canonical isometries $\Omega^{\text{in/out}}$, mapping the Fock space $\mathcal{H}_F$ based on the single-particle space $\mathcal{H}_1$ onto subspaces $\mathcal{H}^{\text{in/out}} \subset \mathcal{H}$ of incoming and outgoing scattering states. Moreover, these isometries intertwine the action of the Poincaré transformations on the respective spaces.*

Since the scattering states have been identified with Fock space, asymptotic creation and annihilation operators act on $\mathcal{H}^{\text{in/out}}$ in a natural manner. This point will be explained in the following section.

LSZ Formalism

Prior to the results of Haag and Ruelle, an axiomatic approach to scattering theory was developed by Lehmann, Symanzik, and Zimmermann (LSZ), based on time-ordered vacuum expectation values of quantum fields. The relative advantage of their approach with respect to Haag–Ruelle theory is that useful reduction formulas for the S-matrix greatly facilitate computations, in particular in perturbation theory. Moreover, these formulas are the starting point of general studies of the momentum space analyticity properties of the S-matrix (dispersion relations), as outlined in Dispersion Relations (cf. also Iagolnitzer (1993)). Within the present general setting, the LSZ method was established by Hepp.

For simplicity of discussion, we consider again a single particle type of mass $m>0$ and integer spin s, subject to condition [3]. According to the results of the preceding section, one then can consistently define asymptotic creation operators on the scattering states, setting

$$\begin{aligned} A(f)^{\text{in/out}}&\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \\ &\doteq \lim_{t\to\mp\infty} A_t(f)\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \\ &= \Big(P_1A(f)\Omega \otimes P_1A_1(f^{(1)})\Omega \otimes \cdots \\ &\quad \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \end{aligned} \quad [10]$$

Similarly, one obtains the corresponding asymptotic annihilation operators,

$$\begin{aligned} A(f)^{\text{in/out}*}&\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \\ &= \lim_{t\to\mp\infty} A_t(f)^*\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \\ &\quad \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} = 0 \end{aligned} \quad [11]$$

where the latter equality holds if the Fourier transforms of the functions $f, f^{(1)}, \ldots, f^{(n)}$, have disjoint supports. We mention as an aside that, by replacing the time-averaging function g in the definition of $A_t(f)$ by a delta function, the above formulas still hold. But the convergence is then to be understood in the weak Hilbert space topology. In this form, the above relations were anticipated by LSZ (asymptotic condition).

It is straightforward to proceed from these relations to reduction formulas. Let B be any local operator. Then one has, in the sense of matrix elements between outgoing and incoming scattering states,

$$\begin{aligned} BA(f)^{\text{in}} - A(f)^{\text{out}}B &= \lim_{t\to\infty}(BA(f_{-t}) - A(f_t)B) \\ &= \lim_{t\to\infty}\Big(\int \mathrm{d}^4x f_{-t}(x)BA(x) - \int \mathrm{d}^4x f_t(x)A(x)B\Big) \end{aligned} \quad [12]$$

$f_t(x) \doteq g_t(x_0)f(x_0)(\mathrm{vec}(x))$. Because of the (essential) support properties of the functions $f_{\pm t}$, the contributions to the latter integrals arise, for asymptotic t, from spacetime points x where the localization

regions of $A(x)$ and B have a negative timelike (first term), respectively, positive timelike (second term) distance. One may therefore proceed from the products of these operators to the time-ordered products $T(BA(x))$, where $T(BA(x))=A(x)B$ if the localization region of $A(x)$ lies in the future of that of B, and $T(BA(x))=BA(x)$ if it lies in the past. It is noteworthy that a precise definition of the time ordering for finite x is irrelevant in the present context – any reasonable interpolation between the above relations will do. Similarly, one can define time-ordered products for an arbitrary number of local operators. The preceding limit can then be recast into

$$\lim_{t\to\infty}\int \mathrm{d}^4x(f_{-t}(x)-f_t(x))T(BA(x)) \qquad [13]$$

The latter expression has a particularly simple form in momentum space. Proceeding to the Fourier transforms of $f_{\pm t}$ and noticing that, in the limit of large t,

$$\left(\widetilde{f_{-t}}(p)-\widetilde{f_t}(p)\right)/(p_0-\omega(\boldsymbol{p})) \longrightarrow -2\pi\mathrm{i}\widetilde{f}(\boldsymbol{p})\,\delta(p_0-\omega(\boldsymbol{p})) \qquad [14]$$

one gets

$$BA(f)^{\mathrm{in}}-A(f)^{\mathrm{out}}B = -2\pi\mathrm{i}\int \mathrm{d}^3p\widetilde{f}(\boldsymbol{p})(p_0-\omega(\boldsymbol{p})) \times T(B\widetilde{A}(-p))\Big|_{p_0=\omega(\boldsymbol{p})} \qquad [15]$$

Here $T(B\widetilde{A}(p))$ denotes the Fourier transform of $T(BA(x))$, and it can be shown that the restriction of $(p_0-\omega(\boldsymbol{p}))T(B\widetilde{A}(-p))$ to the manifold $\{p\in\mathbb{R}^4\colon p_0=\omega(\boldsymbol{p})\}$ (the "mass shell") is meaningful in the sense of distributions on $\mathbb{R}^3$. By the same token, one obtains

$$A(f)^{\mathrm{out}*}B-BA(f)^{\mathrm{in}*} = -2\pi\mathrm{i}\int \mathrm{d}^3p\,\overline{\widetilde{f}(\boldsymbol{p})}\,(p_0-\omega(\boldsymbol{p}))T(\widetilde{A^*}(p)B)\Big|_{p_0=\omega(\boldsymbol{p})} \qquad [16]$$

Similar relations, involving an arbitrary number of asymptotic creation and annihilation operators, can be established by analogous considerations. Taking matrix elements of these relations in the vacuum state and recalling the action of the asymptotic creation and annihilation operators on scattering states, one arrives at the following result, which is central in all applications of scattering theory.

Theorem 2 *Consider the theory of a particle of mass $m>0$ subject to the conditions stated in the preceding sections and let $f^{(1)},\ldots,f^{(n)}$ be any family of test functions whose Fourier transforms have compact and nonoverlapping supports. Then*

$$\begin{aligned}&\left\langle\left(P_1A_1(f^{(1)})\Omega\otimes\cdots\otimes P_1A_k(f^{(k)})\Omega\right)^{\mathrm{out}},\right.\\&\left.\left(P_1A_{k+1}(f^{(k+1)})\Omega\otimes\cdots\otimes P_1A_n(f^{(n)})\Omega\right)^{\mathrm{in}}\right\rangle\\&=(-2\pi\mathrm{i})^n\int\cdots\int \mathrm{d}^3p_1\cdots\mathrm{d}^3p_n\,\overline{\widetilde{f^{(1)}}(\boldsymbol{p}_1)}\cdots\\&\times\overline{\widetilde{f^{(k)}}(\boldsymbol{p}_k)}\widetilde{f^{(k+1)}}(\boldsymbol{p}_{k+1})\cdots\widetilde{f^{(n)}}(\boldsymbol{p}_n)\\&\times\prod_{i=1}^{n}(p_{i_0}-\omega(\boldsymbol{p}_i))\Big\langle\Omega,T\Big(\widetilde{A_1^*}(p_1)\cdots\\&\times\widetilde{A_k^*}(p_k)\widetilde{A_{k+1}}(-p_{k+1})\cdots\\&\times\widetilde{A_n}(-p_n)\Omega\Big)\Big\rangle\Big|_{p_{i_0}=\omega(\boldsymbol{p}_i)}^{i=1,\ldots,n}\end{aligned} \qquad [17]$$

in an obvious notation.

Thus, the kernels of the scattering amplitudes in momentum space are obtained by restricting the (by the factor $\prod_{i=1}^{n}(p_{i_0}-\omega(\boldsymbol{p}_i))$) amputated Fourier transforms of the vacuum expectation values of the time-ordered products to the positive and negative mass shells, respectively. These are the famous LSZ reduction formulas, which provide a convenient link between the time-ordered (Green's) functions of a theory and its asymptotic particle interpretation.

Asymptotic Particle Counters

The preceding construction of scattering states applies to a significant class of theories; but even if one restricts attention to the case of massive particles, it does not cover all situations of physical interest. For an essential input in the construction is the existence of local operators interpolating between the vacuum and the single-particle states. There may be no such operators at one's disposal, however, either because the particle in question carries a nonlocalizable charge, or because the given family of operators is too small. The latter case appears, for example, in gauge theories, where in general only the observables are fixed by the principle of local gauge invariance, and the physical particle content as well as the corresponding interpolating operators are not known from the outset. As observables create from the vacuum only neutral states, the above construction of scattering states then fails if charged particles are present. Nevertheless, thinking in physical terms, one would expect that the observables contain all relevant information in order to determine the features of scattering states, in particular their collision cross section. That this is indeed the case was first shown by Araki and Haag (Araki 1999).

In scattering experiments, the measured data are provided by detectors (e.g., particle counters) and

coincidence arrangements of detectors. Essential features of detectors are their lack of response in the vacuum state and their macroscopic localization. Hence, within the present mathematical setting, a general detector is represented by a positive operator C on the physical Hilbert space $\mathcal{H}$ such that $C\Omega = 0$. Because of the Reeh–Schlieder theorem, these conditions cannot be satisfied by local operators. However, they can be fulfilled by "almost-local" operators. Examples of such operators are easy to produce, putting $C = L^*L$ with

$$L = \int \mathrm{d}^4x\, f(x)\, A(x) \qquad [18]$$

where A is any local operator and f any test function whose Fourier transform has compact support in the complement of the closed forward light cone (and hence in the complement of the energy momentum spectrum of the theory). In view of the properties of f and the invariance of Ω under translations, it follows that $C = L^*L$ annihilates the vacuum and can be approximated with arbitrary precision by local operators. The algebra generated by these operators C will be denoted by $\mathcal{C}$.

When preparing a scattering experiment, the first thing one must do with a detector is to calibrate it, that is, test its response to sources of single-particle states. Within the mathematical setting, this amounts to computing the matrix elements of C in states $\Phi \in \mathcal{H}_1$:

$$\langle \Phi, C\Phi \rangle = \int\int \mathrm{d}^3p\, \mathrm{d}^3q\, \overline{\Phi(p)}\, \Phi(q)\, \langle p|C|q\rangle \qquad [19]$$

Here $p \mapsto \Phi(p)$ is the momentum space wave function of Φ, $\langle \cdot|C|\cdot\rangle$ is the kernel of C in the single-particle space $\mathcal{H}_1$, and we have omitted (summations over) indices labeling internal degrees of freedom of the particle, if any. The relevant information about C is encoded in its kernel. As a matter of fact, one only needs to know its restriction to the diagonal, $p \mapsto \langle p|C|p\rangle$. It is called the sensitivity function of C and can be shown to be regular under quite general circumstances (Araki 1999, Buchholz and Fredenhagen 1982).

Given a state $\Psi \in \mathcal{H}$ for which the expectation value $\langle \Psi, C(x)\Psi \rangle$ differs significantly from 0, one concludes that this state deviates from the vacuum in a region about x. For finite x, this does not mean, however, that Ψ has a particle interpretation at x. For that spacetime point may, for example, be just the location of a collision center. Yet, if one proceeds to asymptotic times, one expects, in view of the spreading of wave packets, that the probability of finding two or more particles in the same spacetime region is dominated by the single-particle contributions. It is this physical insight which justifies the expectation that the detectors $C(x)$ become particle counters at asymptotic times. Accordingly, one considers for asymptotic t the operators

$$C_t(h) \doteq \int \mathrm{d}^3x\, h(x/t)\, C(t, x) \qquad [20]$$

where h is any test function on $\mathbb{R}^3$. The role of the integral is to sum up all single-particle contributions with velocities in the support of h in order to compensate for the decreasing probability of finding such particles at asymptotic times t about the localization center of the detector. That these ideas are consistent was demonstrated by Araki and Haag, who established the following result (Araki 1999).

Theorem 3 *Consider, as before, the theory of a massive particle. Let $C^{(1)}, \ldots, C^{(n)} \in \mathcal{C}$ be any family of detector operators and let $h^{(1)}, \ldots, h^{(n)}$ be any family of test functions on $\mathbb{R}^3$. Then, for any state $\Psi^{\mathrm{out}} \in \mathcal{H}^{\mathrm{out}}$ of finite energy,*

$$\lim_{t\to\infty} \left\langle \Psi^{\mathrm{out}}, C_t^{(1)}(h^{(1)}) \cdots C_t^{(n)}(h^{(n)})\Psi^{\mathrm{out}} \right\rangle$$
$$= \int \cdots \int \mathrm{d}^3p_1 \cdots \mathrm{d}^3p_n\, \langle \Psi^{\mathrm{out}}, \rho^{\mathrm{out}}(p_1) \cdots \rho^{\mathrm{out}}(p_n)\Psi^{\mathrm{out}} \rangle$$
$$\times \prod_{k=1}^{n} h(p_k/\omega(p_k)) \langle p_k|C^{(k)}|p_k\rangle \qquad [21]$$

where $\rho^{\mathrm{out}}(p)$ is the momentum space density (the product of creation and annihilation operators) of outgoing particles of momentum p, and (summations over) possible indices labeling internal degrees of freedom of the particle are omitted. An analogous relation holds for incoming scattering states at negative asymptotic times.

This result shows, first of all, that the scattering states have indeed the desired interpretation with regard to the observables, as anticipated in the preceding sections. Since the assertion holds for all scattering states of finite energy, one may replace in the above theorem the outgoing scattering states by any state of finite energy, if the theory is asymptotically complete, that is, $\mathcal{H} = \mathcal{H}^{\mathrm{in}} = \mathcal{H}^{\mathrm{out}}$. Then choosing, in particular, any incoming scattering state and making use of the arbitrariness of the test functions $h^{(k)}$ as well as the knowledge of the sensitivity functions of the detector operators, one can compute the probability distributions of outgoing particle momenta in this state, and thereby the corresponding collision cross sections.

The question of how to construct certain specific incoming scattering states by using only local observables was not settled by Araki and Haag,

however. A general method to that effect was outlined in Buchholz *et al.* (1991). As a matter of fact, for that method only the knowledge of states in the subspace of neutral states is required. Yet in this approach one would need for the computation of, say, elastic collision cross sections of charged particles the vacuum correlation functions involving at least eight local observables. This practical disadvantage of increased computational complexity of the method is offset by the conceptual advantage of making no appeal to quantities which are *a priori* nonobservable.

Massless Particles and Huygens' Principle

The preceding general methods of scattering theory apply only to massive particles. Yet taking advantage of the salient fact that massless particles always move with the speed of light, Buchholz succeeded in establishing a scattering theory also for such particles (Haag 1992). Moreover, his arguments lead to a quantum version of Huygens' principle.

As in the case of massive particles, one assumes that there is a subspace $\mathcal{H}_1 \subset \mathcal{H}$ corresponding to a representation of $U(\mathcal{P}_+^\uparrow)$ of mass $m=0$ and, for simplicity, integer helicity; moreover, there must exist *local operators interpolating* between the vacuum and the single-particle states. These assumptions cover, in particular, the important examples of the photon and of Goldstone particles. Picking any suitable local operator A interpolating between Ω and some vector in $\mathcal{H}_1$, one sets, in analogy to [4],

$$A_t \doteq \int \mathrm{d}^4x\, g_t(x_0) \times (-1/2\pi)\varepsilon(x_0)\,\delta(x_0^2 - \boldsymbol{x}^2)\partial_0 A(x) \qquad [22]$$

Here $g_t(x_0) \doteq (1/|\ln t|)\, g((x_0 - t)/|\ln t|)$ with g as in [4], and the solution of the Klein–Gordon equation in [4] has been replaced by the fundamental solution of the wave equation; furthermore, $\partial_0 A(x)$ denotes the derivative of $A(x)$ with respect to x_0. Then, once again, the strong limit of $A_t\Omega$ as $t \to \pm\infty$ is $P_1 A\Omega$, with P_1 the projection onto $\mathcal{H}_1$.

In order to establish the convergence of A_t as in the LSZ approach, one now uses the fact that these operators are, at asymptotic times t, localized in the complement of some forward, respectively, backward, light cone. Because of locality, they therefore commute with all operators which are localized in the interior of the respective cones. More specifically, let $\mathcal{O} \subset \mathbb{R}^4$ be the localization region of A and let $\mathcal{O}_\pm \subset \mathbb{R}^4$ be the two regions having a positive, respectively, negative, timelike distance from all points in $\mathcal{O}$. Then, for any operator B which is compactly localized in $\mathcal{O}_\pm$, respectively, one obtains $\lim_{t\to\pm\infty} A_t B\Omega = \lim_{t\to\pm\infty} BA_t\Omega = BP_1 A\Omega$. This relation establishes the existence of the limits

$$A^{in/out} = \lim_{t\to\mp\infty} A_t \qquad [23]$$

on the (by the Reeh–Schlieder property) dense sets of vectors $\{B\Omega : B \in \mathcal{A}(\mathcal{O}_\mp)\} \subset \mathcal{H}$. It requires some more detailed analysis to prove that the limits have all of the properties of a (smeared) free massless field, whose translates $x \mapsto A^{in/out}(x)$ satisfy the wave equation and have c-number commutation relations. From these free fields, one can then proceed to asymptotic creation and annihilation operators and construct asymptotic Fock spaces $\mathcal{H}^{in/out} \subset \mathcal{H}$ of massless particles and a corresponding scattering matrix as in the massive case. The details of this construction can be found in the original article, cf. Haag (1992).

It also follows from these arguments that the asymptotic fields $A^{in/out}$ of massless particles emanating from a region $\mathcal{O}$, that is, for which the underlying interpolating operators A are localized in $\mathcal{O}$, commute with all operators localized in $\mathcal{O}_\mp$, respectively. This result may be understood as an expression of Huygens' principle. More precisely, denoting by $\mathcal{A}^{in/out}(\mathcal{O})$ the algebras of bounded operators generated by the asymptotic fields $A^{in/out}$, respectively, one arrives at the following quantum version of Huygens' principle.

Theorem 4 *Consider a theory of massless particles as described above and let $\mathcal{A}^{in/out}(\mathcal{O})$ be the algebras generated by massless asymptotic fields $A^{in/out}$ with $A \in \mathcal{A}(\mathcal{O})$. Then*

$$\mathcal{A}^{in}(\mathcal{O}) \subset \mathcal{A}(\mathcal{O}_-)'$$

and [24]

$$\mathcal{A}^{out}(\mathcal{O}) \subset \mathcal{A}(\mathcal{O}_+)'$$

Here the prime denotes the set of bounded operators commuting with all elements of the respective algebras (i.e., their commutants).

Beyond Wigner's Concept of Particle

There is by now ample evidence that Wigner's concept of particle is too narrow in order to cover all particle-like structures appearing in quantum field theory. Examples are the partons which show up in nonabelian gauge theories at very small spacetime scales as constituents of hadrons, but which do not appear at large scales due to the confining forces. Their mathematical description

requires a quite different treatment, which cannot be discussed here. But even at large scales, Wigner's concept does not cover all stable particle-like systems, the most prominent examples being particles carrying an abelian gauge charge, such as the electron and the proton, which are inevitably accompanied by infinite clouds of ("on-shell") massless particles.

The latter problem was discussed first by Schroer, who coined the term "infraparticle" for such systems. Later, Buchholz showed in full generality that, as a consequence of Gauss' law, pure states with an abelian gauge charge can neither have a sharp mass nor carry a unitary representation of the Lorentz group, thereby uncovering the simple origin of results found by explicit computations, notably in quantum electrodynamics (Steinmann 2000). Thus, one is faced with the question of an appropriate mathematical characterization of infraparticles which generalizes the concept of particle invented by Wigner. Some significant steps in this direction were taken by Fröhlich, Morchio, and Strocchi, who based a definition of infraparticles on a detailed spectral analysis of the energy–momentum operators. For an account of these developments and further references, cf. Haag (1992).

We outline here an approach, originated by Buchholz, which covers all stable particle-like structures appearing in quantum field theory at asymptotic times. It is based on Dirac's idea of improper particle states with sharp energy and momentum. In the standard (rigged Hilbert space) approach to giving mathematical meaning to these quantities, one regards them as vector-valued distributions, whereby one tacitly assumes that the improper states can coherently be superimposed so as to yield normalizable states. This assumption is valid in the case of Wigner particles but fails in the case of infraparticles. A more adequate method of converting the improper states into normalizable ones is based on the idea of acting on them with suitable localizing operators. In the case of quantum mechanics, one could take as a localizing operator any sufficiently rapidly decreasing function of the position operator. It would map the improper "plane-wave states" of sharp momentum into finitely localized states which thereby become normalizable. In quantum mechanics, these two approaches can be shown to be mathematically equivalent. The situation is different, however, in quantum field theory.

In quantum field theory, the appropriate localizing operators L are of the form [18]. They constitute a (nonclosed) left ideal $\mathcal{L}$ in the C*-algebra $\mathcal{A}$ generated by all local operators. Improper particle states of sharp energy–momentum p can then be defined as linear maps $|\cdot\rangle_p : \mathcal{L} \to \mathcal{H}$ satisfying

$$U(x)|L\rangle_p = e^{ipx}|L(x)\rangle_p, \quad L \in \mathcal{L} \qquad [25]$$

It is instructive to (formally) replace L here by the identity operator, making it clear that this relation indeed defines improper states of sharp energy–momentum.

In theories of massive particles, one can always find localizing operators $L \in \mathcal{L}$ such that their images $|L\rangle_p \in \mathcal{H}$ are states with sharp mass. This is the situation covered in Wigner's approach. In theories with long-range forces there are, in general, no such operators, however, since the process of localization inevitably leads to the production of low-energy massless particles. Yet improper states of sharp momentum still exist in this situation, thereby leading to a meaningful generalization of Wigner's particle concept.

That this characterization of particles covers all situations of physical interest can be justified in the general setting of relativistic quantum field theory as follows. Picking g_t as in [4] and any vector $\Psi \in \mathcal{H}$ with finite energy, one can show that the functionals $\rho_t, t \in \mathbb{R}$, given by

$$\rho_t(L^*L) \doteq \int d^4x\, g_t(x_0)\, \langle \Psi, (L^*L)(x)\Psi \rangle, \quad L \in \mathcal{L} \qquad [26]$$

are well defined and form an equicontinuous family with respect to a certain natural locally convex topology on the algebra $\mathcal{C} = \mathcal{L}^*\mathcal{L}$. This family of functionals therefore has, as $t \to \pm\infty$, weak-* limit points, denoted by σ. The functionals σ are positive on $\mathcal{C}$ but not normalizable. (Technically speaking, they are weights on the underlying algebra $\mathcal{A}$.) Any such σ induces a positive-semidefinite scalar product on the left ideal $\mathcal{L}$ given by

$$\langle L_1 | L_2 \rangle \doteq \sigma(L_1^* L_2), \quad L_1, L_2 \in \mathcal{L} \qquad [27]$$

After quotienting out elements of zero norm and taking the completion, one obtains a Hilbert space and a linear map $L \mapsto |L\rangle$ from $\mathcal{L}$ into that space. Moreover, the spacetime translations act on this space by a unitary representation satisfying the relativistic spectrum condition.

It is instructive to compute these functionals and maps in theories of massive particles. Making use of relation [21] one obtains, with a slight change of notation,

$$\langle L_1 | L_2 \rangle = \int d\mu(p)\, \langle p | L_1^* L_2 | p \rangle \qquad [28]$$

where μ is a measure giving the probability density of finding at asymptotic times in state Ψ a particle of energy–momentum p. Once again, possible summations over different particle types and internal degrees of freedom have been omitted here. Thus,

setting $|L\rangle_p \doteq L|p\rangle$, one concludes that the map $L \mapsto |L\rangle$ can be decomposed into a direct integral of improper particle states of sharp energy–momentum, $|\cdot\rangle = \int_{\oplus} \mathrm{d}\mu(p)^{1/2}|\cdot\rangle_p$. It is crucial that this result can also be established without any *a priori* input about the nature of the particle content of the theory, thereby providing evidence of the universal nature of the concept of improper particle states of sharp momentum, as outlined here.

Theorem 5 *Consider a relativistic quantum field theory satisfying the standing assumptions. Then the maps $L \mapsto |L\rangle$ defined above can be decomposed into improper particle states of sharp energy–momentum p,*

$$|\cdot\rangle = \int_{\oplus} \mathrm{d}\mu(p)^{1/2}|\cdot\rangle_p \qquad [29]$$

where μ is some measure depending on the state Ψ and the respective time limit taken.

It is noteworthy that whenever the space of improper particle states corresponding to fixed energy–momentum p is finite dimensional (finite particle multiplets), then in the corresponding Hilbert space there exists a continuous unitary representation of the little group of p. This implies that improper momentum eigenstates of mass $m = (p^2)^{1/2} > 0$ carry definite (half)integer spin, in accordance with Wigner's classification. However, if $m = 0$, the helicity need not be quantized, in contrast to Wigner's results.

Though a general scattering theory based on improper particle states has not yet been developed, some progress has been made in Buchholz *et al.* (1991). There it is outlined how inclusive collision cross sections of scattering states, where an undetermined number of low-energy massless particles remains unobserved, can be defined in the presence of long-range forces, in spite of the fact that a meaningful scattering matrix may not exist.

Asymptotic Completeness

Whereas the description of the asymptotic particle features of any relativistic quantum field theory can be based on an arsenal of powerful methods, the question of when such a theory has a complete particle interpretation remains open to date. Even in concrete models there exist only partial results, cf. Iagolnitzer (1993) for a comprehensive review of the current state of the art. This situation is in striking contrast to the case of quantum mechanics, where the problem of asymptotic completeness has been completely settled.

One may trace the difficulties in quantum field theory back to the possible formation of superselection sectors (Haag 1992) and the resulting complex particle structures, which cannot appear in quantum-mechanical systems with a finite number of degrees of freedom. Thus, the first step in establishing a complete particle interpretation in a quantum field theory has to be the determination of its full particle content. Here the methods outlined in the preceding section provide a systematic tool. From the resulting data, one must then reconstruct the full physical Hilbert space of the theory comprising all superselection sectors. For theories in which only massive particles appear, such a construction has been established in Buchholz and Fredenhagen (1982), and it has been shown that the resulting Hilbert space contains all scattering states. The question of completeness can then be recast into the familiar problem of the unitarity of the scattering matrix. It is believed that phase space (nuclearity) properties of the theory are of relevance here (Haag 1992).

However, in theories with long-range forces, where a meaningful scattering matrix may not exist, this strategy is bound to fail. Nonetheless, as in most high-energy scattering experiments, only some very specific aspects of the particle interpretation are really tested – one may think of other meaningful formulations of completeness. The interpretation of most scattering experiments relies on the existence of conservation laws, such as those for energy and momentum. If a state has a complete particle interpretation, it ought to be possible to fully recover its energy, say, from its asymptotic particle content, that is, there should be no contributions to its total energy which do not manifest themselves asymptotically in the form of particles. Now the mean energy–momentum of a state $\Psi \in \mathcal{H}$ is given by $\langle \Psi, P\Psi \rangle$, P being the energy–momentum operators, and the mean energy–momentum contained in its asymptotic particle content is $\int \mathrm{d}\mu(p)p$, where μ is the measure appearing in the decomposition [29]. Hence, in case of a complete particle interpretation, the following should hold:

$$\langle \Psi, P\Psi \rangle = \int \mathrm{d}\mu(p)p \qquad [30]$$

Similar relations should also hold for other conserved quantities which can be attributed to particles, such as charge, spin, etc. It seems that such a weak condition of asymptotic completeness suffices for a consistent interpretation of most scattering experiments. One may conjecture that relation [30] and its generalizations hold in all theories admitting a local stress–energy tensor and local currents corresponding to the charges.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Dispersion Relations; Perturbation Theory and its Techniques; Quantum Chromodynamics; Quantum Field Theory in Curved

Spacetime; Quantum Mechanical Scattering Theory; Scattering, Asymptotic Completeness and Bound States; Scattering in Relativistic Quantum Field Theory: The Analytic Program.

Further Reading

Araki H (1999) *Mathematical Theory of Quantum Fields.* Oxford: Oxford University Press.

Buchholz D and Fredenhagen K (1982) Locality and the structure of particle states. *Communications in Mathematical Physics* 84: 1–54.

Buchholz D, Porrmann M, and Stein U (1991) Dirac versus Wigner: towards a universal particle concept in quantum field theory. *Physics Letters B* 267: 377–381.

Dybalski W (2005) Haag–Ruelle scattering theory in presence of massless particles. *Letters in Mathematical Physics* 72: 27–38.

Fredenhagen K, Gaberdiel MR, and Ruger SM (1996) Scattering states of plektons (particles with braid group statistics) in $(2 + 1)$ dinemsional quantum field theory. *Communications in Mathematical Physics* 175: 319–336.

Haag R (1992) *Local Quantum Physics.* Berlin: Springer.

Iagolnitzer D (1993) *Scattering in Quantum Field Theories.* Princeton, NJ: Princeton University Press.

Jost R (1965) *General Theory of Quantized Fields.* Providence, RI: American Mathematical Society.

Steinmann O (2000) *Perturbative Quantum Electrodynamics and Axiomatic Field Theory.* Berlin: Springer.

Streater RF and Wightman AS (1964) *PCT, Spin and Statistics, and All That.* Reading, MA: Benjamin/Cummings.

Scattering in Relativistic Quantum Field Theory: The Analytic Program

J Bros, CEA/DSM/SPhT, CEA Saclay, Gif-sur-Yvette, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction to the Analytic Structures of Quantum Field Theory

The importance of complex variables and of the concept of analyticity in theoretical physics finds one of its best illustrations in the analytic structure of relativistic quantum field theory (QFT). The latter have been investigated from several viewpoints in the last 50 years, according to the successive progress in QFT.

In the two main axiomatic frameworks of QFT, namely the one based on Wightman axioms (for a short presentation, *see* Dispersion Relations and also Axiomatic Quantum Field Theory) and the Haag, Kastler, and Araki theory of "local observables" (*see* Algebraic Approach to Quantum Field Theory), there are general justifications of analyticity properties for relevant "N-point structure functions" both in complexified spacetime variables and in complexified energy–momentum variables.

In the Wightman framework, relativistic quantum fields are operator-valued distributions $\Phi_j(x)$ on four-dimensional Minkowski spacetime that transform covariantly under a unitary representation of the Poincaré group in the Hilbert space of states. The basic quantities of QFT are (tempered) distributions on R^{4N} of the form $<\Psi, \Phi(x_1)\cdots\Phi(x_N)\Psi'>$, which depend on pairs of states Ψ, Ψ', belonging to the Hilbert space of the QFT considered: they can be called N-point structure functions of the field Φ "in x-space," namely in Minkowski spacetime (here, for brevity, we assume that the system is defined in terms of a single quantum field). In parallel, it is important to consider the Fourier transform $\tilde{\Phi}(p) = \int e^{ip\cdot x}\Phi(x)\,dx$ of the field in the Minkowskian energy–momentum space ($p\cdot x \doteq p_0x_0 - p\cdot x$ denoting the Minkowskian scalar product). The corresponding quantities $<\Psi, \tilde{\Phi}(p_1)\cdots\tilde{\Phi}(p_N)\Psi'>$, can then be called N-point structure functions of the field Φ "in p-space," namely in energy–momentum space.

In the algebraic QFT framework, each basic local observable B affiliated to a certain bounded region of spacetime $\mathcal{O}$ generates a Haag–Kastler–Araki quantum field $B(x)$ by the action of the translations of spacetime, namely $B(x) \doteq U(x)BU(x)^{-1}$. Here $U(x)$ denotes the unitary representation of the group of spacetime translations in the Hilbert space of states: $B(x)$ is affiliated to the translated region $\mathcal{O}(x) = \{y; y - x \in \mathcal{O}\}$. Then again one can consider N-point structure functions of the theory of the form $<\Psi, B(x_1)\cdots B(x_N)\Psi'>$ and $<\Psi, \tilde{B}(p_1)\cdots\tilde{B}(p_N)\Psi'>$.

To summarize the situation as it occurs in both cases, one can say the following:

1. A certain postulate of relativistic causality implies the analyticity of structure functions of a certain class, often called "Green functions," in the complex energy–momentum variables $k_j = p_j + iq_j$, in particular for purely imaginary energies.
2. "Stability properties" of the states Ψ, Ψ' such as a "bounded energy content" of these states imply

Notes:

the analyticity of the previous structure functions in the complex spacetime variables, in particular for purely imaginary times.

In both cases, analyticity is obtained as a basic property of the Fourier–Laplace transformation in several variables. Let V^+ denote the forward cone of the Minkowskian space ($V^+ \doteq -V^- \doteq \{x;\ x^2 \doteq x \cdot x > 0,\ x_0 > 0\}$) and let

$$\tilde{f}(p+iq) = \int_{V_a^+} e^{i(p+iq)\cdot x} f(x)dx \qquad [1]$$

$$g(x+iy) = (2\pi)^{-4} \int_{V_p^+} e^{-ip(x+iy)} \tilde{g}(p)dp \qquad [2]$$

be the associated reciprocal Fourier formulas, applied, respectively, to functions $f(x)$ with support contained in the translated forward cone $V_a^+ = -a + V^+$, $a \in V^+$ (or in its closure), and to functions $\tilde{g}(p)$ with support contained in the translated forward cone $V_p^+ = -P + V^+$, $P \in V^+$ of energy–momentum space (or in its closure). Then in view of the convergence properties of the previous integrals, one easily checks that $\tilde{f}(k)$ is holomorphic with possible exponential increase in the imaginary directions controlled by the bound $e^{q\cdot a}$ in the tube domain $\mathcal{T}^+ = \boldsymbol{R}^4 + iV^+$; similarly, $g(z)$ is holomorphic with an increase controlled by the exponential bound $e^{-y\cdot P}$ in the tube domain $\mathcal{T}^- = \boldsymbol{R}^4 + iV^-$.

On the one hand, for each N the structure functions $<\Psi, \tilde{\Phi}(p_1)\cdots\tilde{\Phi}(p_N)\Psi'>$ (or $<\Psi, \tilde{B}(p_1)\cdots\tilde{B}(p_N)\Psi'>$) have conical support properties of the previous type in the variables p_j, as a consequence of the relativistic shape of the energy–momentum spectrum. In both axiomatic frameworks, in fact, one postulates that there is a state of zero energy–momentum Ω, called the vacuum, and that the energy–momentum spectrum Σ, namely the joint spectrum of the generators P_μ of the Lie algebra of the group $U(x)$, is contained in the closure of V^+: this is the so-called spectral condition. A more refined assumption introduced for the requirements in particle physics is that Σ contains discrete parts localized on sheets of (mass-shell) hyperboloids inside V^+. These support properties in p-space imply that the corresponding inverse Fourier transforms $<\Psi, \Phi(x_1)\cdots\Phi(x_N)\Psi'>$ are boundary values of holomorphic functions in appropriate tube domains of the complex space variables $(z_1, \ldots, z_n)$.

On the other hand, in order to exhibit structure functions with conical support properties in x-space, one needs to build appropriate algebraic combinations of functions $<\Psi, \Phi(x_{j_1})\cdots\Phi(x_{j_N})\Psi'>$ with permuted arguments in order to take the benefit of the causality postulate, which is always formulated in terms of the commutator of two field operators. There are two versions of this postulate. In the Wightman framework, causality is expressed by the condition of local commutativity or microcausality,

$$[\Phi(x_1), \Phi(x_2)] = 0 \text{ for } (x_1 - x_2)^2 < 0 \qquad [3]$$

In the algebraic QFT framework, causality is expressed by a similar property in terms of any field $B(x)$ generated by a local observable $B \doteq B(0)$ affiliated to a region of spacetime enclosed in a given "double cone" $\mathcal{O}_b = V_b^+ \cap (-V_b^+)$. The corresponding expression of causality is

$$[B(x_1), B(x_2)] = 0 \quad \text{for } (x_1 - x_2) \notin (V_a^+ \cup (-V_a^+) \qquad [4]$$

for all a such that $a > 2b$.

So, we see that basically, causality and spectral condition generate analyticity respectively in complexified p-space and x-space. However, the situation is more intricate, since for each N there are always several holomorphic branches (two in the case $N=2$) in the variables $(z_1, \ldots, z_n)$ and also in the variables $(k_1, \ldots, k_n)$: each of these two sets is obtained essentially by permutations of the N vector variables. The important point is that these various branches can be seen to "communicate together," thanks to the existence of "coincidence regions" of their boundary values on the reals. Here again the roles played by causality and stability are symmetric (but inverted): while causality produces coincidence regions for the holomorphic functions in complex spacetime, spectral conditions produce coincidence regions for the holomorphic functions in complex energy–momentum space.

In view of a basic theorem of several complex variable analysis, called the edge-of-the-wedge theorem (see below in (4)), the two sets of communicating holomorphic branches actually define by mutual analytic continuation two holomorphic function $H_N^{\Psi,\Psi'}(k_1, \ldots, k_N)$ and $W_N^{\Psi,\Psi'}(z_1, \ldots, z_N)$ in respective domains $D_N^{\Psi,\Psi'}$ and $\Delta_N^{\Psi,\Psi'}$. However, these two primitive domains are not natural holomorphy domains (a phenomenon which is particular to complex geometry in several variables). The problem of finding their holomorphy envelopes, namely the smallest domains $\hat{D}_N^{\Psi,\Psi'}$ and $\hat{\Delta}_N^{\Psi,\Psi'}$ in which any functions holomorphic in the primitive domains can be analytically continued, is the idealistic purpose of what has been called the analytic program of axiomatic QFT. So, we see that there is an analytic program in x-space and there is an analytic program in p-space. In practice, except for the case $N=2$, where the complete answer is known, only a partial knowledge of the holomorphy envelopes has been obtained.

The analytic program in p-space, which is the only one to be described in the rest of this article, was often considered as physically more interesting, in view of the fact that it aims to establish analyticity properties of the scattering kernels on the complex mass shell. As a matter of fact, an important part of it concerns the derivation of the analyticity domains of dispersion relations for two-particle scattering amplitudes. This part is important from the historical viewpoint as well as from conceptual, physical, and pedagogical viewpoints (the reader may find it useful to first check the article Dispersion Relations, which illustrates how a structure function of the form $H_2^{\Psi,\Psi'}(k_1,k_2)$ can be used for that purpose with a suitable choice of the states Ψ and Ψ'). In the general development of the analytic program (in x-space as well as in p-space), it is recommended to consider the infinite set of structure functions $H_N \doteq H_N^{\Omega,\Omega}(k_1,\ldots,k_N)$ and $\mathcal{W}_N \doteq \mathcal{W}_N^{\Omega,\Omega}(z_1,\ldots,z_N)$ where Ω is the privileged vacuum state of the theory, in view of the fact that each of these sets characterizes entirely the field theory considered.

Before shifting to the analytic program in p-space, we would like to mention various points of interest of the analytic program in x-space:

1. Various results of this program have been extensively used for proving fundamental properties of QFT, such as the PCT-invariance theorem, the spin–statistics connection, etc. A good part of these can be found in the books by Streater and Wightman (1980) and by Jost (1965).
2. The functions H_N and $\mathcal{W}_N$ are holomorphic in their respective p-space and x-space "Euclidean subspaces." To make this clear, let us assume that a Lorentz frame has been chosen once for all; the linear subspace of complex spacetime (resp. energy–momentum) vectors of the form $z=(\mathrm{i}y_0,\boldsymbol{x})$ (resp. $k=(\mathrm{i}q_0,\boldsymbol{p})$) is called the "Euclidean subspace" of the corresponding complex Minkowskian space, in view of the fact that the quadratic form $z^2 \doteq z\cdot z=-(y_0^2+\boldsymbol{x}^2)$ (resp. $k^2 \doteq k\cdot k=-(q_0^2+\boldsymbol{p}^2)$) has a definite (negative) sign on that subspace. Then it has been established that (for each N) the restrictions of H_N and $\mathcal{W}_N$ to the corresponding N-vector Euclidean subspaces are the Fourier transforms of each other. This fact participates in the foundation of the Euclidean formulation of QFT or "QFT at imaginary times"; the latter has provided many important results in QFT, in particular for the rigorous study of field models (initiated by Glimm and Jaffe in the 1970s).
3. A more recent extension of QFT called thermal QFT (TQFT), which aims to study the behavior of quantum fields in a thermal bath, can be described in terms of a modified analytic program. In the latter, the spectral condition is replaced by the so-called KMS condition, which prescribes x-space analyticity properties of a particular type for the structure functions $\mathcal{W}_N$: it requires analyticity together with periodicity conditions with respect to imaginary times, the period being the inverse of the temperature (*see* Thermal Quantum Field Theory). The usual analytic structure for the theories with vacuum and spectral conditions is recovered in the zero-temperature limit.
4. In more recent investigations concerning quantum fields on (holomorphic) curved spacetimes, analyticity properties of the structure functions similar to those of thermal QFT can be established. This is the case in particular with de Sitter spacetime, for which a notion of "temperature of geometrical origin" is most simply exhibited.

In this article, an account of the general analytic program of axiomatic QFT in complex energy–momentum space will be presented; it will describe some of the methods which have been used for establishing analyticity properties of the N-point structure functions of QFT and corresponding properties of the $(n\to n')$-particle collision processes, for all n,n' such that $n\geq 2, n'\geq 2, n+n'=N$. (For a more detailed study, in particular concerning the microlocal methods, see the book by Iagolnitzer (1992)).

Concerning the important case $N=4$, this article gives complements to the results described in the article Dispersion Relations. In fact, the program allows one to justify other important analytic structures of the four-point functions and of two-particle scattering functions. They concern

- the field-theoretical basis of analyticity in the complexified variable of angular momentum, first introduced and developed in potential theory (Regge 1959);
- the Bethe–Salpeter (BS-) type structure (based on the additional postulate of asymptotic completeness), which is a relativistic field-theoretical generalization of the Lippmann–Schwinger structure of nonrelativistic scattering theory (for Schrödinger equations with Yukawa-type potentials).

The latter allows one to introduce the concept of composite particle in the field-theoretical framework (including bound states and unstable particles or "resonances") and also the concept of "Regge particle," thanks to complex angular momentum analysis.

Various Aspects of the General Analytic Program of QFT in Complex Energy–Momentum Space

The *N*-Point Structure Functions of QFT

It is proved in the Wightman QFT axiomatic framework that any QFT is completely characterized by the (infinite) sequence of its "N-point functions" or "vacuum expectation values" (also called "Wightman functions")

$$W_N(x_1,\ldots,x_N) \doteq < \Omega, \Phi(x_1)\cdots\Phi(x_N)\Omega >$$

which are tempered distributions on R^{4N} satisfying a set of general properties that can be split up into linear and nonlinear conditions. (This is known as the Wightman reconstruction theorem).

Linear conditions Each individual N-point function satisfies three sets of linear conditions which result, respectively, from:

1. *Poincaré invariance*: typically, for every Poincaré transformation g of Minkowski spacetime

$$W_N(x_1,\ldots,x_N) = W_N(gx_1,\ldots,gx_N)$$

in particular, the W_N are invariant under spacetime translations and therefore defined on the quotient subspace $R^{4(N-1)} \doteq R^{4N}/R^4$ of the differences $x_j - x_k$.

2. *Microcausality*: support conditions on commutator functions of the following form:

$$C^{(j\,j+1)}(x_1,\ldots,x_n) \doteq W_N(x_1,\ldots,x_j,x_{j+1},\ldots,x_N) - W_N(x_1,\ldots,x_{j+1},x_j,\ldots,x_N) = 0$$

in the region of R^{4N} defined by $(x_j - x_{j+1})^2 < 0$.

3. *Spectral condition*: support conditions on the Fourier transform $\tilde{W}_N(p_1,\ldots,p_N) = \delta(p_1+\cdots+p_N)\times\hat{w}_N(p_1,\ldots,p_{N-1})$ of W_N, which assert that $\hat{w}_N(p_1,\ldots,p_{N-1}) = 0$ if either one of the following conditions is fulfilled: $p_1+\cdots+p_j \notin \Sigma$, for $j = 1,\ldots,N-1$.

For each N, one can then construct a set of distributions $R_N^{(\alpha)}(x_1,\ldots,x_N)$, called "generalized retarded functions" (Araki, Ruelle, Steinmann, 1960 (see Iagolnitzer (1992, ref. [EGS])) which are appropriate linear combinations of multiple commutator functions built from W_N and multiplied by products of Heaviside step-functions $\theta(x_{j,0} - x_{k,0})$ of the differences of time coordinates. Each of these distributions $R_N^{(\alpha)}(x_1,\ldots,x_N)$ has its support contained in a convex salient cone C_α. This construction can be seen as a generalization of the decomposition [23] of the commutator $C_{\Psi,\Psi'}$ in the article Dispersion Relations. Then in view of the Laplace-transform theorem in several variables, the Fourier transform $\tilde{R}_N^{(\alpha)}(p_1,\ldots,p_N) = \delta(p_1+\cdots+p_N)\times \tilde{r}_N^{(\alpha)}([p]_N)$ is such that $\tilde{r}_N^{(\alpha)}([p]_N)$ is the boundary value of a holomorphic function $\tilde{r}_N^{(\alpha),(c)}([k]_N)$ defined in a tube $\mathcal{T}_\alpha = R^{4(N-1)} + i\tilde{C}_\alpha$. Here $[k]_N = [p]_N + i[q]_N$ belongs to a $4(N-1)$-dimensional complex linear space $M_N^{(c)}$: this is the set of complex vectors $[k]_N \doteq (k_1,\ldots,k_N)$ such that $k_1+\cdots+k_N = 0$. $\tilde{C}_\alpha$ is the dual cone of C_α in the real ($4(N-1)$-dimensional) $[q]_N$-space. Geometrically, each cone $\tilde{C}_\alpha$ is defined in terms of a certain "cell" of $[q]_N$-space which is defined by prescribing consistent conditions of the form $\varepsilon_J q_J \in V^+$ with $q_J = \sum_{j\in J} q_j$ and $\varepsilon_J = \pm 1$ for all proper subsets J of the set $\{1,2,\ldots,N\}$. This is the expression of the microcausality postulate (summarized in [3] or [4]) in complex energy–momentum space. Concerning the difference between the two formulations [3] and [4], one can see that there is no geometrical difference concerning the analyticity domains, but differences for the type of increase of the structure functions in their tube domains: in the case of [3], they are bounded by powers of the energy–momenta, while in the case of [4] they may have an exponential increase governed by factors of the type $e^{q\cdot a}$.

For each N, the linear space generated by all the distributions $\tilde{r}_N^{(\alpha)}([p]_N)$ is constrained by a set of linear relations (called Steinmann relations) which result from algebraic expressions of discontinuities of the following type, called (generalized) "absorptive parts,"

$$\tilde{r}_N^{(\alpha)}([p]_N) - \tilde{r}_N^{(\alpha')}([p]_N) = < \Omega, [\tilde{R}_{J_1}^{(\alpha_1)}([p]_{(J_1)}), \tilde{R}_{J_2}^{(\alpha_2)}([p]_{(J_2)})]\Omega > \quad [5]$$

for all pairs of adjacent cells $(\alpha,\alpha')_{(J_1,J_2)}$ in the following sense: α and α' only differ by changing the value of $\varepsilon_{J_1} = -\varepsilon_{J_2}$, (J_1,J_2) denoting any given partition of the set $\{1,2,\ldots,N\}$. In [5], the symbols $\tilde{R}_{J_i}^{(\alpha_i)}$ denote generalized retarded operators of lower order and the argument $[p]_{(J)}$ stands for the set of independent 4-momenta $\{p_j; j\in J\}$. Formula [5] may be seen as an N-point generalization of formula [26] of Dispersion Relations for the case when the state $\Psi = \Psi'$ is replaced by Ω.

Then by applying to [5] the same argument based on spectral condition as in the exploitation of eqn [26] in Dispersion Relations, one concludes that the two distributions $\tilde{r}_N^{(\alpha)}$ and $\tilde{r}_N^{(\alpha')}$ coincide on an open set $\mathcal{R}_{\alpha,\alpha'}$ of the form $p_{J_1}^2 = p_{J_2}^2 < M_{J_1}^2$, where $p_{J_1} \doteq \sum_{j\in J_1} p_j = -p_{J_2}$. It then follows from the general "oblique edge-of-the-wedge theorem" (Epstein, 1960; see below) that the two corresponding holomorphic functions $\tilde{r}_N^{(\alpha),(c)}([k]_N)$ and $\tilde{r}_N^{(\alpha'),(c)}([k]_N)$

have a common analytic continuation in the union of their tubes together with a certain complex "connecting set," bordered by $\mathcal{R}_{\alpha,\alpha'}$. Since this argument applies to all pairs $(\alpha,\alpha')_{(J_1,J_2)}$, the following important property holds (see Iagolnitzer (1992, refs. [B2], [EGS])):

Theorem 1

(i) *All the holomorphic functions* $\tilde{r}_N^{(\alpha),(c)}([k]_N)$ *admit a common analytic continuation* $H_N([k]_N)$, *called the N-point structure function (or Green function) of the given quantum field in complex energy–momentum space. It is holomorphic in a "primitive domain"* D_N *of* $M_N^{(c)}$, *which is the union of all tubes* $\mathcal{T}_\alpha$ *together with complex "connecting sets" bordered by all the coincidence regions* $\mathcal{R}_{\alpha,\alpha'}$ *defined previously.*

(ii) *For each N the complex domain* D_N *contains the whole Euclidean subspace* $\mathcal{E}_N$ *of* $M_N^{(c)}$, *which is the set of all complex vectors* $[k]_N=(k_1,\ldots,k_N)$ *such that* $k_j=(k_{j,0},\boldsymbol{k}_j); k_{j,0}=iq_{j,0}, \boldsymbol{k}_j=\boldsymbol{p}_j$ *for* $j=1,2,\ldots,N$. (*This Euclidean subspace depends on the choice of a given Lorentz frame in Minkowski spacetime.*)

Positivity Conditions The Hilbert space framework which underlies the axioms of QFT implies (an infinite set of) positivity inequalities on the N-point structure functions of the fields. As a typical example related to the previous formula [5] when $|J_1|=|J_2|=N/2$ (for N even), one can mention the positive-definiteness property of the absorptive parts for appropriate pairs of adjacent cells $(\alpha_1,\alpha_2=-\alpha_1)_{(J_1,J_2)}$, which simply expresses the positivity of the following Hilbertian squared norm:

$$\left|\int \bar{f}([p]_{(J_2)})f([p]_{(J_1)})[\tilde{r}_N^{(\alpha)}([p]_N) - \tilde{r}_N^{(\alpha')}([p]_N)]\mathrm{d}[p]_{(J_1)}\mathrm{d}[p]_{(J_2)}\right| = \left|\int f([p]_{(J_1)})\tilde{\mathrm{R}}_j^{(\alpha_1)}([p]_{(J_1)})\Omega > \mathrm{d}[p]_{(J_1)}\right|^2 \geq 0 \quad [6]$$

Scattering Kernels of General $(n\to n')$-Particle Collisions and General Reduction Formulas

The presentation of $(2\to 2)$-particle scattering kernels in the article Dispersion Relations can be generalized to arbitrary $(n\to n')$-particle collision processes, involving n incoming massive particles $(n\geq 2)$ and n' outgoing massive particles $(n'\geq 2)$. The big "scattering matrix" or "S-matrix" in the Hilbert space of states is the collection of all partial scattering matrices $S_{n,n'}$ or of the equivalent kernels $S_{n,n'}(p_{n,\mathrm{in}};p_{n',\mathrm{out}})$, defined by a straightforward generalization of formula [20] of the quoted article:

$$S_{n,n'}(\hat{f}_{n,\mathrm{in}},\hat{g}_{n',\mathrm{out}}) = \int_{\mathcal{M}_{n,n'}} \hat{f}_{n,\mathrm{in}}(p_{n,\mathrm{in}})\overline{\hat{g}_{n',\mathrm{out}}}(p_{n',\mathrm{out}}) \times S_{n,n'}(p_{n,\mathrm{in}};p_{n',\mathrm{out}})\mu_m^n(p_{n,\mathrm{in}})\mu_m^{n'}(p_{n',\mathrm{out}}) \quad [7]$$

Here we have considered for simplicity the case of collisions involving a single type of particle with mass m. In the arguments of the wave packets, the kernel, and the measures $(\mu_m^n,\mu_m^{n'})$, $p_{n,\mathrm{in}}$ and $p_{n',\mathrm{out}}$, respectively, denote the sets of incoming and outgoing 4-momenta $(p_1,\ldots,p_n)$ and $(p'_1,\ldots,p'_{n'})$ which all belong to the physical mass shell $H_m^+=\{p;p\in V^+,p^2=m^2\}$. By supplementing these mass-shell constraints with the relativistic law of conservation of total energy–momentum $p_1+\cdots+p_n=p'_1+\cdots+p'_{n'}$, one obtains the definition of the mass-shell manifold $\mathcal{M}_{n,n'}$ of $(n\to n')$-particle collision processes.

We shall reserve the name of scattering kernel (or scattering amplitude), denoted by $T_{n,n'}(p_{n,\mathrm{in}};p_{n',\mathrm{out}})$, to the so-called "connected component" of the S-matrix kernel $S_{n,n'}(p_{n,\mathrm{in}};p_{n',\mathrm{out}})$. By analogy with the definition of T in terms of S for the two-particle collision processes (*see* Dispersion Relations) $T_{n,n'}$ is defined by a recursive algorithm, which amounts to subtract from $S_{n,n'}$ all the components of the $(n\to n')$-collision processes that are decomposable into independent collision processes involving smaller number of particles, according to all admissible partitions of the numbers n and n'.

For any given N, let us consider all the "affiliated" scattering kernels $T_{n,n'}$ such that $n+n'=N$ and whose corresponding collision processes, also called "channels," are deduced from one another by the relevant exchange of incoming particles and outgoing antiparticles (e.g., $\Pi_1+\Pi_2+\Pi_3\to\Pi_4+\Pi_5+\Pi_6$, $\Pi_1+\Pi_2\to\overline{\Pi_3}+\Pi_4+\Pi_5+\Pi_6$, and $\Pi_1+\Pi_3\to\overline{\Pi_2}+\Pi_4+\Pi_5+\Pi_6$). There exist general reduction formulas according to which all these scattering kernels are restrictions to the mass-shell manifold $\mathcal{M}_{(N)}$ of appropriate boundary values of the (so-called) "amputated N-point function" $\hat{H}_N(k_1,\ldots,k_N)\doteq(k_1^2-m^2)\cdots(k_N^2-m^2)\times H_N(k_1,\ldots,k_N)$. More precisely, these reduction formulas can be written as follows:

$$T_{n,n'}(-p_{n,\mathrm{in}};p_{n',\mathrm{out}})_{|\mathcal{M}_{(N)}^{(\alpha)}} = \hat{H}_N^{(\alpha)}(p_1,\ldots,p_N)_{|\mathcal{M}_{(N)}^{(\alpha)}} \quad [8]$$

In the latter, $\hat{H}_N^{(\alpha)}$ denotes a certain boundary value of $\hat{H}_N$ on the reals: it is equal to a generalized retarded function $\tilde{r}_N^{(\alpha)}([p]_N)$ which depends in a specific way on a region of the mass shell, called $\mathcal{M}_{(N)}^{(\alpha)}$, in which

the $(n \to n')$-channel is considered. The important thing to be noted in [8] is the sign convention which attributes the notation $-p_j$ to the momentum of any incoming particle and therefore implies that p_j belongs to the negative sheet of hyperboloid $H_m^- \doteq -H_m^+$. This is the price to pay for expressing symmetrically the energy–momentum conservation law as $p_1 + p_2 + \cdots + p_N = 0$ (according to the QFT formalism), but it also displays, as a nice feature, the fact that all the affiliated scattering kernels $T_{n,n'}$ such that $n + n' = N$ are located on the various connected components of the mass shell $\mathcal{M}_{(N)}(p_j \in H_m;\ j = 1, 2, \ldots, N)$: the choice of the sheet H_m^- or H_m^+ of H_m is exactly linked to the incoming or outgoing character of the particle considered.

Remark 1 The reduction formulas are more usually expressed in terms of the Fourier transforms of the (connected parts of the) N-point amputated chronological functions $\tau_N([p]_N)$ (*see* Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools). As a matter of fact, the latter coincide with the boundary values $\tilde{\tau}_N^{(\alpha)}([p]_N)$ of H_N in the corresponding relevant regions $\mathcal{M}_{(N)}^{(\alpha)}$.

Remark 2 Coming back to the case of two-particle scattering amplitudes (i.e., $n = n' = 2, N = 4$), one can see that the general study presented here implies the consideration of the four-point function $H_4(k_1, k_2, k_3, k_4)$, which is a holomorphic function of three independent complex 4-momenta (since $k_1 + k_2 + k_3 + k_4 = 0$). In that case, the domain D_4 contains 32 tubes $\mathcal{T}_\alpha$ which are specified by triplets of conditions such as $q_1 \in V^+$, $q_2 \in V^+$, $q_3 \in V^+$, or $-q_1 \in V^+$, $q_1 + q_2 \in V^+$, $q_1 + q_3 \in V^+$, and those obtained by permutations of the subscripts $(1, 2, 3, 4)$ and also by a global substitution of the cone V^- to V^+.

Remark 3 The logical path from the postulates of QFT to the analyticity properties of two-particle scattering amplitudes that has been followed in the article Dispersion Relations can be seen as a partial exploitation of the general analyticity properties of the four-point function: one was specially interested there in the analyticity properties of H_4 in a single 4-momentum $k_1 = -k_3$ (at fixed real values of $p_2 = -p_4$). The "partial reduction formula" [27] of Dispersion Relations corresponds to the restriction of eqn [8] (for $N = 4$) to the linear submanifold $(p_1 = -p_3,\ p_2 = -p_4)$. It may also be worthwhile to stress the fact that, in spite of the exponential bounds on H_4 implied by the postulates of algebraic QFT, it has been possible to prove that the scattering function is still bounded by a power of s in its cut-plane (or crossing) domain; the dispersion relations with two subtractions are still justified in that case (Epstein, Glaser, Martin, 1969 (see Martin (1969, preprint))).

Off-Shell Character of D_N: Nontriviality of the Analytic Structure of the Scattering Kernels

One can now see that for each value of $N (N \geq 4)$ the situation created by complex geometry in the space $C^{4(N-1)}$ of $[k]_N$ is a mere generalization of the one described in a simple situation in the article Dispersion Relations.

1. There exists a fundamental $(3N - 4)$-dimensional complex submanifold, namely the complex mass shell $\mathcal{M}_{(N)}^{(c)}$ defined by the equations $k_j^2 = m^2$; $j = 1, \ldots, N$, which connects together the various real mass-shell components $\mathcal{M}_{(N)}^{(\alpha)}$ interpreted as the various physical regions of a set of affiliated $(n \to n')$-collision processes. The problem of proving the "analyticity of $(n \to n')$-scattering functions" thus amounts to constructing such holomorphic functions on the complex manifold $\mathcal{M}_{(N)}^{(c)}$, whose boundary values on the various real regions $\mathcal{M}_{(N)}^{(\alpha)}$ would reproduce the relevant scattering kernels $T_{n,n'}(-p_{n,\text{in}}; p_{n',\text{out}})$.
2. All the tubes $\mathcal{T}_\alpha$ which generate the primitive domain D_N are off-shell domains, namely their intersections with $\mathcal{M}_{(N)}^{(c)}$ are empty. This simply comes from the fact that the conditions $q_j \in V^\pm$ (included in their definition) and $k_j^2 = m^2 > 0$ are incompatible. One can also check that adding the coincidence regions $\mathcal{R}_{\alpha,\alpha'}$ between adjacent tubes does not improve the situation. However, one can state as a relevant scope the following program.
3. *Linear program* (so-called because it only relies on the linear conditions presented in the section "N-point structure functions of QFT"): find parts of the holomorphy envelope of D_N (possibly improved by the exploitation of the Steinmann relations) whose intersections with the complex mass shell $\mathcal{M}_{(N)}^{(c)}$ are nonempty. In the best case, show that such intersections can exist which connect two different regions $\mathcal{M}_{(N)}^{(\alpha)}$ together, which means "proving the crossing property between these two regions."
4. We shall see in the following that, except for the case $N = 4$, the results of this linear program have been rather disappointing as far as reaching the complex mass shell is concerned; however, other interesting analytic structures also coming from positivity conditions and from the additional postulate of asymptotic completeness have been investigated under the general name of

nonlinear program. The "synergy" created by the combination of these two programs remains, to a large extent, to be explored.

Results of Analytic Completion in the "Linear Program"

We can only outline here some of the geometrical methods which allow one to compute parts of the holomorphy envelopes of the domains D_N. One important method, which may be used after applying suitable conformal mappings, reduces to the following basic theorem.

The tube theorem *The holomorphy envelope of a "tube domain" of the form $T_B = \mathbf{R}^n + \mathrm{i}B$, where B is an arbitrary domain in $\mathbf{R}^n$ called the basis of the tube, is the convex tube $T_{\hat{B}} = \mathbf{R}^n + \mathrm{i}\hat{B}$, where $\hat{B}$ is the convex hull of B.*

The *opposite or oblique edge-of-the-wedge theorem* (Epstein 1960 (see Streater and Wightman (1980, ch. 2, ref. 18))) is a refined local version of the tube theorem, in which the basis B is of the form $B = C_1 \cup C_2$, where C_1, C_2 are two disjoint (opposite or nonopposite) cones with apex at the origin and where T_B is replaced by a pair of "local tubes" $(T_{C_1}^{(\mathrm{loc})}, T_{C_2}^{(\mathrm{loc})})$. Here the adjective "local" means that the real parts of the variables are confined in a given open set $\mathcal{U}$ (which can be arbitrarily small). The connectedness of T_B is now replaced by the consideration of any pair of functions (f_1, f_2) holomorphic in these local tubes whose boundary values on their common real set $\mathcal{U}$ coincide. The result is that f_1 and f_2 admit a common analytic continuation f in a local tube $T_C^{(\mathrm{loc})}$, where C is the convex hull of $C_1 \cup C_2$. In the case of opposite cones $(C_1 = -C_2)$, f is then analytic in the real set $\mathcal{U}$, while in the general oblique case f is only analytic in a complex connecting set bordered by $\mathcal{U}$ (namely a set which connects $T_{C_1}^{(\mathrm{loc})}$ and $T_{C_2}^{(\mathrm{loc})}$). There exists an extended version of the edge-of-the-wedge theorem in which the boundary values of f_1 and f_2 are only defined as distributions.

For simplicity, we shall just give a very rough classification of the type of results obtained. We shall distinguish:

- analyticity domains in the space of several (possibly all) variables: they can be of global type or of microlocal type, namely restricted to complex neighborhoods of real points;
- analyticity domains in special families of one-dimensional complex manifolds; and
- combinations of one-dimensional results which generate domains in several variables by a refined use of the tube theorem, called the *Malgrange–Zerner "flat tube theorem,"* or "flat edge-of-the-wedge theorem." In the latter, the local tubes $T_{C_1}^{(\mathrm{loc})}$ and $T_{C_2}^{(\mathrm{loc})}$ of f_1 and f_2 reduce to one-variable domains of the upper half-plane in separate variables $z_1 = x_1 + \mathrm{i}y_1, z_2 = x_2 + \mathrm{i}y_2$ but with a common range of real parts $(x_1, x_2) \in \mathcal{U}$. The data $f_1(z_1, x_2)$ and $f_2(x_1, z_2)$ have coinciding boundary values $(f_1(x_1, x_2) = f_2(x_1, x_2))$ in the limit $(y_1 \to 0, y_2 \to 0)$. The result is again the existence of a common analytic continuation to f_1 and f_2, which is a function of two complex variables $f(z_1, z_2)$ in the intersection of the quadrant $(y_1 > 0,\ y_2 > 0)$ with a complex neighborhood of $\mathcal{U}$. (Note that this result of complex analysis still holds when the real boundary values of the holomorphic functions have singularities, namely are only defined in the sense of distributions).

Global analyticity properties The following property (discovered by Streater for three-point functions) looks like an extension of the tube theorem. The holomorphy envelope of the union of two tubes $\mathcal{T}_\alpha, \mathcal{T}_{\alpha'}$ corresponding to adjacent pairs of cells $(\alpha, \alpha')_{(J_1, J_2)}$ together with a complex connecting set bordered by $\mathcal{R}_{\alpha,\alpha'} = \{[p]_N; p_{J_1}^2 < m_{J_1}^2\}$ is the convex hull $\mathcal{T}_{\alpha,\alpha'}$ of the union of these tubes minus the following analytic hypersurface σ_{J_1} which can be called "a cut": $\sigma_{J_1} = \{[k]_N: k_{J_1}^2 = m_{J_1}^2 + \rho, \rho \geq 0\}$. The interest of this result (although it remains by itself an off-shell result) is that it can generate larger cut-domains by additional analytic completions, which may have intersections with the complex mass shell (see below for the case $N = 4$).

Microlocal analyticity properties In the case of the four-point function $\hat{H}_4$, it is possible to consider opposite cut-domains of the previous type, for which $\sigma_{J_1} = \sigma_{\{1,2\}}$ is the energy-cut of the channel $(1, 2 \to 3, 4)$, and for which the spectral conditions prescribe an "edge-of-the-wedge situation" in the neighborhood of the corresponding mass-shell component $\mathcal{M}_{(1,2\to3,4)}$. The result is that H_4 is proved to be holomorphic in a full complex cut-neighborhood of $\mathcal{M}_{(1,2\to3,4)}$ in the ambient complex energy–momentum space. The intersection of this local domain with the complex mass shell $\mathcal{M}_{(4)}^{(\mathrm{c})}$ is of course a full complex cut-neighborhood of $\mathcal{M}_{(1,2\to3,4)}$ in $\mathcal{M}_{(4)}^{(\mathrm{c})}$, and this proves that the corresponding scattering amplitude is the boundary value of an analytic scattering function defined as the restriction $\hat{F}(s,t) \doteq \hat{H}_{4|\mathcal{M}_{(4)}^{(\mathrm{c})}}$ of $\hat{H}_4$: it is holomorphic in a domain of complex (s,t) space deprived from the s–cut.

In the general case $N > 4$, the results are less spectacular, although a more sophisticated microlocal method involving a "generalized edge-of-the-wedge

theorem" has been applied. This method, which was one of the three methods at the origin of the chapter of mathematics called microlocal analysis (the other two being Hörmander's "analytic wave-front" method and Sato's "microfunctions" method) is based on a local version of the Fourier–Laplace transformation called the FBI transformation (see, e.g., the book on "hypo-analytic structures" by Treves (1992) and in the present context the article "Causality and local analyticity" by Bros and Iagolnitzer (1973) (see Iagolnitzer (1992, ref. [BI1]))).

A first positive result (obtained at first by Hepp in 1965) is the fact that the various real boundary values of $\hat{H}_N$ admit well-defined restrictions as tempered distributions on the corresponding (real) mass shell $\mathcal{M}_{(N)}$; this result is in fact crucial for the rigorous proof of general reduction formulas. However, (according to Bros, Epstein, Glaser, 1972 (see Iagolnitzer (1992, ref. [BEG2])) the local existence of an analytic scattering function in $\mathcal{M}^{(c)}_{(N)}$ is not ensured at all points of the mass shell, but only in certain regions. A rather favourable situation still occurs for $(2 \to 3)$-particle collision amplitudes (i.e., for $N=5$), but in the general case there are large regions of the mass shell where it is only possible to prove (at least in this linear program) that the amplitude is a sum of a limited number of boundary values of analytic functions, defined in local domains of $\mathcal{M}^{(c)}_{(N)}$ (see in this connection, Iagolnitzer (1992)).

Analyticity at fixed total energy in momentum transfer variables A remarkably simple situation had already been exploited before the general analysis of H_N leading to Theorem 1 was carried out. It is the section of the domain of the N-point function in the space of the "initial relative 4-momentum" $k=(k_1-k_2)/2$ of the s-channel with initial 4-momenta (k_1,k_2), when the total energy–momentum $P=-(k_1+k_2)$ with $P^2=s$ is kept fixed and real. The remaining 4-momenta $p_3,\ldots,p_N$ such that $p_3+\cdots+p_N=P$ are also kept fixed and real. Consider the case when P is (positive) timelike and such that $s \geq 4m^2$. Then it can be seen that one obtains analyticity of (a certain "1-vector restriction" of) H_N with respect to the vector variable k in the union of the two opposite tubes $\mathcal{T}^+=\boldsymbol{R}^4+iV^+$, $\mathcal{T}^-=\boldsymbol{R}^4+iV^-$. Moreover, an edge-of-the-wedge situation holds in view of the spectral coincidence region of the form $k_1^2=(-P/2+k)^2<M_1^2$, $k_2^2=(-P/2-k)^2<M_2^2$. The corresponding holomorphy envelope is given by a Jost–Lehmann–Dyson domain (*see* Dispersion Relations), whose section by the complex mass shell $k_1^2=k_2^2=m^2$ turns out to give a "spherical tube domain" of the form $\{k;\, k=p+iq;\, k.P=0,\, k^2=-s/4+m^2;\, |q^2|<b^2\}$. The $(2 \to N-2)$-particle scattering kernel is therefore the boundary value of a scattering function holomorphic in the previous spherical domain of complex k-space. In the special case of the two-particle scattering amplitude $F(s,t)$, one checks that the previous domain yields for each $s, s \geq 4m^2$, an ellipse of analyticity for $\hat{F}(s,t)$ in the t-plane with foci at $t=0$ and $u=4m^2-s-t=0$; this ellipse is called the Lehmann ellipse. (We have considered for simplicity the case of a single type of particle with mass m and two-particle threshold at $2m$.) In fact, the squared momentum transfer t is equal to $(k-k')^2$, if $k'=(k_3-k_4)/2$ denotes the "final relative momentum" of the s-channel, which was here taken to be fixed and real. Moreover, by a similar argument the corresponding absorptive part, namely the discontinuity across the s-cut of the scattering amplitude, can be shown to be holomorphic in a larger ellipse with the same foci called the large Lehmann ellipse.

It is interesting to compare the previous result with the one that one obtains when the fixed vector P is chosen to be spacelike, namely when s has a negative, namely "unphysical" value with respect to the distinguished channel $(1,2 \to 3,4)$. For that case, the exploitation of the primitive domain D_4 shows that for all negative (unphysical) values $\zeta_i=k_i^2<0$; $i=1,2,3,4$, of the squared mass variables, the function $\hat{H}_4$ is holomorphic in a cut-plane of the variable t, where the cuts are the t-cut $(t=4m^2+\rho,\ \rho \geq 0)$ and the u-cut $(u=4m^2-s-t=4m^2+\rho',\ \rho' \geq 0)$. This cut-plane has of course to be compared with the off- shell cut-plane domain Δ_ζ at the basis of the proof of dispersion relations (*see* Dispersion Relations). Here, however, the choice of the squared momentum transfer t as the variable of analyticity allows one to shift to another interpretation in terms of the concept of angular momentum.

Analyticity in the complex angular momentum variable In all the situations previously considered for the case $N=4$, one can see that at fixed real values of the squared energy s and of the squared masses $\zeta=\{\zeta_i;\, i=1,2,3,4\}$, the complex initial and final relative 4-momenta k and k' have directions which vary on the complexified sphere $S^{(c)}$. Moreover, the corresponding restriction of $\hat{H}_4$ to that sphere turns out to be always well defined and analytic on the real part of that sphere: it therefore defines a kernel on the sphere, which, in view of Poincaré invariance, is invariant under the rotations and therefore admits a convergent expansion in Legendre polynomials. Let us call $h_\ell(s;\zeta)$ the corresponding sequence of Legendre coefficients.

In the first case considered above, this sequence coincides (all ζ_i being equal to m^2) with what the physicists call the set of partial waves $f_\ell(s)$ of the scattering amplitude. The analyticity of $\hat{H}_4$ on a complex spherical tube of $S^{(c)}$, namely of $\hat{F}(s,t)$ in the Lehmann ellipse, is then equivalent to a certain exponential decrease property with respect to ℓ of the sequence of partial waves.

In the second case, where s and the ζ_i are negative, it can be seen that the sphere S describes 4-momentum configurations which all belong to a certain Euclidean subspace $\mathcal{E}_4$ of $M_4^{(c)}$. But this situation is much more favourable from the viewpoint of analyticity, since $\hat{H}_4$ can be seen to be holomorphic on the full complex submanifold $S^{(c)} \times S^{(c)}$ minus two sets σ_t and σ_u which correspond to the t- and u-cuts of the complex t-plane. Then this larger analyticity property turns out to be equivalent to the fact that the sequence $h_\ell(s;\zeta)$ admits an interpolation $\tilde{H}(\lambda;s;\zeta)$ holomorphic in a certain half-plane of the form $\mathrm{Re}\,\lambda > \ell_0$ such that for all integers $\ell > \ell_0$ one has: $\tilde{H}(\ell;s;\zeta) = h_\ell(s;\zeta)$. The value of ℓ_0 is linked to the power bound at large momenta that must be satisfied by $\hat{H}_4$ as a consequence of the temperateness property included in the Wightman axiomatic framework (Bros and Viano 2000).

Of course, this nice analytic structure in a complex angular momentum variable could extend to the set of physical partial waves $f_\ell(s)$ if one could establish the analytic continuation of $\hat{F}(s,t)$ in a cut-plane of t containing the Lehmann ellipses, but this seems out of the possibilities at least of the linear program.

The "Nonlinear Program" and Its Two Main Aspects

The extension of the analyticity domains by positivity and the derivation of bounds by unitarity Positivity conditions of the form [6] have been extensively applied to the case $N=4$ (namely for subsets J with two elements). The main result (Martin 1969) consists in the possibility of differentiating the forward dispersion relations with respect to t and, as a consequence, to enlarge the analyticity domain in t at fixed s: the Lehmann ellipse, whose size shrinks to zero when s tends to infinity, can then be replaced by an ellipse (i.e., the Martin ellipse) whose maximal point $t = t_{\max} > 0$ is fixed when s goes to infinity. This justifies the extension of dispersion relations in s to positive values of t; then in a second step the use of unitarity relations for the partial waves allows one to obtain Froissart-type bounds on the scattering amplitudes (see Martin (1969)).

Asymptotic completeness and BS-type structural analysis The BS equations have been at first introduced as identities of formal series in the perturbative approach of QFT, and the idea of considering such identities as exact equations having a conceptual content in the general axiomatic framework of QFT has been introduced and developed by Symanzik in 1960. However, it took a long time before its integration in the analytic program of QFT (Bros 1970 (see Iagolnitzer (1992, ref. [B1])). These developments belong to the nonlinear program since they rely on quadratic integral equations between the various N-point functions, which express the postulate of asymptotic completeness via the use of appropriate reduction formulas.

For brevity, the general set of BS-type equations for the N-point functions with $N > 4$ will not be presented. The simplest BS-type equation, which concerns the four-point function, can be written as follows:

$$\hat{H}_4(K;k,k') = B(K;k,k') + (\hat{H}_4 \circ_s B)(K;k,k') \quad [9]$$

where

$$(\hat{H}_4 \circ_s B)(K;k,k') = \int_\Gamma \hat{H}_4(K;k,k'')B(K;k'',k')G\left(\frac{K}{2}+k''\right) \times G\left(\frac{K}{2}-k''\right) \mathrm{d}_4k'' \quad [10]$$

In the latter, the s-channel is privileged, with $s=K^2, K=-(k_1+k_2)$; $\hat{H}_4$ is seen as a K-dependent kernel (k and k' are the initial and final relative 4-momenta already defined), and the new object B to be studied is also a K-dependent kernel. The function $G(k)$ is holomorphic in k^2 in a cut-plane except for a pole at $k^2=m^2$ which plays a crucial role. (It is essentially the "propagator" or two-point function of the field theory considered). Apart from pathologies due to the Fredholm alternative, the correspondence between $\hat{H}_4$ and B is one-to-one, but the peculiarity concerns the integration cycle Γ of [10]: it is a complex cycle of real dimension 4, which coincides with the Euclidean space of the vector variable k'' when all the 4-momenta are Euclidean, and can always be distorted inside the analyticity domain of $\hat{H}_4$ together with the external variables. The exploitation of the Fredholm equation in complex space with "floating integration cycles" then implies that B is holomorphic at least in the primitive domain of $\hat{H}_4$.

An important geometrical aspect of the integration on the cycle Γ in [10] is the fact that this cycle is "pinched" between the pair of poles of the functions G when K^2 tends to its threshold value ($s=4m^2$).

The type of mathematical concept encountered here is closely related to those used in the study of analyticity properties and Landau singularities of the Feynman amplitudes in the perturbative approach of QFT (in this connection, see the books by Hwa and Teplitz (1966) and by F Pham (2005) and references therein).

The first basic result is that it is equivalent for $\hat{H}_4$ to satisfy an asymptotic completeness equation in the pure two-particle region $4m^2 < s < 9m^2$ and for B to satisfy the following property called two-particle irreducibility: B satisfies dispersion relations in s such that the s-cut begins at the three-particle threshold: $s = 9m^2$.

The consequence of this extended analyticity property of B is that it generates the following type of analyticity properties for $\hat{H}_4$:

1. *The existence of a two-sheeted analytic structure* for $\hat{H}_4$ over a domain of the s-plane containing the interval $4m^2 \leq s < 9m^2$, with a square-root-type branch point at the threshold $s = 4m^2$.
2. *Composite particles.* There exists a Fredholm-type expression

$$\hat{H}_4(K; k, k') = \frac{N(K; k, k')}{D(K^2)} \qquad [11]$$

where N and D are expressed in terms of B via Fredholm determinants, which shows that in its second sheet $\hat{H}_4$ may have poles in $s = K^2$, generated by the zeros of D. These poles are interpreted as resonances or unstable particles. The generation of real poles in the first sheet (i.e., bound states) is also possible under special spectral assumptions of QFT.
3. *Complex angular momentum diagonalization of BS-type equations* (Bros and Viano 2000, 2003). The operation $\circ_s$ in the BS-type equation [9] contains not only an integration over squared-mass variables, but also a convolution product on the sphere S; the latter is transformed into a product by the Legendre expansion of four-point functions described previously in the subsection "Analyticity in the complex angular momentum variable." As a result, there is a partially diagonalized transform of eqn [9] in terms of the functions $\tilde{H}(\lambda; s; \zeta)$ and $\tilde{B}(\lambda; s; \zeta)$, which allows one to write a Fredholm formula similar to [11], namely

$$\tilde{H}(\lambda;\ s; \zeta) = \frac{\tilde{N}(\lambda;\ s; \zeta)}{\tilde{D}(\lambda; s)} \qquad [12]$$

Then under suitable increase assumptions on B, there may exist a half-plane of the form $\mathrm{Re}\,\lambda > \ell_1$ (with $\ell_1 < \ell_0$) such that $\tilde{H}(\lambda; s; \zeta)$ admits poles in the joint variables λ and s, corresponding to the concept of Regge particle: the composite particles introduced in (2) might then be integrated in the Regge particle, although they manifest themselves physically only for integral values ℓ of λ with the corresponding spin interpretation. Of course, this scenario is by no means proven to hold in the general analytic program of QFT, but we have seen that the relevant "embryonary structures" are conceptually built-in, so that the phenomenon might hopefully be produced in a definite quantum field model.
4. *Byproducts of BS-type structural analysis for* $N = 5$ *and* $N = 6$. Relativistic exact structural equations for $(3 \to 3)$-particle collision amplitudes, which generalize the Faddeev structural equations of nonrelativistic potential theory, have been shown to be valid in the energy region of "elastic" collisions (i.e., with total energy bounded by $4m$); relevant Landau singularities of tree diagrams and triangular diagrams have been exhibited as a by-product in this low-energy region (Bros, and also Combescure, Dunlop in two-dimensional field models, 1981 (see Iagolnitzer (1992, refs. [B3], [B4], [CD]))). Moreover, crossing domains on the complex mass shell for $(2 \to 3)$-particle collision amplitudes have been obtained (Bros 1986 (see Iagolnitzer (1992, ref. [B1]))) by conjointly using $(N = 5)$ BS-type equations together with analytic completion properties (*see*, e.g., the "Crossing lemma" in Dispersion Relations).

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Dispersion Relations; Scattering, Asymptotic Completeness and Bound States; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Thermal Quantum Field Theory.

Further Reading

Bros J and Viano GA (2000) Complex angular momentum in general quantum field theory. *Annales Henri Poincaré* **1**: 101–172.

Bros J and Viano GA (2003) Complex angular momentum diagonalization of the Bethe–Salpeter structure in general quantum field theory. *Annales Henri Poincaré* 4: 85–126.

Haag R (1992) *Local Quantum Physics.* Berlin: Springer.

Hwa RC and Teplitz VL (1966) *Homology and Feynman integrals.* New York: Benjamin.

Iagolnitzer D (1992) *Scattering in Quantum Field Theories: The Axiomatic and Constructive Approaches*, Princeton Series in Physics. Princeton: Princeton University Press.

Jost R (1965) *The General Theory of Quantized Fields.* Providence: American Mathematical Society.

Martin A (1969) *Scattering Theory: Unitarity, Analyticity and Crossing*, Lecture Notes in Physics. Berlin: Springer.

Pham F (2005) *Intégrales singulières*. Paris: EDP Sciences/CNRS Éditions.

Streater RF and Wightman AS (1964, 1980) *PCT, Spin and Statistics, and all that*. Princeton: Princeton University Press.

Treves F (1992) *Hypoanalytic Structures*. Princeton: Princeton University Press.

Symmetries and Conservation Laws

L H Ryder, University of Kent, Canterbury, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction: Spacetime Symmetries

Symmetries have played, and continue to play, an important role in fundamental physics, but the part they play is today seen as more complicated and many-sided than it was in the early days of particle physics, just after the Second World War. The area in which symmetries have had their most dramatic consequences is elementary particle physics, or high-energy physics, and the majority of this article is concerned with this subject. The article concludes with some observations about symmetries and conservation laws in general relativity.

In the early days, considerations of symmetry were almost limited to Lorentz transformations: we begin by reviewing this crucially important topic. Invariance of the laws of nature under *translations* in space and time are actually necessary for the existence of science itself; if experiments did not yield the same results today and tomorrow, and in Paris and Moscow and on the Moon, then in effect there would be no laws of nature. Almost as strong

Notes:

a statement could be made about invariance under *rotations*; if space were not isotropic, experimental results would depend on which direction the apparatus was aligned in, and again any laws would be extremely hard to find. Turning to the question of *motion*, Newton and Galileo realized that the laws of dynamics are the same in all inertial frames in relative motion. In the Newton–Galileo scheme, the rule for relating the space and time coordinates of two frames of reference is (for relative motion along the common x-axis)

$$x' = x - vt, \qquad t' = t \tag{1}$$

This *principle of relativity* was reaffirmed by Einstein, but with the crucial modification that the rules for relating coordinates in two frames are given by Lorentz transformations, so that [1] is replaced by

$$x' = \gamma(x - vt), \qquad t' = \gamma\left(t - \frac{vx}{c^2}\right) \tag{2}$$

Time is absolute in [1] but relative in [2]. Einstein was of course motivated by the fact that Maxwell's equations are covariant under Lorentz transformations, but not under Newton–Galileo ones.

The above considerations reveal that the laws of nature should be covariant under ten types of transformation: three translations in space, one in time, three parameters (angles) for rotations and three velocities. These transformations together form a group, the inhomogeneous Lorentz, or Poincaré group. It is a nonabelian group whose ten generators correspond to 4-momentum, angular momentum, and Lorentz boosts. The seminal work on the significance of this group in fundamental physics is that of Wigner in 1939. Assuming that the states of fundamental quantum systems (particles, atoms, molecules) form the basis states for *representations* of this group, these entities are described by two quantities, mass and spin. Spin, moreover, which was already familiar from earlier investigations in quantum physics, was described by the rotation group (SU(2), which is homomorphic to SO(3)) only for states with timelike momentum. For photons, for example, with null momentum, spin is described by the (noncompact) Euclidean group in the plane, with the consequence that there are only two polarization states for this massless particle.

Noether's theorem provides the crucial link between symmetries and *conservation laws*, via the principle of least action. Noether showed that the invariance of the action under a continuous symmetry operation implied the existence of a conserved quantity. The conserved quantities corresponding to invariance under translation in space and time are momentum and energy; conservation of angular momentum follows from invariance under rotations and invariance under Lorentz transformations gives rise to conservation of motion of the center of mass.

Gauge Theories: Electromagnetism and Yang–Mills Theories

A quantity whose conservation has been well known for a long time is electric charge. The question may then be asked: invariance under what symmetry gives rise to conservation of electric charge? A classical complex field has the Lagrangian density

$$L = (\partial_\mu \phi)(\partial^\mu \phi^*) - m^2 \phi^* \phi \tag{3}$$

which is invariant under

$$\phi \rightarrow \exp(-\mathrm{i}Q\,\Lambda)\phi \tag{4}$$

Λ being the parameter for the transformation. Noether's theorem then yields conservation of Q, interpreted as electric charge. With Λ a constant, as above, the Lagrangian possesses a "global" symmetry. This becomes a "local" symmetry when Λ becomes space and time dependent, $\Lambda(r,t)$ or $\Lambda(x^\mu)$. In that case, however, the Lagrangian [3] is no longer invariant under [4], because of the derivative terms. To preserve invariance an extra field A_μ must be introduced, so that [4] then becomes

$$\begin{aligned} \phi &\rightarrow \exp(-\mathrm{i}Q\,\Lambda(x^\mu))\phi \\ A_\mu &\rightarrow A_\mu + \frac{1}{Q}\partial_\mu \Lambda \end{aligned} \tag{5}$$

and the Lagrangian acquires extra terms, involving A_μ. The field A_μ is called a *gauge field* and is identified with the electromagnetic potential. The transformation [5] is called a *gauge transformation*, and since the phase factor $\exp(-\mathrm{i}Q\,\Lambda)$ may be regarded as a unitary 1×1 matrix, we have here a theory with U(1) gauge invariance, which describes electromagnetism and conservation of charge.

The notion of isospin had been introduced by Heisenberg in 1932. Isospin (then called isotopic spin) was a vector-like quantity conserved in strong (nuclear) interactions. Yang and Mills in 1954 made the pioneering suggestion that isospin conservation could also be recast as a gauge theory, by enlarging the U(1) group of electromagnetism to SU(2) (corresponding to rotations in "isospin space"), and at the same time treating the rotation angles as functions of spacetime. Then, eqn [4] will change: if

for example ψ is an isospinor field, then local isospin rotations are given by

$$\psi(x) \rightarrow \exp\left\{-i\frac{\tau}{2}\cdot\theta(x)\right\}\psi(x) = U(x)\psi(x) \quad [6]$$

where τ are the Pauli matrices: $\tau/2$ are the generators of SU(2). The gauge field then has three components $A^i_\mu (i = 1, 2, 3)$ which may be written as a matrix

$$A_\mu = A^i_\mu \frac{\tau^i}{2}$$

transforming as

$$A_\mu \rightarrow A'_\mu = U(x)A_\mu\, U^{-1}(x) - \frac{i}{g}(\partial_\mu\, U(x))U^{-1}(x) \quad [7]$$

where g is the coupling constant, analogous to electric charge. The problem with this idea was that the isospin gauge field, analogous to the photon in electrodynamics, should, like the photon, be massless and have polarization states ± 1 (commonly, but inaccurately – see the work of Wigner (1939) – called spin 1); whereas the Yukawa particle, identified as the π meson, was massive and had spin 0, so could not act as the isospin gauge field.

The Yang–Mills idea really came into its own with the standard model (SM) of particle physics. This (gauge) model has an invariance group $SU(2) \otimes U(1) \otimes SU(3)$, the first two groups corresponding to electroweak interactions (a unification of weak interactions and electromagnetism) and the final SU(3) to quantum chromodynamics (QCD), the gauge theory describing quark interactions, which "glues" them together to make hadrons – protons, neutrons, pions, etc. This model is a dramatically successful one. The QCD sector of the theory requires essentially no further elaboration on the Yang–Mills idea than replacing the group SU(2) by SU(3). This is a straightforward matter of replacing the generators $\tau/2$ of SU(2) with the eight generators (3×3 matrices) of SU(3). $U(x)$ then also becomes a 3×3 matrix. The three degrees of freedom are the three quark "colors," for which there is good experimental evidence, and the gluons, the quanta of the gauge fields, are indeed massless and have good experimental support. In the electroweak sector, however, the gauge fields, the W and Z bosons, were found with the predicted masses of 80.3 and 91.2 GeV respectively (the proton mass, for comparison, is 0.98 GeV). They are certainly not massless, as the straightforward Yang–Mills theory would require, and the explanation for this requires the introduction of the concept of spontaneous symmetry breaking.

Spontaneous Symmetry Breaking

The general idea of spontaneous symmetry breaking is that the *vacuum* – the state of lowest energy – is not invariant under the symmetry in question. A simple and common illustration is a pencil balanced vertically on its tip on a horizontal plane. The pencil is in unstable equilibrium but the system has a symmetry under rotations in the plane about the axis coincident with the pencil. Eventually, the pencil will fall into its lowest-energy state (vacuum), lying on the table in some direction – and the rotational symmetry is then lost. In fact, under rotations the actual lowest-energy (vacuum) state will be changed into another such state. There is a *degenerate vacuum*.

A similar scenario may be constructed in a complex scalar field theory. Consider such a theory with a Lagrangian given by

$$L = (\partial_\mu\phi)(\partial^\mu\phi^*) - m^2\phi^*\phi - \lambda(\phi^*\phi)^2 \quad [8]$$

that is, with a potential energy function given by

$$V(\phi, \phi^*) = m^2\phi^*\phi + \lambda(\phi^*\phi)^2 \quad [9]$$

where m is the mass of the field (quantum) and λ is the coupling of its self-interaction. The ground state is obtained by minimizing V, hence $\partial V/\partial\phi = 0$, giving (assuming that $m^2 > 0$) a minimum at $\phi = \phi^* = 0$. If, however, $m^2 < 0$, there is a local maximum at $\phi = 0$ and a minimum at $|\phi|^2 = -m^2/2\lambda > 0$. In quantum theory language, the vacuum expectation value $<0|\phi|0>$ of the field is nonzero. Goldstone showed that this implied the presence of a massless scalar particle – a Goldstone boson. There was some interest in this result in particle physics, where the hypothesis of "partial conservation of the axial vector current" (PCAC) might result in a Goldstone boson that could be identified with the pion; although not massless, the pion is the lightest hadron, so "almost" massless.

Higgs analyzed what happens to the Goldstone model if electromagnetism is included. The Lagrangian [8] is invariant under the global transformation [4], but if this is made local, as in [5], a gauge field must be introduced and it is found that the massless Goldstone boson disappears and the massless gauge field (photon) becomes massive. Thus, spontaneous symmetry breaking of a gauge theory results in the appearance of a massive, rather than massless, gauge particle. (It is relevant to remark that a massless photon possesses two polarization states, but a massive one possesses three, so the number of spin-polarization states is preserved – the massless photon "eats" the Goldstone boson and becomes massive.) The Higgs model was generalized to the

case of a nonabelian symmetry group by Guralnik, Hagen, and Kibble and invoked by Weinberg in his 1971 model for the electroweak interaction in which the gauge quanta were massive.

Higgs' work was motivated by the theory of superconductivity, where the Meissner effect (expulsion of magnetic flux from a superconductor), when relativistic, implies that the effective mass of a photon in a superconductor is nonzero – this is, the "reason" that the flux does not penetrate. In the theory of Bardeen, Cooper, and Schrieffer (BCS), a superconductor is described by an effective scalar field, a composite of electron pairs (though paired in momentum space rather than coordinate space), and this provides a physical analogy with the model above. The SM of particle physics postulates a Higgs scalar field analogous to the BCS composite scalar field. If this field exists, Higgs particles should also exist, but they have not yet been found. This is an outstanding problem for the SM.

Baryon and Lepton Numbers

The fact that the proton p does not decay into positron plus photon, $e^+ + \gamma$, or muon plus photon, $\mu^+ + \gamma$, implies a conservation law of baryon number B (the proton possessing $B = 1$ and the others $B = 0$). Furthermore, the stability of μ^- and τ^- against decay into $e^- + \gamma$ implies conservation of lepton numbers L_e, L_μ, and L_τ. These are regarded as global, not local, symmetries, so there are no associated gauge fields or interactions. Interestingly, however, these symmetries are not built into the SM, so are not guaranteed by it. More interestingly, these symmetries are actually destroyed in one attempt to go beyond the SM. This is the hypothesis that QCD may be unified with electroweak interactions to produce a "grand unified" theory (GUT). The simplest GUT is the one in which the $SU(2) \otimes U(1) \otimes SU(3)$ symmetry is assumed to be a subgroup of the much tighter symmetry $SU(5)$, and in that theory the proton is unstable:

$$p \rightarrow e^+ + \pi^0 \qquad [10]$$

The predicted lifetime is $10^{30\pm1}$ years, while a recent estimate of the lifetime for this decay mode is $> 5 \times 10^{32}$ years. It may be that GUTs do not exist in nature, but since the decay [10] violates conservation of the quantities B and L_e, even entertaining the idea that the decay might take place begs the question, "are these conservation laws sacrosanct?"

Another recent development which leads to the same question is the subject of neutrino oscillations. A strong motivation for this is the solar neutrino problem; this is the problem that the number of electron neutrinos detected on Earth, originating in the Sun, is less than the number predicted, by a factor close to 3. The mismatch could be at least partly, and perhaps completely, explained if electron neutrinos "oscillated" into muon and/or tau neutrinos on their passage from the Sun to the Earth, since the reaction which detects the neutrinos on Earth is sensitive only to electron neutrinos, and not to the other species. But oscillation is only permitted if L_e, L_μ, and L_τ are not separately conserved quantities. Oscillation can also only take place if the masses of the different neutrinos are different – the oscillation rate depends on Δm^2 – hence not all the neutrinos may be massless.

Discrete Symmetries

Ever since parity violation was discovered in weak interactions (nuclear beta decay) by Wu in 1957, the whole subject of discrete symmetries has presented problems which are still not resolved. The symmetries in question are

- P (space inversion): $(x, y, z) \rightarrow (-x, -y, -z)$
- T (time reversal): $t \rightarrow -t$
- C (particle–antiparticle conjugation): particle $\leftrightarrow$ antiparticle

Are the laws of physics invariant under these operations? The Wu experiment revealed that weak interactions are not invariant under P, but what about other interactions and other operations? In this context, the CPT theorem is highly important. According to this theorem (based on very general assumptions), all laws of nature must be invariant under the combined operation CPT, so that, for example, the fact that weak interactions are not invariant under P means that they are not invariant under the product CT either.

The violation of P invariance in beta decay was soon related to the fact that the neutrino involved (the electron neutrino – or, to be precise, antineutrino) was massless. Spin-1/2 particles like the electron and neutrino obey the Dirac equation, which may be written out as a pair of coupled equations for left- and right-handed states. In the case $m = 0$, however, these equations decouple so it is possible to have a massless spin-1/2 particle which is either left-handed or right-handed. Any interaction involving this particle would automatically violate parity (which turns a left-handed state into a right-handed one). Experiments have verified that the neutrino is indeed left-handed. The SM incorporates this in the sense that the left-handed electron e^-_L and the electron neutrino ν_e are assigned to a

weak isospin SU(2) doublet, while the right-handed electron e^-_R transforms as a singlet. A similar pattern is repeated for the μ and τ particles and their neutrinos. The phenomenon of neutrino oscillations, on the other hand, does not allow all the neutrino states also to be purely left-handed (since they cannot be massless). This poses a potential problem for the SM.

For a few years after 1957 it was believed that beta decay violated C as well as P, but conserved the product CP; and indeed that all weak interactions were CP invariant. In 1964, however, it was found that there is a small element of CP violation in K^0 decay. CP-violating effects are also expected in B^0 decays. The physical origin of CP violation is still not understood, but its importance is that it implies T violation, so that in (at least some) weak interactions, there is an "arrow of time" on the subnuclear scale. (Such an arrow of time is, of course, familiar in thermodynamics.) This is used in a cosmological context to explain baryon–antibaryon asymmetry in the Universe.

Baryon–Antibaryon Asymmetry

In the standard model of cosmology it is shown that applying the known laws of physics to the early Universe (the first few minutes) leads to the conclusion that at an age of 226 s nuclear fusion reactions took place resulting in a mixture of 74% protons and 26% α particles, so that, hundreds of thousands of years later, when galactic condensation took place, it would involve precisely this admixture of hydrogen and helium gases. Just this amount of helium has been found in the Sun, giving great confidence to the "big bang" model. Assuming that at extremely small times the baryon number of the Universe was zero, $B=0$, and assuming also (a big assumption, but one nevertheless made by cosmologists) that the Universe is made of matter and not antimatter, we may then ask, why is this – where has the antimatter gone?

Surprisingly, this question was addressed as early as 1966 by Sakharov, who showed that, starting with an initial state with $B=0$, it would be possible to reach a state with $B\neq 0$ as long as three conditions obtained: B violating interactions, CP and C violating interactions, and lack of thermal equilibrium. GUTs and ordinary weak interactions already provide possibilities for the first two of these conditions. Breakdown of thermal equilibrium will be expected to occur as the Universe expands. When the particle density is high, reactions such as $p+\bar{p}\rightarrow\gamma+\gamma$ will ensure an equal population of baryons and antibaryons, even in the presence of B violating interactions, but as the density increases and this reaction rate becomes less than the expansion rate, thermal equilibrium can no longer be maintained. Thus, GUTs offer an explanation of why there is no antimatter in the Universe. It might be thought that this sort of explanation is implausible, since the B-violating and CP-violating forces are so weak, but actually this is not a problem, since the ratio of baryon number to photon number in the Universe is of the order $N_B/N_\gamma\approx 10^{-9}$; so we may conjure up a scenario in which the B and CP violating forces give rise to a volume of space in which there are, say, 10^9 antibaryons, 10^9+1 baryons and approximately the same number of photons. Then, all the antibaryons become annihilated leaving one baryon and 10^9 photons – as observed.

A recent development in the area of discrete symmetries has been the suggestion by Kostelecky and coworkers that there might exist spontaneous violation of CPT and Lorentz symmetry.

Topological Charges

Conserved quantities of a quite different type have received a lot of attention in recent decades. Their conservation is a consequence of nontrivial boundary conditions for the fields. A famous example is the sine-Gordon "kink." The sine-Gordon equation

$$\frac{\partial^2\phi}{\partial t^2}-\frac{\partial^2\phi}{\partial x^2}+\frac{1}{b^2}\sin(b\phi)=0 \qquad [11]$$

describes a scalar field in one space and one time dimension. It is a nonlinear equation which possesses, among others, the interesting solution

$$f(\xi)=\frac{4}{b}\arctan\exp[\pm(\gamma/\sqrt{b})\xi]$$

where $\xi=x-vt$ and $\gamma=(1-v^2)^{-1/2}$. This corresponds to a solitary wave which moves, preserving its shape and size – in distinction to usual waves, which spread out and dissipate. Waves of this type are called solitons, and solitons have in fact been observed moving along canals. In this case, they are solutions to the Korteveg de Vries equation. Equation [11] clearly possesses the constant solutions

$$\phi=\frac{2\pi n}{b},\quad n=0,\pm1,\pm2,\ldots$$

which, it may be shown, all have zero energy. We may then construct a solution of the above type, but with $n=0$ as $x\rightarrow-\infty$ and $n=N$ as $x\rightarrow+\infty$. This so-called "kink" solution has finite energy and is not continuously deformable into a solution with $n=0$ everywhere, since this would involve overcoming an

infinite energy barrier. The "kink number" may be characterized as a charge: defining the current

$$J^\mu = \frac{b}{2\pi}\varepsilon^{\mu\nu}\partial_\nu\phi$$

with $\varepsilon^{\mu\nu}$ the totally antisymmetric symbol, it is clear that this is identically conserved, $\partial_\mu J^\mu = 0$. This is a consequence of the definition of $\varepsilon^{\mu\nu}$; it is not a consequence of invariance of the sine-Gordon Lagrangian under a symmetry operation, so the current J^μ is not a Noether current. The associated conserved charge is

$$Q = \int J^0\,\mathrm{d}x = \frac{b}{2\pi}\int \frac{\partial\phi}{\partial x}\,\mathrm{d}x$$
$$= \frac{b}{2\pi}[\phi(\infty) - \phi(-\infty)] = N$$

Models of the above type may be written down in a spacetime with more than two dimensions. In that case the above solution depends only on one coordinate, so represents an infinite planar "domain wall," on the two sides of which the field assumes different values. Such domain walls, as well as "cosmic strings," are considered as serious possibilities in cosmology.

Nonabelian gauge theories and the sigma model also provide a fertile ground for topological excitations – field configurations which for topological reasons do not decay. Gauge theories with spontaneous symmetry breaking have two-dimensional solutions corresponding to vortex lines and three-dimensional solutions corresponding to magnetic monopoles. In spacetime ($3+1$ dimensions), there is a solution to the gauge field equations, with no spontaneous symmetry breaking, corresponding to an "instanton," a finite-energy field configuration, localized in time as well as in space (hence the name). The gauge group here is SU(2), whose group space is S^3. Spacetime is "Euclideanized" into R^4, whose boundary is then S^3. Asymptotic field configurations may then be characterized by mappings of S^3 in field space into S^3 in parameter space, and since the third homotopy group of S^3 is nontrivial, $\pi_3(S^3) = Z$, these field configurations belong to different classes and are not deformable into each other. These define "degenerate vacua" of the gauge field equations. In quantum theory, tunneling between these vacua is allowed and 't Hooft has shown how this may give rise to deuteron decay $d \to e^+ + \bar{\nu}_\mu$. Other examples of topologically nontrivial configurations are so-called sphalerons, which may also contribute to baryon number violation in the early Universe, and skyrmions, constructs in the nonlinear sigma model which serve as a model for baryon number.

Supersymmetry

Supersymmetry is a fermion–boson symmetry, postulating that multiplets of fundamental particles contain both fermions and bosons. Thus, for example, since electrons exist there should also be "selectrons" – "scalar" electrons, with spin 0. There should also be photinos, with spin 1/2, to take their place alongside photons, and so on. If supersymmetry were exact, these particles would have the same mass as their partners and would have all been found, but in fact none have yet been discovered, so presumably supersymmetry is a broken symmetry. The feature that makes supersymmetry attractive is that it holds some promise for solving divergence problems in quantum field theory, since the radiative corrections from fermion and boson loops are opposite in sign and may exactly cancel. Supersymmetric models can also help to solve the so-called hierarchy problem in quantum field theory.

If supersymmetry is made into a local symmetry, rather than simply a global one, extra fields must be introduced (as the photon field was introduced above), and it turns out that one of these is a spin-2 field, which may be identified with the *graviton*. Local supersymmetry thus becomes supergravity.

General Relativity

Symmetries and conservation laws take on new aspects when general relativity is considered. Einstein's field equations relate the energy–momentum tensor of matter (and radiation) to the Ricci tensor of spacetime. The Ricci tensor has vanishing *covariant* divergence, which means that the energy–momentum tensor possesses the same property, but conservation of energy and momentum requires that it is the ordinary derivative, not the covariant one, of this tensor that should vanish. It might be expected that this problem could be alleviated by including the contribution of the gravitational field itself in energy–momentum tensor. This is quite reasonable, but then problems of interpretation arise, since at any one point in a general spacetime, a coordinate system might be found which is inertial (this is the force of the equivalence principle), corresponding to no gravitational field, and therefore no energy. The usual procedure is to introduce an energy–momentum "pseudotensor," and to conclude that energy in a gravitational field is not localizable.

The role of symmetries in general relativity is rather different from its role in particle physics, which is set in Minkowski spacetime. In a general spacetime there are no symmetries, but many examples of particular spacetimes with their own symmetries are now known. The symmetry operations involved are

isometries, with corresponding groups of motion (so that the isometry group of Minkowski space is the Poincaré group). These groups are an important subject of study in cosmology; for example, there is a classification of homogeneous cosmological models, labeled according to the Bianchi classification.

See also: Cotangent Bundle Reduction; Effective Field Theories; Electroweak Theory; General Relativity: Overview; Infinite-Dimensional Hamiltonian Systems; Noncommutative Geometry and the Standard Model; Quantum Field Theory: A Brief Introduction; Quasiperiodic Systems; Sine-Gordon Equation; Supergravity; Symmetries in Quantum Field Theory of Lower Spacetime dimensions; Symmetry and Symplectic Reduction; Symmetry Classes in Random Matrix Theory; Topological Defects and Their Homotopy Classification.

Further Reading

Aitchison IJ and Hey AJ (1981) *Gauge Theories in Particle Physics*. Bristol: Adam Hilger.

Cheng T-P and Li L-F (1984) *Gauge Theory of Elementary Particle Physics*. Oxford: Clarendon Press.

Eguchi T, Gilkey PB, and Hanson AJ (1980) Gravitation, gauge theories and differential geometry. *Physics Reports* 66: 213.

Huang K (1998) *Quantum Field Theory: From Operators to Path Integrals*. New York: Wiley.

Kostelecky VA (ed.) (2004) *CPT and Lorentz Symmetry: Proceedings of the Third Meeting*. Singapore: World Scientific.

Landau LD and Lifshitz EM (1971) *The Classical Theory of Fields*. Oxford: Pergamon Press.

Manton N and Sutcliffe P (2004) *Topological Solitons*. Cambridge: Cambridge University Press.

Perkins DH (2000) *Introduction to High Energy Physics*, 4th edn. Cambridge: Cambridge University Press.

Review of Particle Properties (2002), *Physical Review D* 66: 01002, July 2002, Part 1.

Rubakov V (2002) *Classical Theory of Gauge Fields*. Princeton: Princeton University Press.

Ryder LH (1996) *Quantum Field Theory*, 2nd edn. Cambridge: Cambridge University Press.

Stephani H (2004) *Relativity: An Introduction to Special and General Relativity*, 3rd edn. Cambridge: Cambridge University Press.

Weinberg S (1983) *The First Three Minutes*. London: Fontana.

Wess J and Bagger (1983) *Supersymmetry and Supergravity*. Princeton: Princeton University Press.

Wigner E (1939) On unitary representations of the inhomogeneous Lorentz group. *Annals of Mathematics* 40: 149.

Symmetries in Quantum Field Theory: Algebraic Aspects

J E Roberts, Università di Roma "Tor Vergata," Rome, Italy

© 2006 Elsevier Ltd. All rights reserved.

Introduction

This article treats the most important results and concepts relating to symmetry and conservation laws in quantum field theory. It includes such results as Wigner's theorem, Goldstone's theorem, the Bisognano–Wichmann theorem, the quantum Noether theorem, and the theorem on the existence of gauge groups and a field net. It is written within the framework of algebraic quantum field theory, this being the simplest setting capable of expressing all these concepts and results.

Symmetries come in many guises. They are to a physical system what automorphisms are to a mathematical theory. In fact, when a physical system is described in mathematical terms, its symmetries correspond to the automorphisms of the mathematical structure and in particular form a group, its symmetry group. The reader should bear in mind this simple picture throughout its diverse variations. Readers unfamiliar with the mathematical terminology should consult the appendix.

Elementary Quantum Mechanics

Before turning to quantum field theory, let us comment on symmetries in elementary quantum mechanics. These systems have the density matrices, that is, positive operators of trace 1, on an infinite-dimensional separable Hilbert space as states, the self-adjoint operators as observables. The expectation value of the bounded observable A in the state determined by ρ is given by $\mathrm{tr}\,\rho A$. Having specified the mathematical structure, the notion of symmetry follows. With a suggestive notation, it is a pair of mappings $A \mapsto \alpha A, \rho \mapsto \rho\alpha^{-1}$ such that

$$\mathrm{tr}\,\rho\alpha^{-1}\alpha A = \mathrm{tr}\,\rho A$$

for all observables A and states ρ.

If we take ρ and A to be the projections onto $C\phi$ and $C\psi$ for unit vectors ϕ and ψ, then the above condition corresponds to the conservation of

Notes:

transition probabilities $|(\phi, \psi)|^2$. This formed the starting point for Wigner's analysis, who concluded:

Theorem *Every symmetry is of the form* $A \mapsto UAU^{-1}$ *and* $\rho \mapsto U\rho U^{-1}$, *where* U *is a unitary or antiunitary operator.*

As could have been foreseen from the outset, this simple result in no way distinguishes one elementary quantum-mechanical system from another. A more useful notion of symmetry results if the Hamiltonian is reckoned as part of the information describing the system and, therefore, has to be left invariant by a symmetry. The operator U above must therefore satisfy the condition $UHU^{-1} = H$ and it commutes with the Hamiltonian. As the Hamiltonian is the generator of time translations, U is a constant of motion. This is the genesis of the relation between symmetries and conservation laws.

Quantum Field Theories

The simplest types of quantum field theories can be described by von Neumann algebras $\mathfrak{A}(\mathcal{O})$ depending on double cones $\mathcal{O}$ and subject to

$$\mathcal{O}_1 \subset \mathcal{O}_2 \Rightarrow \mathfrak{A}(\mathcal{O}_1) \subset \mathfrak{A}(\mathcal{O}_2)$$

a structure referred to as the net of observables.

An alternative approach would be to use the Wightman formalism. This would need a discussion of pointlike fields and the domains of definition of unbounded operators, thus complicating a general exposition of symmetry.

Comparing this description of a quantum field theory with that of an elementary quantum-mechanical system, the net clearly substitutes observables but nothing has yet been said about states. Since the set of double cones is directed under inclusion, the union of the $\mathfrak{A}(\mathcal{O})$ is a $*$-algebra $\mathfrak{A}$ and a state of our system is a state on this algebra.

Most states are of no physical relevance. A characterization of the states of physical relevance, even say to elementary particle physics, is not known although some progress has been made.

The net structure is the hallmark of a field theory and allows us to distinguish two important classes of symmetries. An internal symmetry α satisfies the condition

$$\alpha(\mathfrak{A}(\mathcal{O})) = \mathfrak{A}(\mathcal{O})$$

for all double cones $\mathcal{O}$. By contrast, a spacetime symmetry is an automorphism α_L implementing a Poincaré transformation L and hence satisfying the condition

$$\alpha_L(\mathfrak{A}(\mathcal{O})) = \mathfrak{A}(L\mathcal{O})$$

for every double cone $\mathcal{O}$. It is usually the case that internal symmetries commute with spacetime symmetries.

The state of prime relevance to elementary particle physics is the vacuum state ω_0. The corresponding Gelfand–Naimark–Segal (GNS) representation π_0 is called the vacuum representation. Now the vacuum state of a quantum field theory is typically unique and as such invariant under a symmetry of the system $\omega_0 \alpha^{-1} = \omega_0$.

Spacetime Symmetries

Since the vacuum state is invariant, we have a unitary representation of the Poincaré group implementing the spacetime symmetries in the vacuum representation. To illustrate the role of representations up to a factor, we take instead the GNS representation of a pure state corresponding to a particle of half-integral spin. Here we need a unitary representation of the covering group of the Poincaré group, inhomogeneous $\mathrm{SL}(2, \mathbb{C})$ to implement the symmetries. The situation for the subgroup of rotations is the same.

The most important property of these representations is positivity of the energy. More precisely, in a representation of relevance to elementary particle physics such as the vacuum representation, the generator P^0 of time translations is a positive operator $P^0 \geq 0$. Expressed in a frame-independent way, the spectrum of spacetime translations is contained in the closed forward light cone. It is one of the basic principles to be exploited in applying quantum field theory to elementary particle physics. Notice that the principle is no longer valid for an equilibrium state.

A similar situation arises in conformal field theory. Here the role of double cones in Minkowski space is played by intervals on the circle and that of the Poincaré group by the Möbius group on the circle $\mathrm{PSL}(2, \mathbb{R})$. Again, the Möbius group cannot always be unitarily implemented and conformal invariance is defined via a continuous unitary representation of its covering group. Most importantly, there is an analog of positivity of the energy. The generator of rotations of the circle is a positive operator.

A remarkable aspect of spacetime symmetries was discovered by Bisognano and Wichmann in an application of modular theory in the field-theoretical context looking not at double cones but at wedges. A wedge $\mathcal{W}$ is a Poincaré transform of the standard wedge $x^1 > |x^0|$. They found that the modular automorphisms of $\mathfrak{A}(\mathcal{W})$ and the vacuum vector Ω_0

have a geometric significance. For the standard wedge, they got the following result.

Theorem *If the net is derived from Wightman fields, the modular operator is* $e^{-2\pi K}$, *where* K *is the generator of boosts in the* 1*-direction and the modular conjugation is* $ZR\Theta$, *where* Θ *is the* TCP*-operator,* R *is the rotation through* π *about the* 1*-axis, and* Z *is the unitary operator equal to* 1 *on the Bose subspace and* $-i$ *on the Fermi subspace.*

The modular data for $\mathfrak{A}(\mathcal{O})$ and Ω_0 also admit a geometric interpretation for the free massless scalar field.

These facts enhance our understanding of spacetime symmetries. The ideas have meanwhile been applied to curved spacetime to select a state with vacuum-like properties using the principle of the geometric action of the modular conjugation.

Gauge Symmetry

Gauge symmetries do not fit into our scheme in that they act trivially on the observable algebra $\mathfrak{A}$. To exhibit a gauge symmetry we need a larger net $\mathfrak{F}$ called the field net. The gauge group will be the group of automorphisms of $\mathfrak{F}$ leaving the subnet $\mathfrak{A}$ pointwise fixed and $\mathfrak{A}$ the subnet of $\mathfrak{F}$ of fixed points under G. This has the merit of indicating the mathematical framework for gauge symmetry but otherwise begs important questions. *A priori* one does not know what properties $\mathfrak{F}$ should have nor how it should be constructed.

The right approach is to understand what intrinsic structure of $\mathfrak{A}$ governs the existence of a nontrivial gauge group. This brings us back to the states or representations relevant to elementary particle physics. A condition for selecting some of these relevant representations is that asymptotically they be like the vacuum in spacelike directions. More precisely, π must be unitarily equivalent to the vacuum representation π_0 on the spacelike complement of every double cone.

The resulting theory of superselection sectors hinges on the property of Haag duality that, for each double cone $\mathcal{O}$,

$$\mathfrak{A}(\mathcal{O}) = \mathfrak{A}(\mathcal{O}')'$$

where $\mathcal{O}'$ denotes the spacelike complement of $\mathcal{O}$. It implies that every representation satisfying the selection criterion is unitarily equivalent to one of the form $\pi_0\rho$, where ρ is an endomorphism of $\mathfrak{A}$ localized in some fixed but arbitrary double cone, that is, $\rho(A) = A$ if $A \in \mathfrak{A}(\mathcal{O}')$. The endomorphisms thus obtained are closed under composition and hence the objects of a full tensor subcategory $\mathcal{T}$ of the category of all endomorphisms and their intertwiners. There is a dimension function d defined on the objects of $\mathcal{T}$, $d(\rho) = 1, 2, \ldots, \infty$. If $\mathcal{T}_f$ denotes the full subcategory whose objects have finite dimension, then the following result holds.

Theorem $\mathcal{T}_f$ *is equivalent to the tensor category of finite-dimensional continuous unitary representations of a canonical compact group* G. *There is a canonical field net* $\mathfrak{F}$ *with Bose–Fermi commutation relations extending* $\mathfrak{A}$ *such that* G *is the group of automorphisms of* $\mathfrak{F}$ *leaving* $\mathfrak{A}$ *pointwise fixed.*

The first step in the proof is to define and analyze the statistics of the representations in question. The statistics of an irreducible representation ρ can be classified as being para-Bose or para-Fermi of order $d(\rho)$. The second step is to show that each ρ of finite dimension has a well-defined conjugate up to equivalence. The third and most difficult step is showing that $\mathcal{T}_f$ can be embedded in the tensor category of Hilbert spaces.

The Local Implementation of Symmetries

Gauge symmetry has its associated conservation laws in that the different sectors of the last section are labeled by conserved quantities such as baryon number, lepton number, or electric charge, generically called charges. The theory is built round the idea of creating charge and elements of the field net carry charges. But there should be a dual approach based on measuring charges. One would like to prove the existence of local conserved currents corresponding to these charges. This has not proved possible but there is a good substitute, described below, which can be regarded as a weak version of a quantum Noether theorem.

If $\mathcal{O}_1 \subset \mathcal{O}_2$ is a strict inclusion of double cones, then the theory is said to satisfy the split property if there is a type I factor $\mathcal{M}$ such that

$$\mathfrak{A}(\mathcal{O}_1) \subset \mathcal{M} \subset \mathfrak{A}(\mathcal{O}_2)$$

where a type I factor is a von Neumann algebra isomorphic to some $\mathcal{B}(\mathcal{H})$. In this case $\mathcal{M}$ can be chosen in a canonical fashion and there is an isomorphism ψ called the universal localizing map of $\mathcal{B}(\mathcal{H})$ onto $\mathcal{M}$, where $\mathcal{H}$ is the underlying Hilbert space. We have $\psi(A) = A$ for $A \in \mathfrak{A}(\mathcal{O}_1)$.

Theorem *If* U *is an implementing representation of the internal symmetry group* G, $\psi(U)$ *will be a representation of* G *in* $\mathcal{M}$ *that continues to implement the symmetry on* $\mathfrak{A}(\mathcal{O}_1)$. *If* G *is a Lie group*

then the infinitesimal generators in the representation are an analog of locally integrated current densities.

Spontaneously Broken Symmetry

The standard physical example of a spontaneously broken symmetry is magnetization. Despite the overall rotational symmetry, a magnet picks out a preferred direction as its direction of magnetization. The chosen state breaks the symmetry.

The phenomenon of spontaneously broken symmetry involves an interplay of symmetries and certain classes of states, vacuum states, ground states, or equilibrium states. If such an ω is induced by a vector cyclic and separating for a local algebra $\mathfrak{A}(\mathcal{O})$, then, as explained in the appendix, given $\mathcal{O}$, modular theory yields a canonical unitary representation V of the internal symmetry group G:

$$gA = V_g A V_g^*, \quad A \in \mathfrak{A}(\mathcal{O})$$

The results concern the breaking of a one-parameter group $\lambda \mapsto \alpha_\lambda$ of symmetries. More precisely, one asks whether $\omega\delta = 0$ or not, where δ is the infinitesimal generator of $\lambda \mapsto \alpha_\lambda$,

$$\delta(F) = \lim_{\lambda \to 0} \lambda^{-1}(\alpha_\lambda(F) - F)$$

where norm convergence is understood and holds on a dense domain. δ, the derivation, is an infinitesimal symmetry. Goldstone first showed that the spontaneous breaking of such symmetries requires the presence of massless bosons. The following result is taken from a more modern treatment. $\mathcal{O}_R$ here denotes the double cone whose base is the ball in $t = 0$ of radius R centered on the origin and $\mathcal{D}$ the domain of δ.

Theorem *Let δ be a derivation on a field net $\mathfrak{F}$ in $s > 1$ spatial dimensions such that for $F \in \mathfrak{F}(\mathcal{O}_R) \cap \mathcal{D}$*

$$|\omega_0 \delta F| \leq c_{R,\varepsilon}(\|F\Omega\| + \|F^*\Omega\|) + \varepsilon\|\delta F\|$$

(i) *If $\liminf_{R\to\infty} c_{R,\varepsilon} R^{-(s-1)/2} = 0$, then $\omega_0\delta = 0$.*

(ii) *If $\liminf_{R\to\infty} c_{R,\varepsilon} R^{-(s-1)/2} < \infty$, then $\omega_0\delta \neq 0$ is only possible if the spectrum of the translations coincides with the forward light cone V_+ and the boundary $\partial V_+/\{0\}$ has non-trivial spectral measure (i.e., there are massless particles in the theory).*

(iii) *If $c_{R,\varepsilon}$ is polynomially bounded in R, then $\omega_0\delta \neq 0$ is only possible if the spectrum of translations coincides with V_+ but there are not necessarily any massless particles.*

Symmetries of the S-matrix

Scattering theory not only allows one to construct the multiparticle scattering states but also shows that internal symmetries and spacetime symmetries continue to act on these states and are therefore symmetries of the S-matrix. We can, however, ask what are all the symmetries of the S-matrix. An answer was provided by Coleman and Mandula, who showed that, when there is nontrivial scattering, there are no further symmetries of the S-matrix.

Appendix

In an effort to make this article more self-contained, this appendix collects together a few simple pertinent concepts and results from the theory of operator algebras. A C^*-algebra is a $*$-algebra $\mathcal{A}$ with a norm $\|\cdot\|$ making it into a Banach algebra and satisfying

$$\|A^*A\| = \|A\|^2$$

for every $A \in \mathcal{A}$. Any C^*-algebra can be realized as a norm closed $*$-subalgebra of the C^*-algebra $\mathcal{B}(\mathcal{H})$ of all bounded operators on a Hilbert space $\mathcal{H}$. A von Neumann algebra $\mathcal{R}$ is a C^*-algebra that is the dual space of a Banach space. This Banach space $\mathcal{R}_*$, the predual of $\mathcal{R}$, is intrinsically defined. The topology on $\mathcal{R}$ determined by duality with $\mathcal{R}_*$ is called the σ-topology. $\mathcal{B}(\mathcal{H})$ is a von Neumann algebra and its predual is the set of trace class operators. Any von Neumann algebra can be realized as a σ-closed unital $*$-subalgebra of some $\mathcal{B}(\mathcal{H})$.

A state on a C^*-algebra $\mathcal{A}$ is a positive linear functional ω of norm 1. If $\mathcal{A}$ has a unit I the normalization condition can be expressed as $\omega(I) = 1$. Of fundamental importance is the relation between representations and states. A representation of $\mathcal{A}$ on a Hilbert space $\mathcal{H}$ is just a structure-preserving mapping or morphism of $\mathcal{A}$ into $\mathcal{B}(\mathcal{H})$. For simplicity, we suppose that $\mathcal{A}$ has a unit. Given a state ω, there is an associated representation π_ω defined by a vector Ω such that $\pi_\omega(\mathcal{A})\Omega$ is dense in the Hilbert space in question, that is, it is a cyclic vector for the representation and

$$\omega(A) = (\Omega, \pi_\omega(A)\Omega), \quad A \in \mathcal{A}$$

that is, the cyclic vector implements the given state. This is referred to as the GNS construction. Given any two such representations, there is a unique unitary operator mapping the one cyclic vector onto the other and realizing the equivalence of the representations.

A state of a von Neumann algebra is said to be normal if it is continuous in the σ-topology. If ω is normal, then $\pi_\omega(\mathcal{R})$ is σ-closed.

An inclusion of unital von Neumann algebras has the split property if there is an intermediate type I factor, that is, if it has the form $\mathcal{R}_1 \subset \mathcal{B}(\mathcal{H}) \subset \mathcal{R}_2$.

The following elementary observation is often used in treating symmetries. If α is an automorphism of $\mathcal{A}$ with $\omega\alpha^{-1} = \omega$, there is a unique unitary operator leaving the cyclic vector Ω invariant and inducing α in the representation π_ω. In other words, $U\Omega = \Omega$ and

$$U\pi_\omega(A)U^{-1} = \pi_\omega(\alpha A)$$

If we apply the above lemma to a group G of symmetries leaving a state invariant, it yields a group $U(g)$ of unitaries satisfying the condition

$$U(gh) = U(g)U(h), \quad g, h \in G$$

since $U(g)$ is uniquely defined by the above conditions.

When there is no invariant state, the situation is more complicated. Suppose there is a group G of symmetries and a representation π of $\mathfrak{A}$ where each g is unitarily implemented. Thus, there is a unitary $U(g)$ with

$$U(g)\pi(A)U(g)^{-1} = \pi(gA), \quad A \in \mathfrak{A}$$

All we can now conclude is that

$$U(gh) = Z(g,h)U(g)U(h)$$

where $Z(g, h)$ is a unitary in $\mathfrak{A}'$, the commutant of $\mathfrak{A}$, satisfying the 2-cocycle identity

$$Z(gh,k)Z(g,h) = Z(g,hk)^g Z(h,k)$$

where ${}^gX = U(g)XU(g)^{-1}$. U is said to be a representation up to a factor. It can be chosen to be a representation if the cocycle Z is a coboundary, that is, if there is a unitary $Y(g)$ in $\mathfrak{A}'$ such that

$$Y(g)^g Y(h) = Y(gh)Z(g,h)$$

In general, little is known about solving problems of this kind, but there are a number of results when π is irreducible and the unitary group of its commutant reduces to the circle.

We turn now to consider the modular theory of von Neumann algebras. A vector Ω is said to be separating for a von Neumann algebra $\mathcal{R}$ if $A\Omega = 0$ and $A \in \mathcal{R}$ implies $A = 0$. If Ω is both cyclic and separating, there is a uniquely determined closed antilinear involution S with $SA\Omega = A^*\Omega$ for $A \in \mathcal{R}$. If $S = J\Delta^{1/2}$ is the polar decomposition of S, then the unitary operators Δ^{it} induce automorphisms δ^{it} of $\mathcal{R}$ and $J\mathcal{R}J = \mathcal{R}'$. J is called the modular conjugation, Δ the modular operator, and δ^{it} the modular automorphisms. The closure of $\{\Delta^{1/4}A\Omega : A \in \mathcal{R}, A \geq 0\}$ is a cone, called the natural cone. Every normal state of $\mathcal{R}$ is implemented by a unique vector in the natural cone. If α is an automorphism of $\mathcal{R}$, there is therefore a unique vector Ω_α in the natural cone such that, for every $A \in \mathcal{R}$,

$$(\Omega, \alpha^{-1}(A)\Omega) = (\Omega_\alpha, A\Omega_\alpha)$$

There is now a canonical unitary operator V_α defined by

$$V_\alpha A\Omega = \alpha(A)\Omega_\alpha$$

V_α maps the natural cone into itself and $\alpha \mapsto V_\alpha$ is an implementing representation of the group of automorphisms of $\mathcal{R}$. Under these circumstances, we do not have to deal with representations up to a factor.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Boundary Conformal Field Theory; Current Algebra; Quantum Fields with Topological Defects; Supergravity; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Two-Dimensional Models.

Further Reading

Bisognano JJ and Wichmann EH (1976) On the duality condition for quantum fields. *Journal of Mathematical Physics* 17: 303–321.

Borchers HJ and Buchholz D (1985) The energy–momentum spectrum in local field theories with broken Lorentz symmetry. *Communications in Mathematical Physics* 97: 169–185.

Buchholz D (1982) The physical state space of quantum electrodynamics. *Communications in Mathematical Physics* 85: 49–71.

Buchholz D, Doplicher S, and Longo R (1986) On Noether's theorem in quantum field theory. *Annals of Physics* 170: 1–17.

Buchholz D, Doplicher S, Longo R, and Roberts JE (1992) A new look at Goldstone's theorem. *Reviews in Mathematical Physics* (Special Issue): 49–83.

Buchholz D, Dreyer O, Florig M, and Summers SJ (2000) Geometric modular action and spacetime symmetry groups. *Reviews in Mathematical Physics* 12: 475–560.

Coleman S and Mandula J (1967) All possible symmetries of the *S*-matrix. *Physical Review* 159: 1251–1256.

Doplicher S and Roberts JE (1990) Why there is a field algebra with a compact gauge group describing the superselection structure in particle physics. *Communications in Mathematical Physics* 131: 51–107.

Guido D and Longo R (1996) The conformal spin and statistics theorem. *Communications in Mathematical Physics* 181: 11–35.

Lopuszański J (1991) *An Introduction to Symmetry and Supersymmetry in Quantum Field Theory*. Singapore: World Scientific.

O'Raifeartaigh L (1965) Lorentz invariance and internal symmetry. *Physical Review* 139: B1052–B1062.

Symmetries in Quantum Field Theory of Lower Spacetime Dimensions

J Mund, Universidade de São Paulo, São Paulo, Brazil
K-H Rehren, Universität Göttingen, Göttingen, Germany

© 2006 Elsevier Ltd. All rights reserved.

Symmetries in Quantum Field Theory

Symmetries have proved to be one of the most powerful concepts in quantum theory, and in quantum field theory in particular. From the beginnings of quantum mechanics, it is well known that the presence of a symmetry allows one to predict relations between different measurements, to classify spectra (energy or other), and to understand the Pauli exclusion principle, to name only a few applications. Much more remarkably, in modern relativistic quantum field theory, designed to describe the interactions of elementary particles, fundamental interactions have been found to be induced by the principle of local gauge invariance.

One distinguishes spacetime symmetries (Poincaré or conformal transformations), which change the position and orientation of the system in space and time, and internal symmetries, which preserve the localization, acting on certain internal degrees of freedom. The Coleman–Mandula (1967) theorem states that internal and spacetime symmetries cannot be mixed, in the sense that the generators of internal symmetries must be Lorentz scalars, hence the total group of symmetries factorizes into a direct product. Supersymmetries are an exception of this theorem because their generators do not form a Lie algebra, and they were in fact designed to circumvent the Coleman–Mandula theorem.

It is well known that the structure of symmetries of quantum systems in low-dimensional spacetime differs significantly from that in four-dimensional spacetime. ("Low" means in our context two or three, depending on the type of charge localization, c.f. below.) To name some examples:

- Two-dimensional quantum systems may have much higher symmetries than four-dimensional ones:
  - In two dimensions, there exist massive integrable models with infinitely many conservation laws and factorizable scattering matrices (*see* Integrability and Quantum Field Theory). These models exhibit solitonic superselection sectors, c.f. below.
  - The conformal group of two-dimensional spacetime is infinite dimensional, allowing for

Notes:

the exact computation of correlation functions by the help of Ward identities (Belavin, Polyakov, and Zamolodchikov 1984). Only the finite-dimensional Möbius group, however, is also a symmetry of the vacuum state. Möbius covariance implies that the theory contains two subtheories of chiral fields defined on the light rays $t - x =$ constant, resp. $t + x =$ constant, and that these can be extended to fields defined on a circle, by adding a "point at infinity" to the light ray (Lüscher and Mack 1976). One arrives thus at one-dimensional chiral quantum field theories on a circle, which will play an important role in the discussion below.

- Continuous symmetries cannot be spontaneously broken in two dimensions. The latter is true not only for relativistic quantum field theory (Coleman 1973), but also in quantum statistical mechanics (Mermin and Wagner 1966) where it is responsible for the absence of ferromagnetism (*see* Symmetry Breaking in Field Theory). Spontaneous symmetry breakdown requires long-range order which is overcome by thermal fluctuations down to zero temperature, because these diverge logarithmically (in the thermodynamical limit) in two dimensions. This theorem thus illustrates how the spacetime dimension-dependent size of phase space has an effect on internal symmetries of quantum systems. A detailed mathematical analysis of the balance between phase space (thermal fluctuations) and long-range order (symmetry breakdown) has been given in a recent discussion of the Goldstone theorem (Buchholz, Doplicher, Longo and Roberts 1992).
- The Coleman–Mandula theorem, excluding a mixing between internal and spacetime symmetries (see above), is valid only in higher dimensions.

In more recent times, it has become apparent that low-dimensional quantum systems do not only admit more symmetries, but they may exhibit internal symmetries of an entirely new type, not describable by groups of transformations. In this article, we shall focus on the various ways in which the new symmetries can arise, and how they can be understood. In order to properly appreciate these issues, let us first recall some basic symmetry concepts in the conventional case.

In the traditional setting, symmetries arise in the form of groups of transformations of the quantum system which leave observable quantities (e.g., vacuum expectation values and correlation functions) invariant. The symmetries form a group of $*$-automorphisms of the algebra of fields:

$$\begin{aligned} \alpha_g(\phi_1\phi_2) &= \alpha_g(\phi_1)\alpha_g(\phi_2) \\ \left(\alpha_g(\phi)\right)^* &= \alpha_g(\phi^*) \\ \alpha_{g_1}\alpha_{g_2} &= \alpha_{g_1 g_2} \end{aligned} \qquad [1]$$

(typically given by linear transformations of field multiplets). In the strongest case, the automorphisms are implemented by unitary operators on the state space

$$U(g)\phi U(g)^* = \alpha_g(\phi) \qquad [2]$$

The implementers form a representation of the group of automorphisms,

$$U(g_1)U(g_2) = U(g_1 g_2) \qquad [3]$$

and there is an invariant vector state (a ground state, or the vacuum state in relativistic quantum field theory),

$$U(g)\Omega = \Omega \qquad [4]$$

However, depending on the dynamics of the quantum system, these relations cannot always be fully realized. One therefore considers several weaker or more general notions of symmetries relevant in four dimensions:

- *Spontaneously broken symmetries.* The transformations are given as automorphisms of an algebra, but which are not unitarily implemented in a given irreducible representation of the algebra. Invariant pure states do not exist.
- *Projective representations.* The symmetries are unitarily implemented, but the implementers fail to satisfy the group law [3]. They give rise to ray (projective) representations or representations of a covering group. In particular, an invariant state vector as in [4] cannot exist in an irreducible representation.
- *Infinitesimal symmetries.* Lie algebras of infinitesimal transformations, given as derivations of an algebra, which cannot be integrated to finite transformations. Derivations may or may not be implemented in a given representation of the algebra by commutators with self-adjoint generators.
- *Supersymmetry.* The infinitesimal transformations form a graded Lie algebra.
- Local gauge symmetries form an infinite-dimensional group which are, however, not realized as automorphisms of the quantum algebra. Quantization of classical gauge interactions usually proceeds by breaking the gauge invariance in some way and restoring it at a later stage.

The Connection between Symmetry and Superselection Sectors

It is often convenient to describe a model in terms of localized fields which do not represent an observable (in the sense of quantum mechanics that an operator corresponds to some measurement prescription). For example, Fermi fields which violate the principle of causality because they anticommute with each other at spacelike distance rather than commute are not observables. Only fields which are quadratic in the Fermi fields (densities of charge, current, energy) are observables. This means that an internal symmetry is used in order to distinguish the observables as those operators which are invariant under the symmetry: in the example, the symmetry transformation multiplies each Fermi field by -1 (by the spin-statistics theorem, this transformation coincides with the univalence of the Lorentz group). We characterize this situation by writing

$$A(O) = F(O)^G \qquad [5]$$

where $A(O)$ and $F(O)$ stand for the algebras of observables and fields localized in some spacetime region O, respectively, G is the internal symmetry group acting by automorphisms on each $F(O)$ without affecting the localization, and $F(O)^G \equiv \{a \in F(O), \alpha_g(a) = a \text{ for all } g \in G\}$ denotes the subalgebra of invariants. The internal symmetry group G which distinguishes the observables according to [5] is usually called the "(global) gauge group."

If the gauge symmetry G is unbroken in the vacuum state, then there is a well-known connection between symmetry and superselection rules (*see* Symmetries and Conservation Laws): namely, the observables act reducibly on the vacuum Hilbert space representation of F because they commute with the unitary operators which implement the symmetry (or with their infinitesimal generators, usually called charges). As a consequence, the validity of the superposition principle is restricted because two eigenstates of different eigenvalues of the charges cannot exhibit interference. In other words, they belong to different superselection sectors. Wick, Wightman, and Wigner (1952) were the first to point out this relation. We therefore call this scenario the "WWW scenario" for brevity.

In the WWW scenario, the decomposition of the Hilbert space is determined by the central decomposition of the internal symmetry group (the eigenvalues of the Casimir operators). In this way, the superselection sectors are in one-to-one correspondence with the irreducible representations of the internal symmetry group.

Superselection sectors of two-dimensional models do not follow this scheme expected by the WWW scenario (see below). This was most strikingly demonstrated through the classification of the unitary highest-weight representations of the Virasoro algebra (Friedan, Qiu, and Shenker) which is nothing other than the classification of the superselection sectors of the observable algebra generated by the chiral stress–energy tensor, and through the determination of their fusion rules by Belavin, Polyakov, and Zamolodchikov (1984).

In two dimensions, one is therefore lacking a compelling *a priori* ansatz, like the WWW scenario, for describing the system in terms of auxiliary nonobservable charged fields. At this point, one may argue that from an operational point of view, a quantum field theory, and in particular its symmetries, should be understood entirely in terms of its observables. (This viewpoint is emphasized in the algebraic approach to QFT, *see* Algebraic Approach to Quantum Field Theory.) We shall therefore now ask the opposite question: suppose we are given an algebra A of local observables (without knowledge of a field algebra and its gauge group). We define the superselection sectors intrinsically as (the unitary equivalence classes of) the positive-energy representations of A. Then the question is: do these sectors arise through a WWW scenario from some field algebra and a gauge symmetry, and if so, can the latter be reconstructed from the given observables alone?

The answer in four dimensions is positive, thanks to a deep result due to Doplicher and Roberts (1990). Let us sketch the line of reasoning leading to this result in some detail, because it shows how the connection between (global) gauge symmetry on the one hand and spacetime geometry on the other hand emerges through the principle of causality (locality) of relativistic quantum field theory, and because it makes apparent what is different in low-dimensional spacetime.

The analysis is based on the general structure theory of superselection sectors due to Doplicher, Haag, and Roberts (DHR, 1971). The latter starts with a selection criterion invoking the concept of a localized charge: a superselection sector which by measurements within the causal complement of some spacetime region O cannot be distinguished from the vacuum sector. The heuristic idea is, of course, that the sector is obtained from the vacuum sector by placing some charge in the region O (e.g., by the application of a localized charged field operator to the vacuum vector).

It has been shown (Buchholz and Fredenhagen 1982) that positive-energy representations of

massive theories always satisfy this selection criterion with a localization region O of the form of a narrow cone extending in spacelike direction. (In massless theories with long-range interactions, such as QED, the situation is more complicated because the charge creates an electric field whose flux at infinity does not vanish (Gauss' law) and is not Lorentz invariant.) DHR assume that the localization region is even compact, and can be chosen arbitrarily within the unitary equivalence class of the representation.

Exploiting a strong version of locality (Haag duality) for the vacuum representation of the observables, DHR proceed to define an associative composition (or fusion) law for positive-energy representations. This law is commutative only up to unitary equivalence. The crucial point is that the unitary intertwiner establishing this equivalence (the statistics operator) can be chosen in a unique way provided any pair of spacelike disconnected localization regions can be continuously deformed into any other such pair.

This point marks the separation between high and low dimensions. In two dimensions, in each pair of spacelike disconnected regions, one region is to the left of the other, thus distinguishing the pair (O_1, O_2) from (O_2, O_1). Consequently, they cannot be deformed into each other, and there arise two statistics operators. The same holds in three dimensions when the localization regions are spacelike cones, and O_1, O_2 are taken within (the causal complement of) some larger spacelike cone. If the spacetime dimension is at least 4, or if in three dimensions the localization regions are compact, then the statistics operator is unique and, as a consequence, coincides with its inverse.

The (non-)uniqueness of the statistics operator has far-reaching consequences concerning our original question about the underlying gauge symmetry. Namely, the DHR analysis proceeds to show that the set of positive-energy representations equipped with the composition law, and the linear spaces of inertwiners between different representations, together form the mathematical structure of a C^* tensor category. The statistics operators which are distinguished intertwiners give additional structure to this category: this structure is called a (permutation) symmetry if the statistics operators coincide with their inverse, and it is called a braiding otherwise. (It gives rise to a representation of the permutation group or the braid group, respectively.) In other words, the spacetime topology, through the intervention of the uniqueness of the statistics operator, causes the tensor category to be symmetric in high dimensions, and braided in low dimensions.

At a more elementary level, one may think of statistics operators as reflecting commutation relations between the searched-for charged fields. Making an ansatz for the commutation relations at spacelike separation, essentially the same topological argument as before implies, together with Poincaré invariance, that the coefficients appearing in this relation should form a representation of the permutation group, or of the braid group, respectively. The DHR approach, however, is entirely intrinsic, avoiding any *a priori* assumption of charged fields.

The duality theorem due to Doplicher and Roberts (1990) now states that every symmetric C^* tensor category (with some further qualifications valid in the DHR setting) is isomorphic to the category of unitary representations of a compact group, in which the composition law is the tensor product and the (permutation) symmetry is the natural one. Moreover, the category uniquely determines the group, and by a crossed product construction (an action of the category on the algebra A) one reconstructs a field algebra F such that [5] holds. If fermionic sectors are present, then there is some arbitrariness in the commutation relations among the corresponding fermionic fields, which can be exploited to produce the normal commutation relations (fermionic fields anticommute among each other, and bosonic fields commute with any field at spacelike separation). This fixes the field algebra F up to unitary equivalence. The conclusion is that the WWW scenario is the most general in four dimensions (apart from the reservations due to long-range forces, see above).

Generalized Symmetries in Low Dimensions

In view of the success of this program in four dimensions and the advantage of the WWW scenario for model building, the obvious challenge is to search for an analogous understanding of superselection sectors (charges) in low dimensions in terms of an algebra of charged fields and a gauge symmetry distinguishing the observables. This gauge symmetry cannot, in general, be a group for several reasons:

- As stated before, the tensor category of superselection sectors possesses only a braiding, rather than a (permutation) symmetry, hence the duality theorem fails.
- One can associate a (statistical) dimension d_π to each superselection sector $[\pi]$ which is multiplicative under the composition law (fusion), and additive under direct sums. In a symmetric

category, the dimensions are necessarily positive integers. Indeed, in the WWW scenario, they coincide with the naive dimension of the associated representation of the gauge group. But in the low-dimensional models, the dimensions turn out to be nonintegers in general.

- Moore and Seiberg (1988) have axiomatized the superselection structure of chiral and two-dimensional conformal field theories in terms of a system of recoupling and braiding coefficients controlling the fusion of sectors and its noncommutativity. (In fact, this system is basically equivalent to the DHR category.) For models such as SU(2) current algebras at level k, these coefficients turn out to coincide with the recoupling and braiding coefficients one can associate with a quantum group deformation (Drinfel'd 1986) of SU(2) with deformation parameter $q = -\exp i\pi/k$. Representations of quantum groups (quasitriangular Hopf algebras, *see* Hopf Algebras and q-Deformation Quantum Groups) have a tensor product defined in terms of a noncocommutative coproduct. Moreover, they possess a quantum dimension which is a q-deformation of an integer. The quantum dimensions precisely match the statistical dimensions of the superselection sectors. All this strongly suggests that quantum groups appear as generalized symmetries in two dimensions, at least in a large class of models.

A natural testing ground for the search for appropriate generalized symmetry concepts in low dimensions is the abundance of models in chiral and two-dimensional conformal QFT (*see* Two-Dimensional Models). As mentioned before, conformal symmetry in two dimensions has far-reaching consequences, especially the existence of chiral quantum fields which are defined on a one-dimensional light ray. As a null direction in the two-dimensional spacetime, this ray unites both the spacelike property of carrying a causal structure, and the timelike property that the generator of translations has positive spectrum (energy). These two features together with Möbius covariance are so powerful that they allow for the exact construction of large classes of models. The most elementary ones (minimal models) are completely described by the chiral stress–energy density field, that is, the local generator of the conformal symmetry. Other models also contain currents which are the local generators of internal symmetries. These models exhibit many nontrivial superselection structures, which illustrate the wide range of possible deviations from higher-dimensional QFT, and at the same time exhibit possible approaches to appropriate symmetry concepts in low dimensions.

Attempts to classify the possible algebraic structures of generalized internal symmetries in a model-independent setting start from the idea that the representation category of the internal symmetries of a given model should be equivalent to the tensor category of its superselection sectors. Several algebraic structures have been proposed as candidates, complying with this idea. They all assume specific modifications or deformations of eqns [1]–[5] above, highly constrained by self-consistency. Among these proposals are:

- quantum groups (see e.g., Fröhlich and Kerler 1993),
- weak quasiquantum groups (Mack and Schomerus 1992) and rational Hopf algebras (Fuchs *et al.* 1994),
- weak C^* Hopf algebras (Rehren 1997, Böhm and Szlachányi 1996) or quantum groupoids (Nikshych and Vainerman 1998), and
- braided groups (Majid 1991).

In several cases, the respective "symmetry algebra" can be reconstructed from the tensor category of superselection sectors, and a field algebra with linear transformation behavior can be constructed which contains the observables as invariant elements as in [5]. However, the situation is unsatisfactory for various reasons. First, the class of QFT models for which these constructions have been performed is quite restricted (most constructions work only for rational models, i.e., models with a finite set of charges); second, the reconstructed symmetry algebra is not unique and finally, the constructed field algebras have features which diverge significantly from the WWW scenario. For example, it is not always warranted that the quantum symmetries are consistent with the *-structure, indispensable for Hilbert space positivity (a necessary prerequisite for the probability interpretation of quantum theory). Moreover, typically there are global gauge transformations which are implemented by localized field operators, thus exhibiting a mixing of local and global concepts. It also happens that this holds for elements in the center of the symmetry algebra, which implies that the field algebra is not local relative to its gauge invariant elements, that is, the charged fields do not commute with the gauge-invariant elements at spacelike separation. In other constructions, the field algebra is not associative, or there are no finite field multiplets.

Historically, the first candidate for a "symmetry algebra" compatible with braid group statistics has

been the structure of a quantum group, as mentioned above. However, in physically interesting models, the quantum group is not semisimple and thus has too many (namely, indecomposable) representations. Solutions to this problem have been:

1. A BRS approach in an indefinite-metric framework (Hadjiivanov *et al.* 1991),
2. "Truncation," that is, discarding the "unphysical" representations. Fröhlich and Kerler (1993) have done this consistently in a categorical framework. In fact, they have given a complete classification of the possible braided tensor categories generated by a single irreducible object with statistical dimension d satisfying $1 < d < 2$, in terms of categories constructed from the "truncated" representations of $U_q(\mathrm{sl}_2)$. Truncation can also be performed by dividing the quantum group itself through the ideal which is annihilated by all "physical" representations, leading to a weak quasiquantum group (Mack and Schomerus 1992).
3. Relaxing the axioms, thus admitting the more general structures mentioned above.

All the above approaches assume a given generalized symmetry concept and show to what extent field algebras complying with it can be constructed. They thus concern nonobservable objects, and it is no contradiction if different symmetry concepts can be associated with the same observable data.

A more radical concept of global gauge symmetry, applicable to the low-dimensional case, has been developed by Longo and Rehren (1995). Its point of departure is the notion of a conditional expectation, which has the same abstract properties as a group average. In the WWW scenario, the Haar measure of the compact gauge group defines an average

$$\mu : F \ni \phi \mapsto \int \mathrm{d}\mu(g)\, \alpha_g(\phi) \in A \qquad [6]$$

which is a positive linear map respecting the localization, and the observables are invariant, $\mu(a) = a$. In fact, the observables are exactly the image of this map, that is, [5] is equivalently formulated, but without reference to the group transformations, as

$$A(O) = \mu(F(O)) \qquad [7]$$

Turning to the observables A of a quantum field theory in low dimensions, one looks for a quantum field theory F, containing A and equipped with a conditional expectation μ such that [7] holds, and which preserves the vaccum state. F may not satisfy local commutativity, but it should be local relative to the observables in the sense mentioned before. In rational chiral CFT, such extensions can be classified (and indeed constructed) in terms of the superselection category of A, giving direct access to the decomposition of the vacuum Hilbert space of F into superselection sectors of A. The advantage here is that no problems with Hilbert space structure can arise (because the approach is entirely in terms of operator algebras); a drawback is that in general F is not unique, and nonvacuum representations of F also have to be considered in order to generate all sectors of A.

The method can be used to classify and construct both nonlocal chiral extensions as candidates for sector-generating field algebras for a theory A of chiral observables, and local two-dimensional quantum field theories containing two given chiral subtheories, that is, observable algebras of two-dimensional models (Kawahigashi and Longo 2004). The chiral sector structure of the latter models is described by a "modular invariant." In many cases, this means that their thermal partition functions are invariant under the group $\mathrm{PSL}(2,\mathbb{Z})$ of modular transformations of the temperature (see below).

At this point, another link between spacetime and internal symmetries may be noted. The modular theory of von Neumann algebras (*see* Tomita–Takesaki Modular Theory) associates a one-parameter group of automorphisms (called the "modular group") with a state and an algebra "in standard position." In quantum field theory, for the vacuum state and an algebra of observables localized in certain wedge regions of Minkowski spacetime, this group can be identified with a boost subgroup of the Lorentz group (Bisognano and Wichmann 1975). Similarly, in chiral CFT on the circle, the modular group associated with the observables in an interval and the vacuum coincides with a subgroup of the Möbius group. For nonlocal theories, there may be an obstruction, however. On the other hand, if a subalgebra is stable under the modular group of some algebra, then there is a conditional expectation from the larger algebra onto the smaller algebra. Combining these general theorems, the Möbius covariance of the inclusions $A(O) \subset F(O)$ implies the existence of a conditional expectation, that is, the above generalization of the average over the internal symmetry. Moreover, assuming a generalized notion of compactness ("finite index") for the generalized internal symmetry, the Bisognano–Wichmann property holds also for nonlocal theories (Longo and Rehren 2004).

Of course, there is also a WWW scenario in chiral theories, that is, one may restrict a local theory to its invariants under some group of internal gauge

symmetries ("orbifold models"). It then happens that the invariants not only have the expected superselection sectors in correspondence with the representations of the gauge group, but in addition "twisted" sectors appear which, together with the former, constitute a "quantum double" structure. The twisted sectors arise by restriction of solitonic sectors of the original theory, which are in one-to-one correspondence with the elements of the gauge group (Müger 2005). Solitonic sectors are localizable with respect to two different vacua, and do not admit an unrestricted composition law.

Special Issues

A particularly simple situation is the case of anyons, that is, when all sectors have statistical dimension 1. Then the sectors form an abelian group $\hat{G}$ under fusion, and one can construct a WWW scenario with global gauge group G the dual of $\hat{G}$. The ensuing quantum fields satisfy generalized commutation relations at spacelike separation, given by an abelian representation of the braid group, where the coefficients can be arbitrary complex phases (responsible for the name "anyons"). However, it is known that there can arise an obstruction, which enforces the "local" global gauge transformations (mentioned before) to be present. In this case, the gauge symmetry can also be described by a quasiquantum group. It is noteworthy that free anyon fields have been constructed in two-dimensional spacetime, while in three dimensions there can be no (cone-) localized massive anyon fields which are free in the sense that they generate only single-particle states from the vacuum (Mund 1998).

The charge structure of massive quantum field theories in two dimensions is very different both from that encountered in conformal quantum field theories, and from the charge structure in high dimensions. It has been observed long ago that, in contrast to four dimensions, the strong locality property (Haag duality) which is necessary to set up the DHR analysis of superselection sectors, fails for the algebra of invariants under an internal gauge group in two dimensions. This algebraic feature can be traced back to the fact that the causal complement of a point is disconnected in two dimensions, or, in physical terms, that "a charge cannot be transported around a detector" without passing through its region of causal dependence. Müger (1998) has shown that any algebra of observables which satisfies Haag duality, cannot possess any nontrivial DHR superselection sectors at all, and that the only sectors which can exist are solitonic sectors. This general result nicely complies with the experience with integrable models, as mentioned before.

There are also some results giving interesting insight, which can be obtained intrinsically in terms of the observables. One of them concerns "central" observables (generalized Casimir operators).

Casimir operators in the WWW scenario are functions of the generators of the internal symmetry which usually are integrals over densities belonging to the field algebra F (Noether's theorem). Since they also commute with the generators, they can be approximated by local observables, and are therefore defined in each representation of the latter. By Schur's lemma, they are multiples of the identity in each irreducible sector. Since the eigenvalues of Casimir operators distinguish the representations of the gauge group, they also distinguish the sectors.

In chiral CFT extended to the circle (see above), one can find global "charge measuring operators" C_i, one for each sector π_i, in the center of the observable algebra (Fredenhagen *et al.* 1992) which have similar properties. They arise as a consequence of an algebraic obstruction to define the charged sectors on the circle, related to a nontrivial effect if a charge is "transported once around the circle," and form an operator representation of the fusion rules within the global algebra of observables. Under rather natural conditions clarified by Kawahigashi, Longo, and Müger (2001), the matrix of eigenvalues $\pi_j(C_i)$ is nondegenerate, that is, the generalized Casimir operators completely distinguish the superselection sectors. In this case, the superselection category is a modular category (*see* Braided and Modular Tensor Categories): the matrix with entries $d_{\pi_i}\pi_j(C_i)$ and the diagonal matrix with entries $\pi_j(U)$ (where U is the Möbius rotation by 2π) are multiples of the generators S and T of the "modular group" $PSL(2,\mathbb{Z})$, in a matrix representation labeled by the superselection sectors of the chiral observables. The physical significance of this matrix representation is that it relates thermal expectation values for different values of the temperature (Cardy 1986, Kac and Peterson 1984, Verlinde 1988)

These examples, together with the failure of the Coleman–Mandula theorem, may illustrate the intricate relations among spacetime geometry, covariance, and internal symmetry (charge structure) in low dimensions. In relativistic quantum field theory, the link is provided by the principle of locality, which "turns geometry into algebra."

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Braided and Modular Tensor Categories; Hopf Algebras and *q*-Deformation

Quantum Groups; Integrability and Quantum Field Theory; Quantum Field Theory: A Brief Introduction; Quantum Fields with Topological Defects; Symmetries and Conservation Laws; Symmetries in Quantum Field Theory: Algebraic Aspects; Symmetry Breaking in Field Theory; Tomita–Takesaki Modular Theory; Two-Dimensional Conformal Field Theory and Vertex Operator Algebras; Two-Dimensional Models.

Further Reading

Böhm G and Szlachányi K (1996) A coassociative C^*-quantum group with non-integral dimensions. *Letters in Mathematical Physics* 35: 437–448.

Coleman S and Mandula J (1967) All possible symmetries of the *S*-matrix. *Physical Review* 159: 1251–1256.

Doplicher S and Roberts JE (1990) Why there is a field algebra with a compact gauge group describing the superselection structure in particle physics. *Communications in Mathematical Physics* 131: 51–107.

Fredenhagen K, Rehren K-H, and Schroer B (1992) Superselection sectors with braid group statistics and exchange algebras II: geometric aspects and conformal covariance. *Reviews in Mathematical Physics* SI1: 113–157.

Fröhlich J and Kerler T (1993) *Quantum Groups, Quantum Categories, and Quantum Field Theory*. Lecture Notes in Mathematics, vol. 1542. Berlin: Springer.

Fuchs J, Ganchev A, and Vecsernyés P (1994) On the quantum symmetry of rational field theories. *Theoretical Mathematical Physics* 98: 266–276.

Hadjiivanov LK, Paunov RR, and Todorov IT (1991) Quantum group extended chiral *p*-models. *Nuclear Physics B* 356: 387–438.

Longo R and Rehren K-H (1995) Nets of subfactors. *Rev. Mathematical Physics* 7: 567–597.

Mack G and Schomerus V (1992) Quasi Hopf quantum symmetry in quantum theory. *Nuclear Physics B* 370: 185–230.

Majid S (1991) Braided groups and algebraic quantum field theories. *Letters in Mathematical Physics* 22: 167–175.

Moore G and Seiberg N (1989) Classical and quantum conformal field theory. *Communications in Mathematical Physics* 123: 177–254.

Müger M (1998) Superselection structure of massive quantum field theories in $1+1$ dimensions. *Reviews in Mathematical Physics* 10: 1147–1170.

Mund J (1998) No-go theorem for "free" relativistic anyons in $d=2+1$. *Letters in Mathematical Physics* 43: 319–328.

Rehren K-H (1997) Weak C^* Hopf symmetry. In: Doebner H-D and Dobrev VK (eds.) *Group Theoretical Methods in Physics*, pp. 62–69. (q-alg/9611007). Sofia: Heron Press.

Schomerus V (1995) Construction of field algebras with quantum symmetries from local obsèrvables. *Communications in Mathematical Physics* 169: 193–236.

Wick GC, Wightman AS, and Wigner EP (1952) The intrinsic parity of elementary particles. *Physical Review* 88: 101–105.

t **Hooft–Polyakov Monopoles** *see* Solitons and Other Extended Field Configurations

Thermal Quantum Field Theory

C D Jäkel, Ludwig-Maximilians-Universität München, München, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Quantum field theory was initially invented in order to describe high-energy elementary particles, thereby unifying quantum mechanics and special relativity. In other words, quantum field theory was addressed to the so-called vacuum sector, that is, roughly speaking physics at zero temperature and zero particle density.

The same applies to the various mathematically rigorous versions of quantum field theory that have been developed since the mid-1950s. Indeed, in Wightman's axiomatic setting, quantum field theory is describes in terms of a set of the so-called vacuum expectation values. The "algebraic approach" to quantum field theory developed by Araki, Haag, Kastler, and their collaborators is more flexible in nature. In fact, right from the beginning, the new algebraic tools were successfully applied to lattice models and other nonrelativistic systems with infinitely many degrees of freedom (see *Operator algebras and quantum statistical mechanics* by O Bratteli and D W Robinson). But the need to treat large systems of relativistic particles was apparently not felt. Even in Haag's recent monograph, *Local Quantum Physics*, the subjects of algebraic quantum field theory and algebraic quantum statistical mechanics are treated separately.

It is remarkable that constructive field theory was ahead of its time in this respect. The famous $P(\phi)_2$ model (first constructed by Glimm and Jaffe) was adapted to thermal states by Høegh-Krohn as early as 1974 (see Høegh-Krohn (1974)). His paper was properly named "Relativistic quantum statistical mechanics in two-dimensional space-time," but only recently has it received proper attention.

At the same time, around 1974, cosmology and heavy-ion collisions drew the interest of phyiscists towards the quantum statistical mechanics of hot relativistic quantum systems. Well-known papers from this early stage include those by Weinberg, Bernard, and Dolan and Jackiw. While most of the papers used Euclidean path integrals, Umezawa and his school developed a real-time framework called "thermo-field dynamics," which involved a doubling of the degrees of freedom. The excellent review by Landsman and van Weert (1987) covers these early attempts; it also explains the basic connection to the algebraic approach.

In the following years, it became evident that statistical mechanics (in its standard formulation) is barely sufficient to derive the properties of bulk matter from the underlying microscopic description provided by quantum field theory. Thus, various people began to establish mathematically rigorous foundations for the description of thermal field theory. The most successful approach was launched by D Buchholz (with various collaborators), who, from about 1985 onwards, started applying the KMS condition (which describes a thermal equilibrium state in the operator-algebraic framework of local quantum physics) to relativistic quantum field theory. In 1994, Buchholz and Bros managed to integrate the holomorphic structure of Wightman field theory into Haag's operator-algebraic framework, which led them to the notion of a relativistic KMS condition.

The advanced mathematical concepts involved in the formulation of entropy densities for thermal quantum fields (see Narnhofer (1994)) do not allow us to present this topic. The reader is referred to the excellent book *Quantum Entropy and Its Use* by M Ohya and D Petz for an introduction to the subject. A discussion of the so-called thermalization

Notes:

effects that occur as a result of a curved spacetime is provided in Quantum Field Theory in Curved Spacetime. Another subject, which is missing almost completely, is perturbation theory. This subject has been covered extensively in three well-known textbooks by Kapusta, Le Bellac, and Umezawa.

Observables and States

Following Heisenberg, we start from the basic assumption that quantum theory can be formulated in terms of observables which form an algebra $\mathcal{A}$, that is, a vector space with a (noncommutative) multiplication law. Although our emphasis on the abstract algebraic structure may look strange, there is a profound reason for starting out with an abstract algebra of observables: as soon as one considers systems with infinitely many degrees of freedom, one encounters a possibility to realize the abstract elements of the algebra $\mathcal{A}$ as operators on a Hilbert space in various inequivalent ways. The famous equivalence between the Heisenberg and the Schrödinger picture simply breaks down. States which are macroscopically different (e.g., thermal equilibrium states for different temperatures) give rise – in a natural way, which will be discussed in the sequel – to unitarily inequivalent representations of the abstract algebra of observables $\mathcal{A}$, while states which only differ microscopically can be accommodated by density matrices within the same Hilbert space. In other words, a physical state is described macroscopically by specifying a representation, and microscopically by a density matrix in this representation.

In a Lagrangian approach, the algebra of observables $\mathcal{A}$ may be thought of as being generated by the underlying fields, currents, etc. This leads to the so-called polynomial algebras. It is mathematically convenient to assume that $\mathcal{A}$ is an algebra of bounded operators, generated by the bounded functions of the underlying quantum fields. If $\phi(x)$ is any such field and if $f \in \mathcal{S}(\mathbb{R}^{d+1})$ is any real test function with support in a bounded region of spacetime, then the corresponding operator

$$W(f) = \exp\left(\mathrm{i} \int \mathrm{d}x\, f(x)\phi(x)\right)$$

would be a typical element of $\mathcal{A}$. The set of operators $\{W(f) \mid \operatorname{supp} f \subset \mathcal{O}\}$ will generate a subalgebra $\mathcal{A}(\mathcal{O})$ of $\mathcal{A}$. The underlying fields can be recovered by taking (functional) derivatives, once a representation of $\mathcal{A}$ on a Hilbert space is specified.

The spacetime symmetry of Minkowski space manifests itself in the existence of a representation

$$\alpha : (\Lambda, x) \mapsto \alpha_{\Lambda,x} \in \operatorname{Aut}(\mathcal{A}), \quad (\Lambda, x) \in \mathcal{P}_+^\uparrow$$

of the (orthochronous) Poincaré group $\mathcal{P}_+^\uparrow$. Here $\alpha_{\Lambda,x}$ is an automorphism of $\mathcal{A}$, that is, a mapping from $\mathcal{A}$ to $\mathcal{A}$ which preserves the algebraic structure. Once a Lorentz frame is fixed by choosing a timelike vector $e \in V_+$, the time evolution $t \mapsto \alpha_{1,te}$ will be denoted by $t \mapsto \tau_t$.

For the free field, the group of automorphisms $(\Lambda, x) \mapsto \alpha_{\Lambda,x}$ is defined by

$$\alpha_{\Lambda,x}(W(f)) := W\big(f(\Lambda^{-1}(. - x))\big)$$

As before, $f \in \mathcal{S}(\mathbb{R}^{d+1})$ is a Schwarz function over the Minkowski space $\mathbb{R}^{d+1}$.

While the invariance of the equations of motion is reflected in the existence of a representation of the Poincaré group in terms of automorphisms in the Heisenberg picture, at least the invariance with respect to Lorentz boosts is spontaneously broken in the Schrödinger picture for a thermal equilibrium state.

The usual notions of vector states and density matrices associated with a given Hilbert space (usually Fock space) are *a priori* not general enough to cover all cases of interest in thermal field theory. The following algebraic definition of a state substantially generalizes the notion of a state: A state ω is a positive, linear, and normalized functional, that is, a linear map $\omega : \mathcal{A} \to \mathbb{C}$ such that

$$\omega(a^* a) \geq 0 \quad \text{and} \quad \omega(1) = 1$$

Once a state ω is distinguished on physical grounds, the GNS reconstruction theorem provides a Hilbert space $\mathcal{H}_\omega$ and a representation π_ω of $\mathcal{A}$, that is, a map from $\mathcal{A}$ to the set of bounded operators $\mathcal{B}(\mathcal{H}_\omega)$, which preserves the algebraic relations.

It is instructive to consider the GNS representation of the Pauli matrices $\{\sigma_0 = 1, \sigma_1, \sigma_2, \sigma_3\}$. Given a state (a diagonal 2×2 matrix ρ with positive entries and $\operatorname{tr} \rho = 1$), the left regular representation (a construction well known from group theory)

$$\pi(\sigma_i)|\sqrt{\rho}> = |\sigma_i \sqrt{\rho}>, \quad i = 0, 1, 2, 3$$

defines a reducible representation on $\mathbb{C}^4$, unless one of the entries in the diagonal of ρ is zero (which corresponds to a pure state). In the latter case, the GNS Hilbert space is $\mathbb{C}^2$. By construction, $<\sqrt{\rho}\,|\pi(\sigma_i)|\sqrt{\rho}> = \operatorname{tr} \rho\sigma_i$, $i = 1, 2, 3$.

Thermal Equilibrium

The variety of nonequilibrium states ranges from mild perturbations of equilibrium states through steady states, whose properties are governed by external heat baths, or hydrodynamic flows up to totally chaotic states which no longer

admit a description in terms of thermodynamic notions. Buchholz *et al.* (2002) have initiated an investigation of nonequilibrium states that are locally (but not globally) close to thermal equilibrium. Unfortunately, we will not be able to cover this topic. Instead, we will concentrate on states which deviate from a true equilibrium state only microscopically.

Characterization of Thermal Equilibrium States

When the time evolution $t \mapsto \tau_t \in \mathrm{Aut}(\mathcal{A})$ is changed by a local perturbation, which is slowly switched on and slowly switched off again, then an equilibrium state ω returns to its original form at the end of this procedure. This heuristic condition of adiabatic invariance can be expressed by the stability requirement

$$\lim_{t\to\infty} \int_{-t}^{t} \mathrm{d}t\, \omega([a, \tau_t(b)]) = 0 \quad \forall a, b \in \mathcal{A} \qquad [1]$$

In a pioniering work Haag, Kastler, and Trych-Pohlmeyer showed that the characterization [1] of an equilibrium state leads to a sharp mathematical criterion, first encountered by Haag, Hugenholtz, and Winnink and more implicitly by Kubo, Martin, and Schwinger:

Definition 1 A state ω_β over $\mathcal{A}$ is called a KMS state for some $\beta > 0$, if for all $a, b \in \mathcal{A}$, there exists a function $F_{a,b}$ which is continuous in the strip $0 \leq \Im z \leq \beta$ and analytic and bounded in the open strip $0 < \Im z < \beta$, with boundary values given by

$$F_{a,b}(t) = \omega_\beta(a\tau_t(b)) \quad \text{and}$$
$$F_{a,b}(t + \mathrm{i}\beta) = \omega_\beta(\tau_t(b)a) \quad \forall t \in \mathbb{R} \qquad [2]$$

Before we start analyzing the properties of KMS states, we should mention an alternative characterization of thermal equilibrium states: passivity. The amount of work a cycle can perform when applied to a moving thermodynamic equilibrium state is bounded by the amount of work an ideal windmill or turbine could perform; this property is called semipassivity (Kuckert 2002): a state ω is called semipassive (passive) if there is an "efficiency bound" $E \geq 0$ ($E = 0$) such that

$$-(W\Omega_\omega, H_\omega W\Omega_\omega) \leq E \cdot (W\Omega_\omega, |P_\omega| W\Omega_\omega)$$
$$\forall W \in \pi_\omega(\mathcal{A})''$$

with $W^{-1} = W^*$, $[H_\omega, W] \in \pi_\omega(\mathcal{A})''$, and $[P_\omega, W] \in \pi_\omega(\mathcal{A})''$. Here (H_ω, P_ω) denote the generators implementing the spacetime translations in the GNS representation $(\mathcal{H}_\omega, \Omega_\omega, \pi_\omega)$. Generalizing the notion of complete passivity, the state ω is called completely semipassive if all its finite tensorial powers are semipassive with respect to one fixed efficiency bound E. It has been shown by Kuckert (2002) that a state is completely semipassive in all inertial frames if and only if it is completely passive in some inertial frame. The latter implies that ω is a KMS state or a ground state (a result due to Pusz and Woronowicz).

Let us now turn to properties of thermal equilibrium states which are specific for relativistic models. It was first recognized by Bros and Buchholz (1994) that KMS states of a relativistic theory have stronger analyticity properties in configuration space than those imposed by the traditional KMS condition:

Definition 2 A KMS state ω_β satisfies the relativistic KMS condition (Bros and Buchholz 1994) if there exists a unit vector e in the forward light cone V_+ such that for every pair of local elements a, b of $\mathcal{A}$ the function $F_{a,b}$

$$F_{a,b}(x_1, x_2) = \omega_\beta(\alpha_{x_1}(a)\alpha_{x_2}(b))$$

extends to an analytic function in the tube domain $\mathcal{T}_{\beta e/2} \times \mathcal{T}_{\beta e/2}$, where $\mathcal{T}_{\beta e/2} = \{z \in \mathbb{C} \mid \Im z \in V_+ \cap (\beta e/2 - V_+)\}$.

The relativistic KMS condition can be understood as a remnant of the relativistic spectrum condition in the vacuum sector. It has been rigorously established (Bros and Bruchholz 1994) for the KMS states constructed by Buchholz and Junglas (1989) and by C Gérard and the author for the $P(\phi)_2$ model. In the thermal Wightman framework (Bros and Buchholz 1996) it has been shown that the relativistic KMS condition implies existence of model-independent analyticity properties of thermal n-point functions. These properties also appear in perturbative computations of the thermal Wightman functions (Steinmann 1995).

We now turn to the properties of the set of KMS states. For given β, the convex set $\mathcal{S}_\beta$ of all KMS states is known to form a simplex; the extreme points in the set $\mathcal{S}_\beta$ are called extremal KMS states. As a consequence, the extremal states in $\mathcal{S}_\beta$ can be distinguished with the help of "classical" (central) observables, that is, by observables which commute with all other observables.

If ω is an extremal KMS state and γ is an automorphism which commutes with the time evolution $t \mapsto \tau_t$, then the state ω' defined by

$$\omega'(a) := \omega(\gamma(a)), \quad a \in \mathcal{A}$$

is again an extremal KMS state to the same parameter values. If $\omega' \neq \omega$, one says that the symmetry is spontaneously broken.

Lorentz invariance with respect to boosts is always broken by a KMS state, since the KMS condition distinguishes a rest frame. A KMS state might also break spatial translation or rotation invariance. However, by averaging over the different configurations one can usually construct a translation- and rotation-invariant state. The situation is drastically different with respect to supersymmetry. Buchholz and Ojima (1997) have shown that supersymmetry is broken in any thermal state and it is impossible to proceed from it by "symmetrization" to states on which an action of supercharges can be defined.

Existence of Thermal Equilibrium States

Buchholz and Junglas (1989) demonstrated that the existence of KMS states can be guaranteed for a large class of quantum field-theoretic models. The basic assumption to be met concerns the phase-space properties of the model. A generalized trace norm (the so-called "nuclear norm") is used to estimate the "number" of degrees of freedom in phase space.

The first step is to construction a subspace $\mathcal{H}(\Lambda)$ of the vacuum Hilbert space $\mathcal{H}_{\text{vac.}}$, which represents excitations of the vacuum strictly localized inside of a bounded spacetime region $\hat{\mathcal{O}}$. Due to the strong correlations present in the vacuum state of any relativistic model, as a consequence of the Reeh–Schlieder property (see the section "Analyticity of n-point functions") this is a delicate procedure, which involves the so-called "split property." This property ensures that there exists a product vector η in vacuum Hilbert space $\mathcal{H}_{\text{vac.}}$ such that

$$(\eta; \pi_{\text{vac.}}(ab)\eta) = \omega_{\text{vac.}}(a) \cdot \omega_{\text{vac.}}(b)$$
$$\forall a \in \mathcal{A}(\mathcal{O}),\ b \in \mathcal{A}(\hat{\mathcal{O}})^{\text{c}} \qquad [3]$$

Here $\mathcal{O} \subset \hat{\mathcal{O}}$ denotes a slightly smaller open spacetime region (such that the closure $\overline{\mathcal{O}}$ is inside the interior of $\hat{\mathcal{O}}$) and $\mathcal{A}(\hat{\mathcal{O}})^{\text{c}} := \{A \in \mathcal{A} \mid [A, B] = 0\ \forall B \in \mathcal{A}(\hat{\mathcal{O}})\}$. The existence of a product vector can be ensured if the nuclear norm satisfies some mild bounds which are expected to hold in all models of physical interest. Given a product vector η which satisfies [3], the sought after subspace is

$$\mathcal{H}(\Lambda) := \overline{\pi_{\text{vac.}}(\mathcal{A}(\mathcal{O}))''\Omega_{\text{vac.}}}$$

The crucial step in the proof of existence of KMS states is to show that

$$\text{tr}\, E(\Lambda)\text{e}^{-\beta H}E(\Lambda) < \infty \quad \text{for } \beta > 0$$

if the nuclearity condition holds. Here $E(\Lambda)$ denotes the projection onto the subspace $\mathcal{H}(\Lambda)$ representing localized excitations and H denotes the Hamiltonian in the vacuum representation $\pi_{\text{vac.}}$. Next it is shown that the function

$$t \mapsto \omega_{\beta,\Lambda}(a\tau_t(b)) = \frac{1}{Z}\text{tr}\ E(\Lambda)\text{e}^{-\beta H}E(\Lambda)\pi_{\text{vac.}}(a\tau_t(b))$$

allows an analytic extension to a strip of width β which satisfies the KMS boundary condition [2] for $|t| < \delta$ if $a, b \in \mathcal{A}(\mathcal{O}_\circ)$ and $\mathcal{O}_\circ + te \subset \mathcal{O}$ for $|t| < \delta$. In the final step, Buchholz and Junglas were able to demonstrate that bounds on the nuclear norm are even sufficient to control the thermodynamic limit.

Given a thermal field theory, a slight variation of the method used by Buchholz and Junglas allows one to construct a KMS state for a new temperature (Jäkel 2004), that is, to change the temperature of a thermal state.

Thermal Representations

Given a KMS state ω_β, the GNS construction gives rise to a Hilbert space $\mathcal{H}_\beta$ and a representation π_β, called a thermal representation, of $\mathcal{A}$. The algebra $\mathcal{R}_\beta := \pi_\beta(\mathcal{A})''$ possesses a cyclic (due to the GNS construction) and separating (due to the KMS condition) vector Ω_β such that

$$\omega_\beta(a) = (\Omega_\beta, \pi_\beta(a)\Omega_\beta) \quad \forall a \in \mathcal{A}$$

The KMS condition implies that ω_β is invariant under time translations, that is, $\omega_\beta \circ \tau_t = \omega_\beta$ for all $t \in \mathbb{R}$. Thus,

$$U(t)\pi_\beta(a)\Omega_\beta = \pi_\beta(\tau_t(a))\Omega_\beta, \quad a \in \mathcal{A}$$

defines a strongly continuous unitary group $\{U(t)\}_{t\in\mathbb{R}}$ implementing the time evolution in the representation π_β. By Stone's theorem there exists a self-adjoint generator L such that

$$U(t) = \text{e}^{iLt}, \quad t \in \dot{\mathbb{R}} \qquad [4]$$

For $0 \leq \beta < \infty$, the Liouville operator L is not bounded from below; its spectrum is symmetric and consists typically of the whole real line. However, the negative part of L is "suppressed" with respect to the algebra of observables $\mathcal{R}_\beta := \pi_\beta(\mathcal{A})''$ in the following sense (Haag 1992): let $\mathbb{1}_{]-\infty,-\kappa]}$ be the spectral projection of L for the interval $]-\infty, -\kappa] \subset \text{Sp}(L)$, then

$$\|\mathbb{1}_{]-\infty,-\kappa]}A\Omega_\beta\| \leq \text{e}^{-\beta\kappa}\|A\| \quad \forall A \in \mathcal{R}_\beta$$

We now turn to structural aspects which are characteristic for a relativistic model, namely the existence of strong spatial correlations and the connection between the decay of these correlations and the spectral properties of the Liouville operator.

Let ω_β be a state, which satisfies the relativistic KMS condition. It follows (using a theorem of Glaser) that for $a \in \mathcal{A}$ the function $\Phi_a : \mathbb{R}^4 \to \mathcal{H}_\beta$,

$$x \mapsto \pi_\beta(\alpha_x(a))\Omega_\beta$$

can be analytically continued from the real axis into the domain $\mathcal{T}_{\beta e/2}$ such that it is weakly continuous for $\Im z \searrow 0$. If the usual additivity assumption $\cup_i \mathcal{O}_i = \mathcal{O} \Rightarrow \vee_i \mathcal{R}_\beta(\mathcal{O}_i) = \mathcal{R}_\beta(\mathcal{O})$ for the local von Neumann algebras holds, then

$$\mathcal{H}_\beta = \overline{\pi_\beta(\mathcal{A}(\mathcal{O}))\Omega_\beta} \qquad [5]$$

for any open spacetime region $\mathcal{O} \subset \mathbb{R}^{d+1}$. Junglas has shown that the thermal Reeh–Schlieder property [5] follows as well from the standard KMS condition, if ω_β is locally normal with respect to the vacuum representation.

The decay of spatial correlations depends on infrared properties of the model, and the essential ingredients for the following cluster theorem are the continuity properties of the spectrum of L near zero.

Theorem 3 *Let Ω_β denote the unique (up to a phase) normalized eigenvector with eigenvalue {0} of the Liouvillean L and let P^+ denote the projection onto the strictly positive part of the spectrum of L. Assume that there exist positive constants $m > 0$ and $C_1(\mathcal{O}) > 0$ such that*

$$\|e^{-\lambda L} P^+ \pi_\beta(a)\Omega_\beta\| \leq C_1(\mathcal{O}) \cdot \lambda^{-m}\|a\| \quad \forall a \in \mathcal{A}(\mathcal{O})$$

Here $\mathcal{O} \subset \mathbb{R}^{d+1}$ is an open and bounded spacetime region. Now consider two spacelike separated spacetime regions $\mathcal{O}_1, \mathcal{O}_2$, which can be embedded into $\mathcal{O}$ by translation and such that $\mathcal{O}_1 + \delta e \subset \mathcal{O}_2'$, $\delta >> \beta$. then, for $a \in \mathcal{A}(\mathcal{O}_1)$ and $b \in \mathcal{A}(\mathcal{O}_2)$,

$$|\omega_\beta(ba) - \omega_\beta(b)\omega_\beta(a)| \leq C_2 \cdot \delta^{-2m}\|a\|\,\|b\|$$

The constant $C_2(\beta, \mathcal{O}) \in \mathbb{R}^+$ may depend on the temperature β^{-1} and the size of the region $\mathcal{O}$ but is independent of δ, a, and b.

From explicit calculations one expects that $m = 1/2$ for free massless bosons in $3+1$ spacetime dimensions. Consequently, the exponent given on the right-hand side is optimal since it is well known that in this case the correlations decay only like δ^{-1}.

A description of thermal representations would be inadequate without pointing out one of the deepest connections between pure mathematics and physics that emerged in the last century: consider a von Neumann algebra $\mathcal{R}$ which possesses a cyclic and separating vector Ω. Then polar decomposition of the closeable operator $S : A\Omega \mapsto A^*\Omega$, $A \in \mathcal{R}$, provides an antiunitary operator J (the modular conjugation) and a self-adjoint operator $\Delta^{1/2}$. The connection to physics was established independently by Takesaki and Winnink, showing that the pair $(\mathcal{R}, \sigma)$ satisfies the KMS condition for $\beta = -1$, if one sets $\sigma_t(A) = \Delta^{it} A \Delta^{-it}$ for $A \in \mathcal{R}$.

Taking advantage of the Reeh–Schlieder property [5], one can associate modular objects to certain spacetime regions $\mathcal{O}$. In general, a physical interpretation of these modular objects is missing. But for two-dimensional thermal models, which factorize in light-cone coordinates, the modular group corresponding to the algebra of a spacelike wedge admits a simple description: at large distances (compared to β) from the boundary, the flow pattern is essentially the same as time translations. These are results due to Borchers and Yngvason (1999).

Analyticity Properties of *n*-Point Functions

The correlation functions describe the full physical content of the theory: all observable quantities can in principle be derived from them. This is so because according to the Wightman reconstruction theorem (which is closely related to the GNS construction) knowledge of the correlation functions allows the reconstruction of the full representation of the field algebra. The Wightman distributions $\{\mathcal{W}_\beta^{(n)}\}_{n \in \mathbb{N}}$,

$$\mathcal{W}_\beta^{(n)}(t_2 - t_1, x_2 - x_1, \ldots, t_n - t_{n-1}, x_n - x_{n-1}) := (\Omega_\beta, \phi_\beta(t_1, x_1) \cdots \phi_\beta(t_n, x_n)\Omega_\beta) \qquad [6]$$

where $\pi_\beta(W(f)) =: \exp(i \int dt\, dx\, f(t,x)\phi_\beta(t,x))$, satisfy a number of key properties: locality, positivity, Poincaré covariance, and temperedness. These properties have been formulated for thermal field by Bros and Buchholz (1996), and this section is entirely based on their work.

The relativistic KMS condition implies that the Wightman distributions $\{\mathcal{W}_\beta^{(n)}\}_{n \in \mathbb{N}}$ of a translation-invariant equilibrium state admit in the corresponding set of spacetime variables $(t_2 - t_1, x_2 - x_1), \ldots, (t_n - t_{n-1}, x_n - x_{n-1})$ an analytic continuation into the union of domains

$$(\alpha_1 \mathcal{T}_{\beta e}) \times \cdots \times (\alpha_{n-1} \mathcal{T}_{\beta e})$$

for $\alpha_i > 0$, $i = 1, \ldots, n-1$ and $\sum_{i=1}^{n-1} \alpha_i = 1$. The tube domains $\mathcal{T}_{\beta e}$ were specified in Definition 2. For $\beta \to \infty$, the tube $\mathcal{T}_{\beta e}$ tends to the vacuum tube $\mathcal{T}_{\text{vac.}} = \mathbb{R}^{d+1} + iV_+$; thus, one recovers the spectrum condition for the vacuum expectation values.

Let us now turn to the Fourier transformed Wightman correlation functions. Translation invariance implies

$$\tilde{\mathcal{W}}_\beta^{(n)}(\nu_1, p_1, \ldots, \nu_n, p_n)\delta(\nu_1 + \cdots + \nu_n)\delta(p_1 + \cdots + p_n)$$

The Wightman distribution $\tilde{\mathcal{W}}_\beta^{(n)}$ satisfies on the linear manifold $(\nu_1,p_1)+\cdots+(\nu_n,p_n)=0$ the KMS relation in the energy variables: for any pair of multi-indices (I,J) the identity

$$\tilde{\mathcal{W}}_\beta^{(n)}(J,I)=\mathrm{e}^{-\beta\nu_I}\tilde{\mathcal{W}}_\beta^{(n)}(I,J)$$

holds, where $\tilde{\mathcal{W}}_\beta^{(n)}(J,I)$ is an abbreviation for $\tilde{\mathcal{W}}_\beta^{(n)}(\{p_i\}_{i\in I},\{p_j\}_{j\in J})$ and $\nu_I=\sum_{i\in I}\nu_i$.

We now specialize to the two-point function $\mathcal{W}_\beta^{(2)}$. The corresponding commutation function $\mathcal{C}(x)$ is given by

$$\mathcal{C}(x_1-x_2)=\mathcal{W}_\beta^{(n)}(x_1,x_2)-\mathcal{W}_\beta^{(2)}(x_2,x_1)$$

Locality implies that $\operatorname{supp}\mathcal{C}\subset\overline{V_+}\cup\overline{V_-}$. The retarded and the advanced propagator r and a, formally given by

$$r(x)=\mathrm{i}\theta(x_\circ)\mathcal{C}(x),\quad a(x)=-\mathrm{i}\theta(-x_\circ)\mathcal{C}(x)$$

satisfy the relation

$$r-a=-\mathrm{i}\mathcal{C}$$

which corresponds to a partition of the support of $\mathcal{C}$ in its convex components: $\operatorname{supp}r\subset\overline{V_+}$ and $\operatorname{supp}a\subset\overline{V_-}$. For the free scalar field of mass m the commutator function is

$$\mathcal{C}^{(m)}(x)=\frac{1}{(2\pi)^2}\int_{\mathbb{R}^4}\mathrm{d}p\,\mathrm{e}^{-\mathrm{i}px}\tilde{\mathcal{C}}^{(m)}(p)$$

with

$$\tilde{\mathcal{C}}^{(m)}(p)=\frac{1}{2\pi}\operatorname{sgn}(\nu)\delta(\nu^2-p^2-m^2)$$

and subsequently the retarded and advanced propagators $r^{(m)}$ and $a^{(m)}$ are structural functions of the field algebra, which are determined by the c-number commutation relations of the fields. Thus, they are independent of the temperature, in contrast to the two-point function:

$$\tilde{\mathcal{W}}_\beta^{(2)}(p)=\frac{\tilde{\mathcal{C}}^{(m)}(p)}{1-\mathrm{e}^{-\beta\nu}}\qquad[7]$$

Let now $\tilde{\tau}(p)$ be the Fourier transform of the time-ordered function $\tau(x)$. The relation

$$\tilde{\tau}(p)=\frac{-\mathrm{i}\tilde{r}(p)+-\mathrm{i}\tilde{a}(p)\mathrm{e}^{-\beta\nu}}{1-\mathrm{e}^{-\beta\nu}}$$

shows that $\tilde{\tau}(p)$ and $-\mathrm{i}\tilde{r}(p)$ only "coincide up to an exponential tail" at very high energies (Bros and Buchholz 1996).

Particle Aspects

The condition of locality (together with the relativistic KMS condition) leads to strong constraints on the general form of the thermal two-point functions that allow one to apply the techniques of the Jost–Lehmann–Dyson representation. As has been shown by Bros and Buchholz (1996), the interacting two-point function $\mathcal{W}_\beta$ can be represented in the form

$$\mathcal{W}_\beta(t,x)=\int_0^\infty\mathrm{d}m\,\mathcal{D}_\beta(x,m)\mathcal{W}_\beta^{(2)}(t,x,m)$$

Here $\mathcal{D}_\beta(x,m)$ is a distribution in x,m which is symmetric in x, and

$$\mathcal{W}_\beta^{(2)}(t,x,m)=(2\pi)^{-1}\int\mathrm{d}\nu\mathrm{d}p\;\mathrm{e}^{\mathrm{i}(\nu t-px)}\tilde{\mathcal{W}}_\beta^{(2)}(\nu,p)$$

is the two-point correlation function of the free thermal field of mass m. In contrast to the vacuum case, the damping factors $\mathcal{D}_\beta(x,m)$ depend in a nontrivial way on the spatial variables x. The damping factors describe the dissipative effects of the thermal system on the propagation of sharply localized excitations. Bros and Buchholz suggested that the damping factor $\mathcal{D}_\beta(x,m)$ can be decomposed into a discrete and an absolute continuous part

$$\mathcal{D}_\beta(x,m)=\delta(m-m_0)\mathcal{D}_{\beta,\mathrm{d}}(x)+\mathcal{D}_{\beta,\mathrm{c}}(x,m)$$

and that the δ-contribution in the damping factors is due to stable constituent particles of mass m_0 out of which the thermal states are formed, whereas the collective quasiparticle-like excitations only contribute to the continuous part of the damping factors (Bros and Buchholz 1996).

In the case of spontaneously broken internal symmetries Bros and Buchholz (1998) have shown that the damping factors $\mathcal{D}_\beta^\pm(x,m)$ which appear in the representation of current-field correlations functions

$$\begin{aligned}&(\Omega_\beta,j_0(t,x)\phi_\beta(0,0)\Omega_\beta)\\&\quad=\int_0^\infty\mathrm{d}m\Big(\mathcal{D}_\beta^+(x,m)\partial_t\mathcal{W}_\beta^{(2)}(t,x,m)\\&\qquad+\mathcal{D}_\beta^-(x,m)\mathcal{W}_\beta^{(2)}(t,x,m)\Big)\end{aligned}$$

indeed contain a discrete (in the sense of measures) zero-mass contribution and are slowly decreasing in $|x|$ for small values of m. Thus, these damping factors coincide locally with the Källén–Lehmann weights appearing in the case of spontaneous symmetry breaking in the vacuum sector (Bros and Buchholz 1998). It is easily seen in examples that there is no sharp energy–momentum dispersion law for the Goldstone particles. Thus, the Källén–Lehmann representation is better suited than Fourier transformation to uncover the particle aspects of thermal equilibrium states.

Models of Thermal Field Theory

In the simplest case, the classical Lagrangian density of the so-called $P(\phi)_2$ models is given by

$$\mathcal{L} = (\partial_\nu \phi)(\partial^\nu \phi) - m^2\phi^2 - \frac{\lambda}{4}\phi^4 \qquad [8]$$

Here $\phi(t,x)$ denotes a real scalar field over spacetime. The construction of the corresponding quantized thermal field presented in this section (Gérard and Jäkel 2005) is based on the original ideas of Høegh-Krohn (1974).

Free Fields

Let $\mathfrak{h}_m$ denote the L^2-closure of $C_0^\infty(\mathbb{R})$ with respect to the norm $\|f\| = (f, (1/2\epsilon)f)$, where $\epsilon(k) = \sqrt{k^2 + m^2}$ denotes the one-particle energy for a single neutral scalar boson and the scalar product is the usual L^2-scalar product. The subspaces associated to a double cone $\mathcal{O}$ are given by

$$\mathfrak{h}_m(\mathcal{O}) := \{h \in \mathfrak{h}_m | \operatorname{supp} \Re h, \operatorname{supp} \nu^{-1}\Im h \subset O\}$$

where O denotes the basis of the double cone $\mathcal{O}$. The corresponding free quantum field is described by the Weyl algebra $\mathcal{W}(\mathfrak{h}_m) := \{W(f) \,|\, f \in \mathfrak{h}_m\}$, together with the time evolution $\{\tau_t^\circ\}_{t\in\mathbb{R}}$,

$$\tau_t^\circ(W(f)) = W(e^{it\epsilon}f), \quad f \in \mathfrak{h}_m$$

If $m > 0$, the KMS condition allows just one unique (quasifree) (τ°, β)-KMS state:

$$\omega_\beta^\circ(W(f)) := e^{-(1/4)(f,(1+2\rho)f)_m}, \quad \rho := (e^{\beta\epsilon} - 1)^{-1}$$

The GNS representation associated to the pair $(\mathcal{W}(\mathfrak{h}_m), \omega_\beta^\circ)$ is the well-known Araki–Woods representation, given by

$$\mathcal{H}_{\mathrm{AW}} := \Gamma(\mathfrak{h}_m \oplus \overline{\mathfrak{h}_m}), \quad \Omega_{\mathrm{AW}} := \Omega_{\mathcal{F}},$$
$$\pi_{\mathrm{AW}}(W(h)) := W_{\mathcal{F}}\left((1+\rho)^{1/2}h \oplus \overline{\rho}^{1/2}\overline{h}\right), \quad h \in \mathfrak{h}_m$$

Here $\overline{\mathfrak{h}_m}$ is the Hilbert space conjugate to $\mathfrak{h}_m$, $W_{\mathcal{F}}(.)$ denotes the usual Weyl operator on the Fock space $\Gamma(\mathfrak{h}_m \oplus \overline{\mathfrak{h}_m})$ and $\Omega_{\mathcal{F}} \in \Gamma(\mathfrak{h}_m \oplus \overline{\mathfrak{h}_m})$ is the Fock vacuum. The Liouvillean L_{AW} (see [4]) can be identified with $d\Gamma(\epsilon \oplus -\epsilon)$.

The local von Neumann algebra generated by $\{\pi_{\mathrm{AW}}(W(h)) \,|\, h \in \mathfrak{h}_m(\mathcal{O})\}$ is denoted by $\mathcal{R}_{\mathrm{AW}}(\mathcal{O})$. The algebra of observables for the free quantum field (and, as we will see, the $P(\phi)_2$ model) is the norm closure

$$\mathcal{A} := \overline{\bigcup_{\mathcal{O}\subset\mathbb{R}^2} \mathcal{R}_{\mathrm{AW}}(\mathcal{O})}^{C^*}$$

of the local von Neumann algebras.

The Thermal $P(\phi)_2$ Model

In $1+1$ spacetime dimensions Wick ordering is sufficient to eliminate the UV divergences of polynomial interactions. As it turns out, the leading order in the UV divergences is independent of the temperature (in agreement with the results found in Kopper *et al.* (2001)). Thus, it is a matter of convenience whether one uses the thermal covariance function C_β,

$$C_\beta(h_1, h_2) := \left(h_1, \frac{(1 + e^{-\beta\epsilon})}{2\epsilon(1 - e^{-\beta\epsilon})}h_2\right)_{L^2(\mathbb{R})}$$
$$h_1, h_2 \in \mathcal{S}(\mathbb{R})$$

or the vacuum covariance function $C_{\mathrm{vac.}}$ to define the Wick ordering:

$$:\phi(f)^n:_C = \sum_{m=0}^{[n/2]} \frac{n!}{m!(n-2m!)}\phi(f)^{n-2m}\left(-\frac{1}{2}C(f,f)\right)^m$$

Now let $P(\lambda)$ be a real-valued polynomial, which is bounded from below. Then Euclidean techniques can be used to define the operator sum

$$H_l := L_{\mathrm{AW}} + \int_{-l}^{l} :P(\phi_\beta(x)):_{C_\beta} dx$$

in the Araki–Woods representation and to show that H_l is essentially self-adjoint Gérard and Jäkel (2005). Thus, (the closure of) H_l can be used to define a perturbed time evolution $t \mapsto \tau_t^l$ on $\mathcal{A}$ and the vector

$$\Omega_l := \frac{e^{-(\beta/2)H_l}\Omega_{\mathrm{AW}}}{\|e^{-(\beta/2)H_l}\Omega_{\mathrm{AW}}\|}$$

induces a KMS state ω_l for the dynamical system $(\pi_{\mathrm{AW}}(\mathcal{A})'', \tau^l)$.

A finite propagation speed argument (using Trotter's product formula) shows that

$$\tau_t^l(A) := e^{itH_l}Ae^{-itH_l}, \quad t \in \mathbb{R} \qquad [9]$$

is independent of l for $A \in \mathcal{R}_{\mathrm{AW}}(\mathcal{O}), t \in \mathbb{R}$ fixed and l sufficiently large. Thus, there exists a limiting dynamics τ such that

$$\lim_{l\to\infty} \|\tau_t^l(A) - \tau_t(A)\| = 0 \qquad [10]$$

for all $A \in \mathcal{R}_{\mathrm{AW}}(\mathcal{O}), \mathcal{O}$ bounded. This norm convergence extends to the norm closure $\mathcal{A}$ of the local von Neumann algebras.

The existence of weak* limit points (which are states) of the (generalized) sequence $\{\omega_l\}_{l>0}$ is a consequence of the Banach–Alaoglu theorem. The fact that all limit states satisfy the KMS condition with respect to the pair $(\mathcal{A}, \tau)$ follows from [10]. To

prove that the sequence $\{\omega_l\}_{l>0}$ has only one accumulation point,

$$\omega_\beta = \lim_{l\to\infty} \omega_l \qquad [11]$$

is more delicate. Following Høegh-Krohn, Nelson symmetry is used in Gérard and Jäkel (2005) to relate the interacting thermal theory on the real line to the $P(\phi)_2$ model on the circle S^1 of length at temperature 0. The existence of the limit [11] then follows from the uniqueness of the vacuum state on the circle. The relativistic KMS condition can be derived by Nelson symmetry as well, using the fact that the discrete spectrum of the model on the circle satisfies the spectrum condition. Since the limit [11] exists on the norm closure $\mathcal{A}$ of the weakly closed local algebras, it follows from a result of Takesaki and Winnink that ω_β is locally normal with respect to the Araki–Woods representation (which itself is locally normal with respect to the Fock representation). Consequently,

$$\mathcal{R}_\beta(\mathcal{O}) := \pi_\beta(\mathcal{A}(\mathcal{O}))'' \cong \mathcal{R}_{\mathrm{AW}}(\mathcal{O}), \quad \mathcal{O} \text{ bounded}$$

that is, $\mathcal{R}_\beta(\mathcal{O})$ is (isomorphic to) the unique hyperfinite factor of type III_1. Moreover, the local Fock property implies that the split property holds.

Perturbation Theory

Steinmann (1995) has shown that perturbative expansions for the Wightman distributions of the $:\phi^4:_4$ model can be derived directly in the thermodynamic limit, using as only inputs the equations of motion and the (thermal) Wightman axioms. The result can be represented as a sum over generalized Feynman graphs.

The method consists in solving the differential equations for the correlation functions which follow from the field equation, by a power series expansion in the coupling constant, using the axiomatic properties of the Wightman functions as subsidiary conditions. The Wightman axioms are expected to hold separately in each order of perturbation theory, with the exception of the cluster property.

As expected, the UV renormalization can be chosen to be temperature independent, that is, one can use the same counterterms as in the vacuum case. But the infrared divergencies are more severe, they cannot be removed by minor adjustments of the renormalization procedure. Various elaborate resummation techniques have been proposed to (at least partially) remove the infrared singularities.

Another approach has been pursued by Kopper *et al.* (2001). They have investigated the perturbation expansion of the $:\phi^4:_4$ model in the imaginary-time formalism, using Wilson's flow equations. The result is once again that all correlation functions become ultraviolet-finite in all orders of the perturbation expansion, once the theory has been renormalized at zero temperature by usual renormalization prescriptions.

Asymptotic Dynamics of Thermal Fields

Timelike asymptotic properties of thermal correlation functions cannot be interpreted in terms of free fields due to persistent dissipative effects of a thermal system. This well-known fact manifests itself in a softened pole structure of the Green's functions in momentum space and is at the root of the failure of the conventional approach to thermal perturbation theory (Bros and Bruchholz 2002). In fact, assuming a sharp dispersion law, one would be forced to conclude that the scattering matrix is trivial (a famous no-go theorem by Narnhofer *et al.* (1983)).

However, there seems to be a possibility to find an effective theory, which is much simpler and still reproduces the correct asymptotic behavior of the full theory. Disregarding low-energy excitations, Bros and Buchholz (2002) have shown that the δ-contributions in the damping factors give rise to asymptotically leading terms which have a rather simple form: they are products of the thermal correlation function of a free field and a damping factor describing the dissipative effects of the model-dependent thermal background. This result is based on the assumption that the truncated n-point functions satisfy

$$\lim_{\mathrm{T}\to\infty} \mathrm{T}^{3(n-1)/2-\kappa} \mathcal{W}_\beta^{(n)}(t_1, x_1, \ldots, t_n, x_n)^{\mathrm{trunc.}} = 0$$
$$\kappa > 0$$

while the δ-contribution in the damping factors exhibit, for large timelike separations T, a $T^{-3/2}$ type behavior (in $3+1$ spacetime dimensions).

Bros and Buchholz (2002) have shown that the asymptotically dominating parts of the correlation functions can be interpreted in terms of quasifree states acting on the algebra generated by a Hermitian field ϕ_0 satisfying the commutation relations

$$[\phi_0(t_1, x_1), \phi_0(t_2, x_2)] = \Delta_{m_0}(t_1 - t_2, x_1, x_2) Z(x_1 - x_2)$$

Here Δ_{m_0} is the usual commutator function of a free scalar field of mass m_0 and Z is an operator-valued distribution commuting with ϕ_0 such that $\hat{\omega}_\beta(Z(x_1 - x_2)) = \mathcal{D}_{\beta,\mathrm{d}}(x_1 - x_2)$. (Here $\hat{\omega}_\beta$ denotes a KMS state for the algebra generated by ϕ_0.) Intuitively speaking, the field ϕ_0 carries an additional stochastic degree of freedom, which manifests itself in a central element that appears in the commutation relations and couples to the thermal background.

As ϕ_0 describes the interacting field asymptotically, one may expect that ϕ_0 satisfies the field

equation of the interacting field in an asymptotic sense. Buchholz and Bros (2002) have demonstrated that this assumption allows one to derive an explicit expression for the discrete part of the damping factors $\mathcal{D}_{\beta,\mathrm{d}}(x)$ in simple models.

See also: Axiomatic Quantum Field Theory; Quantum Field Theory in Curved Spacetime; Scattering in Relativistic Quantum Field Theory: The Analytic Program; Tomita–Takesaki Modular Theory.

Further Reading

Borchers H-J and Yngvason J (1999) Modular groups of quantum fields in thermal states. *Journal of Mathematical Physics* 40: 601–624.

Birke L and Fröhlich J (2002) KMS, etc., *Reviews in Mathematical Physics* 14: 829–871.

Bros J and Buchholz D (1994) Towards a relativistic KMS condition. *Nuclear Physics B* 429: 291–318.

Bros J and Buchholz D (1996) Axiomatic analyticity properties and representations of particles in thermal quantum field theory. *Annales de l'Institut Henri Poincaré* 64: 495–521.

Bros J and Buchholz D (1998) The unmasking of thermal Goldstone bosons. *Physical Review D* 58: 125012-1.

Bros J and Buchholz D (2002) Asymptotic dynamics of thermal quantum fields. *Nuclear Physics B* 627: 289–310.

Buchholz D and Junglas P (1989) On the existence of equilibrium states in local quantum field theory. *Communications in Mathematical Physics* 121: 255–270.

Buchholz D and Ojima I (1997) Spontaneous collapse of supersymmetry. *Nuclear Physics B* 498: 228–242.

Buchholz D, Ojima I, and Roos H (2002) Thermodynamic properties of non-equilibrium states in quantum field theory. *Annals of Physics* 297: 219–242.

Cuniberti G, De Micheli E, and Viano GA (2001) Reconstructing the thermal green functions at real times from those at imaginary times. *Communications in Mathematical Physics* 216: 59–83.

Derezinski J, Jaksic V, and Pillet CA (2003) Perturbations of W^*-dynamics, Liouvilleans and KMS states. *Reviews in Mathematical Physics* 15: 447–489.

Fröhlich J (1975) The reconstruction of quantum fields from Euclidean Green's functions at arbitrary temperatures. *Helvetica Physica Acta* 48: 355–363.

Gérard C and Jäkel C (2005) Thermal quantum fields with spatially cutoff interactions in $1+1$ space-time dimensions. *Journal of Functional Analysis* 220: 157–213.

Gérard C and Jäkel C (2005) Thermal quantum fields without cutoffs in $1+1$ space-time dimensions. *Reviews in Mathematical Physics* 17: 113–173.

Haag R (1992) *Local Quantum Physics*. Berlin: Springer.

Høegh-Krohn R (1974) Relativistic quantum statistical mechanics in two-dimensional space-time. *Communications in Mathematical Physics* 38: 195–224.

Jäkel CD (2004) The relation between KMS states for different temperatures. *Annales de l'Institut Henri Poincaré* 5: 1–30.

Kopper C, Müller VF, and Reisz T (2001) Temperature independant renormalization of finite temperature field theory. *Annales de l'Institut Henri Poincaré* 2: 387–402.

Kuckert B (2002) Covariant thermodynamics of quantum systems: passivity, semipassivity, and the Unruh effect. *Annals of Physics* 295: 216–229.

Landsman NP and van Weert Ch G (1987) Real- and imaginary-time field theory at finite temperature and density. *Physics Reports* 145: 141–249.

Narnhofer H (1994) Entropy density for relativistic quantum field theory. *Reviews in Mathematical Physics* 6: 1127–1145.

Narnhofer H, Requardt M, and Thirring W (1983) Quasiparticles at finite temperature. *Communications in Mathematical Physics* 92: 247–268.

Steinmann O (1995) Perturbative quantum field theory at positive temperatures: an axiomatic approach. *Communications in Mathematical Physics* 170: 405–415.

Thermohydraulics *see* Newtonian Fluids and Thermohydraulics

Tomita–Takesaki Modular Theory

S J Summers, University of Florida, Gainesville, FL, USA

© 2006 Elsevier Ltd. All rights reserved.

Basic Structure

The origins of Tomita–Takesaki modular theory lie in two unpublished papers of M Tomita in 1967 and a slim volume by Takesaki (1970). It has developed into one of the most important tools in the theory of operator algebras and has found many applications in mathematical physics.

Although the modular theory has been formulated in a more general setting, it will be presented in the form in which it most often finds application in mathematical physics (for generalizations, details, and further references concerning the material covered in this article, the reader is referred to the Further Reading section). Let $\mathcal{M}$ be a von Neumann algebra on a Hilbert space $\mathcal{H}$ containing a vector Ω which is cyclic and separating for $\mathcal{M}$. Define the operator S_0 on $\mathcal{H}$ as follows:

$$S_0 A\Omega = A^*\Omega, \quad \text{for all } A \in \mathcal{M}$$

This operator extends to a closed antilinear operator S defined on a dense subset of $\mathcal{H}$. Let Δ be the unique positive, self-adjoint operator and J the unique antiunitary operator occurring in the polar decomposition

$$S = J\Delta^{1/2} = \Delta^{-1/2}J$$

Δ is called the modular operator and J the modular conjugation (or modular involution) associated with the pair $(\mathcal{M}, \Omega)$. Note that J^2 is the identity operator and $J = J^*$. Moreover, the spectral calculus may be applied to Δ so that Δ^{it} is a unitary operator for each $t \in \mathbb{R}$ and $\{\Delta^{it} \mid t \in \mathbb{R}\}$ forms a strongly continuous unitary group. Let $\mathcal{M}'$ denote the set of all bounded linear operators on $\mathcal{H}$ which commute with all elements of $\mathcal{M}$. The modular theory begins with the following remarkable theorem.

Theorem 1 *Let $\mathcal{M}$ be a von Neumann algebra with a cyclic and separating vector Ω. Then $J\Omega = \Omega = \Delta\Omega$, and the following equalities hold:*

$$J\mathcal{M}J = \mathcal{M}'$$

and

$$\Delta^{it}\mathcal{M}\Delta^{-it} = \mathcal{M}, \quad \text{for all } t \in \mathbb{R}$$

Note that if one defines $F_0 A'\Omega = A'^*\Omega$, for all $A' \in \mathcal{M}'$, and takes its closure F, then one has the relations

$$\Delta = FS, \quad \Delta^{-1} = SF, \quad F = J\Delta^{-1/2}$$

Notes:

Modular Automorphism Group

By Theorem 1, the unitaries $\Delta^{it}, t \in \mathbb{R}$, induce a one-parameter automorphism group $\{\sigma_t\}$ of $\mathcal{M}$ by

$$\sigma_t(A) = \Delta^{it} A \Delta^{-it}, \quad A \in \mathcal{M}, \quad t \in \mathbb{R}$$

This group is called the modular automorphism group of $\mathcal{M}$ (relative to Ω). Let ω denote the faithful normal state on $\mathcal{M}$ induced by Ω:

$$\omega(A) = \frac{1}{\|\Omega\|^2} \langle \Omega, A\Omega \rangle, \quad A \in \mathcal{M}$$

From Theorem 1 it follows that ω is invariant under $\{\sigma_t\}$, that is, $\omega(\sigma_t(A)) = \omega(A)$ for all $A \in \mathcal{M}$ and $t \in \mathbb{R}$.

The modular automorphism group contains information about both $\mathcal{M}$ and ω. For example, the modular automorphism group is an inner automorphism on $\mathcal{M}$ if and only if $\mathcal{M}$ is semifinite. It is trivial if and only if ω is a tracial state on $\mathcal{M}$. Indeed, for any $B \in \mathcal{M}$, one has $\sigma_t(B) = B$ for all $t \in \mathbb{R}$ if and only if $\omega(AB) = \omega(BA)$ for all $A \in \mathcal{M}$. Let $\mathcal{M}^\sigma$ denote the set of all such B in $\mathcal{M}$.

The KMS Condition

The modular automorphism group satisfies a condition which had already been used in mathematical physics to characterize equilibrium temperature states of quantum systems in statistical mechanics and field theory – the Kubo–Martin–Schwinger (KMS) condition. If $\mathcal{M}$ is a von Neumann algebra and $\{\alpha_t \mid t \in \mathbb{R}\}$ is a σ-weakly continuous one-parameter group of automorphisms of $\mathcal{M}$, then the state ϕ on $\mathcal{M}$ satisfies the KMS condition at (inverse temperature) β $(0 < \beta < \infty)$ with respect to $\{\alpha_t\}$ if for any $A,B \in \mathcal{M}$ there exists a complex function $F_{AB}(z)$ which is analytic on the strip $\{z \in \mathbb{C} \mid 0 < \operatorname{Im} z < \beta\}$ and continuous on the closure of this strip such that

$$F_{A,B}(t) = \phi(\alpha_t(A)B)$$
$$F_{A,B}(t + i\beta) = \phi(B\alpha_t(A))$$

for all $t \in \mathbb{R}$. In this case, $\phi(\alpha_{i\beta}(A)B) = \phi(BA)$, for all A, B in a σ-weakly dense, α-invariant $*$-subalgebra of $\mathcal{M}$. Such KMS states are α-invariant, that is, $\phi(\alpha_t(A)) = \phi(A)$, for all $A \in \mathcal{M}, t \in \mathbb{R}$, and are stable and passive (cf. Bratteli and Robinson (1981) and Haag (1992)).

Every faithful normal state satisfies the KMS condition at $\beta = 1$ (henceforth called the modular condition) with respect to the corresponding modular automorphism group.

Theorem 2 *Let $\mathcal{M}$ be a von Neumann algebra with a cyclic and separating vector Ω. Then the induced state ω on $\mathcal{M}$ satisfies the modular condition with respect to the modular automorphism group $\{\sigma_t \mid t \in \mathbb{R}\}$ associated to the pair $(\mathcal{M}, \Omega)$.*

The modular automorphism group is, therefore, endowed with the analyticity associated with the KMS condition, and this is a powerful tool in many applications of the modular theory to mathematical physics. In addition, the physical properties and interpretations of KMS states are often invoked when applying modular theory to quantum physics.

Note that while the nontriviality of the modular automorphism group gives a measure of the nontracial nature of the state, the KMS condition for the modular automorphism group provides the missing link between the values $\omega(AB)$ and $\omega(BA)$, for all $A,B \in \mathcal{M}$ (hence the use of the term "modular," as in the theory of integration on locally compact groups).

The modular condition is quite restrictive. Only the modular group can satisfy the modular condition for $(\mathcal{M}, \Omega)$, and the modular group for one state can satisfy the modular condition only in states differing from the original state by the action of an element in the center of $\mathcal{M}$.

Theorem 3 *Let $\mathcal{M}$ be a von Neumann algebra with a cyclic and separating vector Ω, and let $\{\sigma_t\}$ be the corresponding modular automorphism group. If the induced state ω satisfies the modular condition with respect to a group $\{\alpha_t\}$ of automorphisms of $\mathcal{M}$, then $\{\alpha_t\}$ must coincide with $\{\sigma_t\}$. Moreover, a normal state ψ on $\mathcal{M}$ satisfies the modular condition with respect to $\{\sigma_t\}$ if and only if $\psi(\cdot) = \omega(h \cdot) = \omega(h^{1/2} \cdot h^{1/2})$ for some unique positive injective operator h affiliated with the center of $\mathcal{M}$.*

Hence, if $\mathcal{M}$ is a factor, two distinct states cannot share the same modular automorphism group. The relation between the modular automorphism groups for two different states will be described in more detail.

One Algebra and Two States

Consider a von Neumann algebra $\mathcal{M}$ with two cyclic and separating vectors Ω and Φ, and denote by ω and ϕ, respectively, the induced states on $\mathcal{M}$. Let $\{\sigma_t^\omega\}$ and $\{\sigma_t^\phi\}$ denote the corresponding modular groups. There is a general relation between the modular automorphism groups of these states.

Theorem 4 *There exists a σ-strongly continuous map $\mathbb{R} \ni t \mapsto U_t \in \mathcal{M}$ such that*

(i) *U_t is unitary for all $t \in \mathbb{R}$;*
(ii) *$U_{t+s} = U_t \sigma_t^{\omega}(U_s)$ for all $s,t \in \mathbb{R}$; and*
(iii) *$\sigma_t^{\phi}(A) = U_t \sigma_t^{\omega}(A) U_t^*$ for all $A \in \mathcal{M}$ and $t \in \mathbb{R}$.*

The 1-cocycle $\{U_t\}$ is commonly called the cocycle derivative of ϕ with respect to ω and one writes $U_t = (D\phi : D\omega)_t$. There is a chain rule for this derivative, as well: If ϕ,ψ, and ρ are faithful normal states on $\mathcal{M}$, then $(D\psi : D\phi)_t = (D\psi : D\rho)_t (D\rho : D\phi)_t$, for all $t \in \mathbb{R}$. More can be said about the cocycle derivative if the states satisfy any of the conditions in the following theorem.

Theorem 5 *The following conditions are equivalent:*

(i) *ϕ is $\{\sigma_t^{\omega}\}$-invariant;*
(ii) *ω is $\{\sigma_t^{\phi}\}$-invariant;*
(iii) *there exists a unique positive injective operator h affiliated with $\mathcal{M}^{\sigma^{\omega}} \cap \mathcal{M}^{\sigma^{\phi}}$ such that $\omega(\,\cdot\,) = \phi(h\,\cdot\,) = \phi(h^{1/2}\,\cdot\,h^{1/2})$;*
(iv) *there exists a unique positive injective operator h' affiliated with $\mathcal{M}^{\sigma^{\omega}} \cap \mathcal{M}^{\sigma^{\phi}}$ such that $\phi(\,\cdot\,) = \omega(h'\,\cdot\,) = \omega(h'^{1/2}\,\cdot\,h'^{1/2})$;*
(v) *the norms of the linear functionals $\omega + \mathrm{i}\phi$ and $\omega - \mathrm{i}\phi$ are equal; and*
(vi) *$\sigma_t^{\omega}\sigma_s^{\phi} = \sigma_s^{\phi}\sigma_t^{\omega}$, for all $s,t \in \mathbb{R}$.*

The conditions in Theorem 5 turn out to be equivalent to the cocycle derivative being a representation.

Theorem 6 *The cocycle $\{U_t\}$ intertwining $\{\sigma_t^{\omega}\}$ with $\{\sigma_t^{\phi}\}$ is a group representation of the additive group of reals if and only if ϕ and ω satisfy the conditions in Theorem 5. In that case, $U(t) = h^{-\mathrm{it}}$.*

The operator $h' = h^{-1}$ in Theorem 5 is called the Radon–Nikodym derivative of ϕ with respect to ω (often denoted by $\mathrm{d}\phi/\mathrm{d}\omega$), due to the following result, which, if the algebra $\mathcal{M}$ is abelian, *is* the well-known Radon–Nikodym theorem from measure theory.

Theorem 7 *If ϕ and ω are normal positive linear functionals on $\mathcal{M}$* such that *$\phi(A) \leq \omega(A)$, for all positive elements $A \in \mathcal{M}$, then there exists a unique element $h^{1/2} \in \mathcal{M}$ such that $\phi(\,\cdot\,) = \omega(h^{1/2}\,\cdot\,h^{1/2})$ and $0 \leq h^{1/2} \leq 1$.*

The analogies with measure theory are not accidental, although these are not discussed in detail here. Indeed, any normal trace on a (finite) von Neumann algebra $\mathcal{M}$ gives rise to a noncommutative integration theory in a natural manner. Modular theory affords an extension of this theory to the setting of faithful normal functionals η on von Neumann algebras $\mathcal{M}$ of any type, enabling the definition of noncommutative L^p spaces, $L^p(\mathcal{M},\eta)$.

Modular Invariants and the Classification of von Neumann Algebras

As already mentioned, the modular structure carries information about the algebra. This is best evidenced in the structure of type III factors. As this theory is rather involved, only a sketch of some of the results can be given.

If $\mathcal{M}$ is a type III algebra, then its crossed product $\mathcal{N} = \mathcal{M} \rtimes_{\sigma^{\omega}} \mathbb{R}$ relative to the modular automorphism group of any faithful normal state ω on $\mathcal{M}$ is a type II_{∞} algebra with a faithful semifinite normal trace τ such that $\tau \circ \theta_t = \mathrm{e}^{-t}\tau$, $t \in \mathbb{R}$, where θ is the dual of σ^{ω} on $\mathcal{N}$. Moreover, the algebra $\mathcal{M}$ is isomorphic to the cross product $\mathcal{N} \rtimes_{\theta} \mathbb{R}$, and this decomposition is unique in a very strong sense. This structure theorem entails the existence of important algebraic invariants for $\mathcal{M}$, which has many consequences, one of which is made explicit here.

If ω is a faithful normal state of a von Neumann algebra $\mathcal{M}$ induced by Ω, let Δ_{ω} denote the modular operator associated to $(\mathcal{M},\Omega)$ and $\mathrm{sp}\,\Delta_{\omega}$ denote the spectrum of Δ_{ω}. The intersection

$$S'(\mathcal{M}) = \cap\, \mathrm{sp}\,\Delta_{\omega}$$

over all faithful normal states ω of $\mathcal{M}$ is an algebraic invariant of $\mathcal{M}$.

Theorem 8 *Let $\mathcal{M}$ be a factor acting on a separable Hilbert space. If $\mathcal{M}$ is of type* III, *then $0 \in S'(\mathcal{M})$; otherwise, $S'(\mathcal{M}) = \{0,1\}$ if $\mathcal{M}$ is of type I_{∞} or II_{∞} and $S'(\mathcal{M}) = \{1\}$ if not. Let $\mathcal{M}$ now be a factor of type* III.

(i) *$\mathcal{M}$ is of type $\mathrm{III}_{\lambda}, 0 < \lambda < 1$, if and only if $S'(\mathcal{M}) = \{0\} \cup \{\lambda^n \mid n \in \mathbb{Z}\}$.*
(ii) *$\mathcal{M}$ is of type III_0 if and only if $S'(\mathcal{M}) = \{0,1\}$.*
(iii) *$\mathcal{M}$ is of type III_1 if and only if $S'(\mathcal{M}) = [0,\infty)$.*

In certain physically relevant situations, the spectra of the modular operators of all faithful normal states coincide, so that Theorem 8 entails that it suffices to compute the spectrum of any conveniently chosen modular operator in order to determine the type of $\mathcal{M}$. In other such situations, there are distinguished states ω such that $S'(\mathcal{M}) = \mathrm{sp}\,\Delta_{\omega}$. One such example is provided by asymptotically abelian systems. A von Neumann algebra $\mathcal{M}$ is said to be "asymptotically abelian" if there exists a sequence $\{\alpha_n\}_{n\in\mathbb{N}}$ of automorphisms of $\mathcal{M}$ such that the limit of $\{A\alpha_n(B) - \alpha_n(B)A\}_{n\in\mathbb{N}}$ in

the strong operator topology is zero, for all $A, B \in \mathcal{M}$. If the state ω is α_n-invariant, for all $n \in \mathbb{N}$, then $\mathrm{sp}\,\Delta_\omega$ is contained in $\mathrm{sp}\,\Delta_\phi$, for all faithful normal states ϕ on $\mathcal{M}$, so that $S'(\mathcal{M}) = \mathrm{sp}\,\Delta_\omega$. If, moreover, $\mathrm{sp}\,\Delta_\omega = [0, \infty)$, then $\mathrm{sp}\,\Delta_\omega = \mathrm{sp}\,\Delta_\phi$, for all ϕ as described.

Self-Dual Cones

Let $j: \mathcal{M} \to \mathcal{M}'$ denote the antilinear $*$-isomorphism defined by $j(A) = JAJ, A \in \mathcal{M}$. The natural positive cone $\mathcal{P}^\natural$ associated with the pair $(\mathcal{M}, \Omega)$ is defined as the closure, in $\mathcal{H}$, of the set of vectors

$$\{Aj(A)\Omega \mid A \in \mathcal{M}\}$$

Let $\mathcal{M}_+$ denote the set of all positive elements of $\mathcal{M}$. The following theorem collects the main attributes of the natural cone.

Theorem 9

(i) *$\mathcal{P}^\natural$ coincides with the closure in $\mathcal{H}$ of the set $\{\Delta^{1/4} A\Omega \mid A \in \mathcal{M}_+\}$.*
(ii) *$\Delta^{it}\mathcal{P}^\natural = \mathcal{P}^\natural$ for all $t \in \mathbb{R}$.*
(iii) *$J\Phi = \Phi$ for all $\Phi \in \mathcal{P}^\natural$.*
(iv) *$Aj(A)\mathcal{P}^\natural \subset \mathcal{P}^\natural$ for all $A \in \mathcal{M}$.*
(v) *$\mathcal{P}^\natural$ is a pointed, self-dual cone whose linear span coincides with $\mathcal{H}$.*
(vi) *If $\Phi \in \mathcal{P}^\natural$, then Φ is cyclic for $\mathcal{M}$ if and only if Φ is separating for $\mathcal{M}$.*
(vii) *If $\Phi \in \mathcal{P}^\natural$ is cyclic, and hence separating, for $\mathcal{M}$, then the modular conjugation and the natural cone associated with the pair $(\mathcal{M}, \Phi)$ coincide with J and $\mathcal{P}^\natural$, respectively.*
(viii) *For every normal positive linear functional ϕ on $\mathcal{M}$, there exists a unique vector $\Phi_\phi \in \mathcal{P}^\natural$ such that $\phi(A) = \langle \Phi_\phi, A\Phi_\phi \rangle$ for all $A \in \mathcal{M}$.*

In fact, the algebras $\mathcal{M}$ and $\mathcal{M}'$ are uniquely characterized by the natural cone $\mathcal{P}^\natural$ [4]. In light of (viii), if α is an automorphism of $\mathcal{M}$, then

$$V(\alpha)\Phi_\phi = \Phi_{\phi \circ \alpha^{-1}}$$

defines an isometric operator on $\mathcal{P}^\natural$, which by (v) extends to a unitary operator on $\mathcal{H}$. The map $\alpha \mapsto V(\alpha)$ defines a unitary representation of the group of automorphisms $\mathrm{Aut}(\mathcal{M})$ on $\mathcal{M}$ in such a manner that $V(\alpha)AV(\alpha)^{-1} = \alpha(A)$ for all $A \in \mathcal{M}$ and $\alpha \in \mathrm{Aut}(\mathcal{M})$. Indeed, one has the following:

Theorem 10 *Let $\mathcal{M}$ be a von Neumann algebra with a cyclic and separating vector Ω. The group $\mathcal{V}$ of all unitaries V satisfying*

$$V\mathcal{M}V^* = \mathcal{M}, \quad VJV^* = J, \quad V\mathcal{P}^\natural = \mathcal{P}^\natural$$

is isomorphic to $\mathrm{Aut}(\mathcal{M})$ under the above map $\alpha \mapsto V(\alpha)$, which is called the "standard implementation" of $\mathrm{Aut}(\mathcal{M})$.

Often of particular physical interest are (anti-)automorphisms of $\mathcal{M}$ leaving ω invariant. They can only be implemented by (anti)unitaries which leave the pair $(\mathcal{M}, \Omega)$ invariant. In fact, if U is a unitary or antiunitary operator satisfying $U\Omega = \Omega$ and $U\mathcal{M}U^* = \mathcal{M}$, then U commutes with both J and Δ.

Two Algebras and One State

Motivated by applications to quantum field theory, the study of the modular structures associated with one state and more than one von Neumann algebra has begun (see Borchers (2000) for references and details). Let $\mathcal{N} \subset \mathcal{M}$ be von Neumann algebras with a common cyclic and separating vector Ω, and $\Delta_\mathcal{N}, J_\mathcal{N}$ and $\Delta_\mathcal{M}, J_\mathcal{M}$ denote the corresponding modular objects. The structure $(\mathcal{M}, \mathcal{N}, \Omega)$ is called a $\pm$-half-sided modular inclusion if $\Delta_\mathcal{M}^{it}\mathcal{N}\Delta_\mathcal{M}^{-it} \subset \mathcal{N}$, for all $\pm t \geq 0$.

Theorem 11 *Let $\mathcal{M}$ be a von Neumann algebra with cyclic and separating vector Ω. The following are equivalent:*

(i) *There exists a proper subalgebra $\mathcal{N} \subset \mathcal{M}$ such that $(\mathcal{M}, \mathcal{N}, \Omega)$ is a $\mp$-half-sided modular inclusion.*
(ii) *There exists a unitary group $\{U(t)\}$ with positive generator such that*

$$U(t)\mathcal{M}U(t)^{-1} \subset \mathcal{M}, \quad \text{for all } \pm t \geq 0,$$
$$U(t)\Omega = \Omega, \quad \text{for all } t \in \mathbb{R}$$

Moreover, if these conditions are satisfied, then the following relations must hold:

$$\Delta_\mathcal{M}^{it} U(s) \Delta_\mathcal{M}^{-it} = \Delta_\mathcal{N}^{it} U(s) \Delta_\mathcal{N}^{-it} = U(e^{\mp 2\pi t} s)$$

and

$$J_\mathcal{M} U(s) J_\mathcal{M} = J_\mathcal{N} U(s) J_\mathcal{N} = U(-s)$$

for all $s, t \in \mathbb{R}$. In addition, $\mathcal{N} = U(\pm 1)\mathcal{M}U(\pm 1)^{-1}$, and if $\mathcal{M}$ is a factor, it must be type III_1.

The richness of this structure is further suggested by the next theorem.

Theorem 12

(i) *Let $(\mathcal{M}, \mathcal{N}_1, \Omega)$ and $(\mathcal{M}, \mathcal{N}_2, \Omega)$ be $-$-half-sided, resp. $+$-half-sided, modular inclusions satisfying the condition $J_{\mathcal{N}_1} J_{\mathcal{N}_2} = J_\mathcal{M} J_{\mathcal{N}_2} J_{\mathcal{N}_1} J_\mathcal{M}$. Then the modular unitaries $\Delta_\mathcal{M}^{it}, \Delta_{\mathcal{N}_1}^{is}, \Delta_{\mathcal{N}_2}^{iu}, s, t, u \in \mathbb{R}$, generate a faithful continuous unitary representation of the identity component of the*

group of isometries of two-dimensional Minkowski space.

(ii) *Let* $\mathcal{M}, \mathcal{N}, \mathcal{N} \cap \mathcal{M}$ *be von Neumann algebras with a common cyclic and separating vector* Ω. *If* $(\mathcal{M}, \mathcal{M} \cap \mathcal{N}, \Omega)$ *and* $(\mathcal{N}, \mathcal{M} \cap \mathcal{N}, \Omega)$ *are* $-$*-half-sided, resp.* $+$*-half-sided, modular inclusions such that* $J_{\mathcal{N}} \mathcal{M} J_{\mathcal{N}} = \mathcal{M}$, *then the modular unitaries* $\Delta_{\mathcal{M}}^{it}, \Delta_{\mathcal{N}}^{is}, \Delta_{\mathcal{N} \cap \mathcal{M}}^{iu}, s, t, u \in \mathbb{R}$, *generate a faithful continuous unitary representation of* $\mathrm{SL}(2, \mathbb{R})/\mathbb{Z}_2$.

This has led to a further useful notion. If $\mathcal{N} \subset \mathcal{M}$ and Ω is cyclic for $\mathcal{N} \cap \mathcal{M}$, then $(\mathcal{M}, \mathcal{N}, \Omega)$ is said to be a "$\pm$-modular intersection" if both $(\mathcal{M}, \mathcal{M} \cap \mathcal{N}, \Omega)$ and $(\mathcal{N}, \mathcal{M} \cap \mathcal{N}, \Omega)$ are $\pm$-half-sided modular inclusions and

$$J_{\mathcal{N}} \left[\lim_{t \to \mp\infty} \Delta_{\mathcal{N}}^{it} \Delta_{\mathcal{M}}^{-it} \right] J_{\mathcal{N}} = \lim_{t \to \mp\infty} \Delta_{\mathcal{M}}^{it} \Delta_{\mathcal{N}}^{-it}$$

where the existence of the strong operator limits is assured by the preceding assumptions. An example of the utility of this structure is the following theorem.

Theorem 13 *Let* $\mathcal{N}, \mathcal{M}, \mathcal{L}$ *be von Neumann algebras with a common cyclic and separating vector* Ω. *If* $(\mathcal{M}, \mathcal{N}, \Omega)$ *and* $(\mathcal{N}', \mathcal{L}, \Omega)$ *are* $-$*-modular intersections and* $(\mathcal{M}, \mathcal{L}, \Omega)$ *is a* $+$*-modular intersection, then the unitaries* $\Delta_{\mathcal{M}}^{it}, \Delta_{\mathcal{N}}^{is}, \Delta_{\mathcal{L}}^{iu}, s, t, u \in \mathbb{R}$, *generate a faithful continuous unitary representation of* $\mathrm{SO}^{\uparrow}(1, 2)$.

These results and their extensions to larger numbers of algebras were developed for application in algebraic quantum field theory, but one may anticipate that half-sided modular inclusions will find wider use. Modular theory has also been applied fruitfully in the theory of inclusions $\mathcal{N} \subset \mathcal{M}$ of properly infinite algebras with finite or infinite index.

Applications in Quantum Theory

The Tomita–Takesaki theory has found many applications in quantum field theory and quantum statistical mechanics. As mentioned earlier, the modular automorphism group satisfies the KMS condition, a property of physical significance in the quantum theory of many-particle systems, which includes quantum statistical mechanics and quantum field theory. In such settings, for a suitable algebra of observables $\mathcal{M}$ and state ω, an automorphism group $\{\sigma_{\beta t}\}$ representing the time evolution of the system satisfies the modular condition. Hence, on the one hand, $\{\sigma_{\beta t}\}$ is the modular automorphism group of the pair $(\mathcal{M}, \Omega)$, and, on the other, ω is an equilibrium state at inverse temperature β, with all the consequences which both of these facts have.

But it has become increasingly clear that the modular objects Δ^{it}, J, of certain algebras of observables and states encode additional physical information. In 1975, it was discovered that if one considers the algebras of observables associated with a finite-component quantum field theory satisfying the Wightman axioms, then the modular objects associated with the vacuum state and algebras of observables localized in certain wedge-shaped regions in Minkowski space have geometric content. In fact, the unitary group $\{\Delta^{it}\}$ implements the group of Lorentz boosts leaving the wedge region invariant (this property is now called modular covariance), and the modular involution J implements the space-time reflection about the edge of the wedge, along with a charge conjugation. This discovery caused some intense research activity (see Baumgartel and Wollenberg 1992, Borchers 2000, Haag 1992).

Positive Energy

In quantum physics the time development of the system is often represented by a strongly continuous group $\{U(t) = e^{itH} \mid t \in \mathbb{R}\}$ of unitary operators, and the generator H is interpreted as the total energy of the system. There is a link between modular structure and positive energy, which has found many applications in quantum field theory. This result was crucial in the development of Theorem 11 and was motivated by the 1975 discovery mentioned above, now commonly called the Bisognano–Wichmann theorem.

Theorem 14 *Let* $\mathcal{M}$ *be a von Neumann algebra with a cyclic and separating vector* Ω, *and let* $\{U(t)\}$ *be a continuous unitary group satisfying* $U(t) \mathcal{M} U(-t) \subset \mathcal{M}$, *for all* $t \geq 0$. *Then any two of the following conditions imply the third*:

(i) $U(t) = e^{itH}$, *with* $H \geq 0$;
(ii) $U(t)\Omega = \Omega$, *for all* $t \in \mathbb{R}$; *and*
(iii) $\Delta^{it} U(s) \Delta^{-it} = U(e^{-2\pi t} s)$ *and* $JU(s)J = U(-s)$, *for all* $s, t \in \mathbb{R}$.

Modular Nuclearity and Phase Space Properties

Modular theory can be used to express physically meaningful properties of quantum "phase spaces" by a condition of compactness or nuclearity of certain maps. In its initial form, the condition was formulated in terms of the Hamiltonian, the global energy operator of theories in Minkowski space. The above indications that the modular operators carry information about the energy of the system were reinforced when it was shown that a

formulation in terms of modular operators was essentially equivalent.

Let $\mathcal{O}_1 \subset \mathcal{O}_2$ be nonempty bounded open subregions of Minkowski space with corresponding algebras of observables $\mathcal{A}(\mathcal{O}_1) \subset \mathcal{A}(\mathcal{O}_2)$ in a vacuum representation with vacuum vector Ω, and let Δ be the modular operator associated with $(\mathcal{A}(\mathcal{O}_2), \Omega)$ (by the Reeh–Schlieder theorem, Ω is cyclic and separating for $\mathcal{A}(\mathcal{O}_2)$). For each $\lambda \in (0, 1/2)$ define the mapping $\Xi_\lambda : \mathcal{A}(\mathcal{O}_1) \to \mathcal{H}$ by $\Xi_\lambda(A) = \Delta^\lambda A\Omega$. The compactness of any one of these mappings implies the compactness of all of the others. Moreover, the l^p (nuclear) norms of these mappings are interrelated and provide a measure of the number of local degrees of freedom of the system. Suitable conditions on the maps in terms of these norms entail the strong statistical independence condition called the split property. Conversely, the split property implies the compactness of all of these maps. Moreover, the existence of equilibrium temperature states on the global algebra of observables can be derived from suitable conditions on these norms in the vacuum sector.

The conceptual advantage of the modular compactness and nuclearity conditions compared to their original Hamiltonian form lies in the fact that they are meaningful also for quantum systems in curved spacetimes, where global energy operators (i.e., generators corresponding to global timelike Killing vector fields) need not exist.

Modular Position and Quantum Field Theory

The characterization of the relative "geometric" position of algebras based on the notions of modular inclusion and modular intersection was directly motivated by the Bisognano–Wichmann theorem. Observable algebras associated with suitably chosen wedge regions in Minkowski space provided examples whose essential structure could be abstracted for more general application, resulting in the notions presented in the preceding sections.

Theorem 12(ii) has been used to construct, from two algebras and the indicated half-sided modular inclusions, a conformal quantum field theory on the circle (compactified light ray) with positive energy. Since the chiral part of a conformal quantum field model in two spacetime dimensions naturally yields such half-sided modular inclusions, studying the inclusions in Theorem 12(ii) is equivalent to studying such field theories. Theorems 12(i) and 13 and their generalizations to inclusions involving up to six algebras have been employed to construct Poincaré-covariant nets of observable algebras (the algebraic form of quantum field theories) satisfying the spectrum condition on $(d+1)$-dimensional Minkowski space for $d = 1, 2, 3$. Conversely, such quantum field theories naturally yield such systems of algebras.

This intimate relation would seem to open up the possibility of constructing interacting quantum field theories from a limited number of modular inclusions/intersections.

Geometric Modular Action

The fact that the modular objects in quantum field theory associated with wedge-shaped regions and the vacuum state in Minkowski space have geometric significance ("geometric modular action") was originally discovered in the framework of the Wightman axioms. As an algebraic quantum field theory (AQFT) does not rely on the concept of Wightman fields, it was natural to ask (i) when does geometric modular action hold in AQFT and (ii) which physically relevant consequences follow from this feature?

There are two approaches to the study of geometric modular action. In the first, attention is focused on modular covariance, expressed in terms of the modular groups associated with wedge algebras and the vacuum state in Minkowski space. Modular covariance has been proven to obtain in conformally invariant AQFT, in any massive theory satisfying asymptotic completeness, and also in the presence of other, physically natural assumptions. To mention only three of its consequences, both the spin–statistics theorem and the PCT theorem, as well as the existence of a continuous unitary representation of the Poincaré group acting covariantly upon the observable algebras and satisfying the spectrum condition follow from modular covariance.

In a second approach to geometric modular action, the modular involutions are the primary focus. Here, no *a priori* connection between the modular objects and isometries of the spacetime is assumed. The central assumption, given the state vector Ω and the von Neumann algebras of localized observables $\{\mathcal{A}(\mathcal{O})\}$ on the spacetime, is that there exists a family $\mathcal{W}$ of subsets of the spacetime such that $J_{\mathcal{W}_1}\mathcal{R}(\mathcal{W}_2)J_{\mathcal{W}_1} \in \{\mathcal{R}(\mathcal{W}) \mid \mathcal{W} \in \mathcal{W}\}$, for every $\mathcal{W}_1, \mathcal{W}_2 \in \mathcal{W}$. This condition makes no explicit appeal to isometries or other special attributes and is thus applicable, in principle, to quantum field theories on general curved spacetimes.

It has been shown for certain spacetimes, including Minkowski space, that under certain additional technical assumptions, the modular involutions encode enough information to determine the dynamics of the theory, the isometry group of the spacetime, and a continuous unitary representation of the isometry group which acts covariantly upon the observables and leaves the state invariant. In certain

cases including Minkowski space, it is even possible to derive the spacetime itself from the group $\mathcal{J}$ generated by the modular involutions $\{J_W \mid W \in \mathcal{W}\}$.

The modular unitaries Δ_W^{it} enter in this approach through a condition which is designed to assure the stability of the theory, namely that $\Delta_W^{it} \in \mathcal{J}$, for all $t \in \mathbb{R}$ and $W \in \mathcal{W}$. In Minkowski space, this additional condition entails that the derived representation of the Poincaré group satisfies the spectrum condition.

Further Applications

As previously observed, through the close connection to the KMS condition, modular theory enters naturally into the equilibrium thermodynamics of many-body systems. But in recent work on the theory of nonequilibrium thermodynamics it also plays a role in making mathematical sense of the notion of quantum systems in local thermodynamic equilibrium. Modular theory has also proved to be of utility in recent developments in the theory of superselection rules and their attendant sectors, charges and charge-carrying fields.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Quantum Central-Limit Theorems; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Thermal Quantum Field Theory; Positive Maps on C^*-Algebras; Two-Dimensional Models; von Neumann Algebras: Introduction, Modular Theory, and Classification Theory.

Further Reading

Baumgärtel H and Wollenberg M (1992) *Causal Nets of Operator Algebras*. Berlin: Akademie-Verlag.

Borchers HJ (2000) On revolutionizing quantum field theory with Tomita's modular theory. *Journal of Mathematical Physics* 41: 3604–3673.

Bratteli O and Robinson DW (1981) *Operator Algebras and Quantum Statistical Mechanics II*. Berlin: Springer.

Connes A (1974) Caractérisation des algèbres de von Neumann comme espaces vectoriels ordonnés. *Annales de l'Institut Fourier* 24: 121–155.

Haag R (1992) *Local Quantum Physics*. Berlin: Springer.

Kadison RV and Ringrose JR (1986) *Fundamentals of the Theory of Operator Algebras*, vol. II. Orlando: Academic Press.

Pedersen GK (1979) *C^*-Algebras and Their Automorphism Groups*. New York: Academic Press.

Stratila S (1981) *Modular Theory in Operator Algebras*. Tunbridge Wells: Abacus Press.

Takesaki M (1970) *Tomita's Theory of Modular Hilbert Algebras and Its Applications*, Lecture Notes in Mathematics, vol. 128. Berlin: Springer.

Takesaki M (2003) *Theory of Operator Algebras II*. Berlin: Springer.

Topological Defects and Their Homotopy Classification

T W B Kibble, Imperial College, London, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Symmetry-breaking phase transitions occur in a wide variety of systems – from condensed matter to the early universe. One of the common features of such transitions is the appearance, in the broken-symmetry phase, of topological defects, trapped regions in which the symmetry is restored, or at least changed. Examples are vortices in superfluids, domain walls in ferromagnets, and disclination lines in liquid crystals. Often these defects are stable for topological reasons, and play an important role in the dynamics of the system. An astonishingly rich variety of defects can be found in various systems. They can usefully be classified using the tools of homotopy theory.

Spontaneous Symmetry Breaking

Let us consider a quantum-mechanical system with a symmetry group G. This means that each $g \in G$ is represented on the Hilbert space of quantum states by a unitary operator $\hat{U}(g)$, which commutes with the Hamiltonian. Spontaneous symmetry breaking occurs if this symmetry is not shared by the ground state or vacuum state $|0\rangle$ of the system. In other words, for some $g \in G, \hat{U}(g)|0\rangle \neq |0\rangle$. Then the ground state is necessarily degenerate: $\hat{U}(g)|0\rangle$ must have the same energy as $|0\rangle$.

Spontaneous symmetry breaking is usually describable in terms of an order-parameter field, which vanishes above the transition and is nonzero below it. We can find a scalar field $\hat{\phi}(\boldsymbol{r})$, or multiplet of fields $\hat{\phi} = (\hat{\phi}_i, i = 1, \ldots, n)$ transforming according to some representation D of G (assumed not to contain the trivial representation), whose expectation value in the ground state is nonzero:

$$\langle 0|\hat{\phi}(\boldsymbol{r})|0\rangle = \phi_0 \neq 0 \qquad [1]$$

This is the order parameter. Since

$$\langle 0|\hat{U}^{\dagger}(g)\hat{\phi}(\boldsymbol{r})\hat{U}(g)|0\rangle = D(g)\phi_0 \qquad [2]$$

it follows that the only elements of G that can be symmetries of the ground state are those in the

Notes:

stability subgroup H of ϕ_0 (the group of unbroken symmetries in this ground state):

$$H = \{g \in G:\ D(g)\phi_0 = \phi_0\} \qquad [3]$$

In terms of this subgroup, we can find a useful characterization of the manifold $\mathcal{M}$ of degenerate ground states. As noted above, for each $g \in G$, $\hat{U}(g)|0\rangle$ is also a ground state. However, these are not all distinct, because clearly $\hat{U}(gh)|0\rangle = \hat{U}(g)|0\rangle$ for all $h \in H$. Hence, the distinct ground states are in one-to-one correspondence with the left cosets gH of H in G, and $\mathcal{M}$ may be identified with the quotient space G/H, the space of left cosets.

For example, suppose G is the rotation group SO(3), and $\hat{\boldsymbol{\phi}}$ belongs to the three-dimensional vector representation. If $\boldsymbol{\phi} \neq 0$ in the ground state, we may choose $\boldsymbol{\phi}_0 = (0, 0, v)$. Then, clearly, $H = \mathrm{SO}(2)$, the group of rotations about the z-axis, and $\mathcal{M} = \mathrm{SO}(3)/\mathrm{SO}(2) = S^2$, the 2-sphere. It is useful to think of $\mathcal{M}$ as the subset of the order-parameter space comprising the possible expectation values $\boldsymbol{\phi} = \langle\hat{\boldsymbol{\phi}}\rangle$ for the various degenerate ground states. For example, in this case, $\mathcal{M} = \{\boldsymbol{\phi}: \boldsymbol{\phi}^2 = v^2\}$.

Defect Formation

It is often possible to characterize the dynamics at finite temperature in terms of a function of the order parameter, the effective potential $V(\phi)$, which is necessarily invariant under G, and whose minima define the equilibrium states. At low temperatures, it has a form like $V = \lambda(\boldsymbol{\phi}^2 - v^2)^2$, whose minima occur at nonzero values of $\boldsymbol{\phi}$. But above the critical temperature T_c, the only minimum is at $\boldsymbol{\phi} = \mathbf{0}$, so the equilibrium state is symmetric under G. In the high-temperature phase, there may be large fluctuations in $\hat{\boldsymbol{\phi}}$, but its mean value will be zero.

Now, when the system is cooled through the phase transition, $\hat{\boldsymbol{\phi}}$ will acquire a nonzero expectation value, gradually approaching one of the degenerate ground states characterized by a point of $\mathcal{M}$. But the choice of which one is unpredictable; the symmetry breaking is spontaneous. Moreover, in a large system, there is no reason why the same choice should be made everywhere. For example, a ferromagnet cooling through its Curie point may acquire a spontaneous magnetization in different directions in different parts of the sample.

Of course, there is an energetic penalty to having a spatially varying order parameter, so it will tend to become more uniform as the temperature is lowered. But the question arises whether there may be any topological obstruction to this process. It can happen that if we choose points on $\mathcal{M}$ in a continuous manner everywhere around the periphery of some region, it is topologically impossible to complete the process throughout its interior. Continuity may require that there are points where ϕ leaves the surface $\mathcal{M}$. For example, if our ferromagnet has two opposite possible directions of easy magnetization, described by $\boldsymbol{\phi}_0$ and $-\boldsymbol{\phi}_0$, then $\mathcal{M}$ consists essentially of these two points. Regions where $\boldsymbol{\phi} \approx \boldsymbol{\phi}_0$ and where $\boldsymbol{\phi} \approx -\boldsymbol{\phi}_0$ must be separated by domain walls across which $\boldsymbol{\phi}$ varies smoothly from one to the other.

Homotopy Groups

To classify the various possible types of defect, we need to consider the homotopy groups of the manifold $\mathcal{M}$ of degenerate ground states. In this section, we briefly review the necessary definitions.

A path in $\mathcal{M}$ is a map $\phi: I \to \mathcal{M}$ from the unit interval $I = [0, 1] \subset \mathbb{R}$. We choose a base point $m_0 \in \mathcal{M}$ (which may be identified with ϕ_0), and consider loops in $\mathcal{M}$, paths such that $\phi(0) = \phi(1) = m_0$. We say that two loops are homotopic, and write $\phi \sim \psi$, if one can be continuously deformed into the other within $\mathcal{M}$, that is, if there exists a map $\chi: I^2 \to \mathcal{M}$ such that

$$\chi(0, t) = \phi(t) \quad \text{and} \quad \chi(1, t) = \psi(t) \qquad [4]$$

for all t, and

$$\chi(s, 0) = \chi(s, 1) = m_0 \qquad [5]$$

for all s. This is an equivalence relation. The set $\pi_1(\mathcal{M})$ is the set of equivalence classes $[\phi]$ of loops under this relation.

On the set of loops, we may define a product $\phi\psi$, comprising the loop ϕ followed by ψ (see **Figure 1**). Explicitly,

$$(\phi\psi)(t) = \begin{cases} \phi(2t), & 0 \le t \le \frac{1}{2} \\ \psi(2t-1), & \frac{1}{2} < t \le 1 \end{cases} \qquad [6]$$

It is easy to show that if $\phi \sim \phi'$ and $\psi \sim \psi'$, then $\phi\psi \sim \phi'\psi'$. Hence, this defines a product on $\pi_1(\mathcal{M})$, by $[\phi][\psi] = [\phi\psi]$. So equipped, $\pi_1(\mathcal{M})$ becomes the

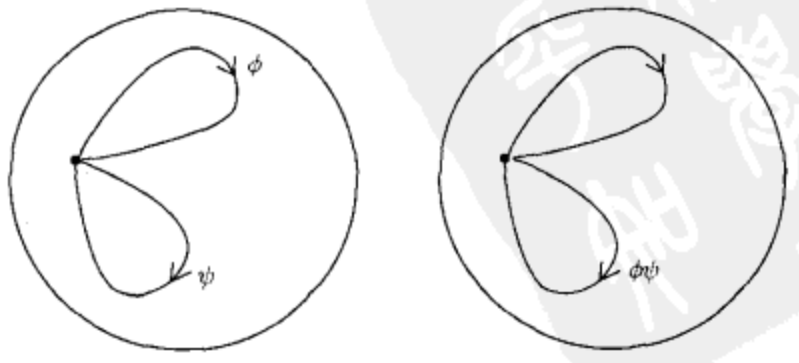


Figure 1 The product of loops.

fundamental group or first homotopy group of $\mathcal{M}$. Note that the identity is the equivalence class $[\phi_0]$ of the trivial loop with $\phi_0(t) \equiv m_0$, while the inverse is $[\phi]^{-1} = [\tilde{\phi}]$, where the map $\tilde{\phi}$ is the reverse of ϕ: $\tilde{\phi}(t) = \phi(1-t)$.

Strictly speaking, we should write $\pi_1(\mathcal{M}, m_0)$ in place of $\pi_1(\mathcal{M})$. However, for any path-connected space, the groups $\pi_1(\mathcal{M}, m_0)$ and $\pi_1(\mathcal{M}, m_0')$ are always isomorphic, and, more importantly, the same is true for any coset space $\mathcal{M} = G/H$, where G is a Lie group and H a closed subgroup. For a general manifold $\mathcal{M}$, $\pi_1(\mathcal{M})$ is not necessarily abelian, but it is so if $\mathcal{M}$ is a Lie group, or more generally a Riemannian symmetric space. The space $\mathcal{M}$ is said to be simply connected if $\pi_1(\mathcal{M}) = 0$, the group comprising only the identity element, $0 = \{[\phi_0]\}$. (Although $\pi_1(\mathcal{M})$ is not always abelian, it is conventional for homotopy groups to use an additive notation and represent the trivial group by 0 rather than 1.)

The nth homotopy group $\pi_n(\mathcal{M})$ may be defined similarly, as a set of equivalence classes of maps $\phi: I^n \to \mathcal{M}$ such that ϕ maps the entire boundary ∂I^n to the base point m_0. Two such maps are homotopic ($\phi \sim \psi$) if there exists a map $\chi: I^{n+1} \to \mathcal{M}$ such that

$$\chi(0,t) = \phi(t) \quad \text{and} \quad \chi(1,t) = \psi(t) \qquad [7]$$

for all $t = (t_1, \ldots, t_n)$, and, for each $s \in I$, $\chi(s,t) = m_0$ for all $t \in \partial I^n$. The product $\phi\psi$ is defined by

$$(\phi\psi)(t_1, \ldots, t_n) = \begin{cases} \phi(2t_1, t_2, \ldots, t_n), & 0 \le t_1 \le \frac{1}{2} \\ \psi(2t_1 - 1, t_2, \ldots, t_n), & \frac{1}{2} < t_1 \le 1 \end{cases} \qquad [8]$$

The choice of t_1 rather than any other t_j is arbitrary; all choices yield homotopic product maps. The product again defines a product on $\pi_n(\mathcal{M})$, which thereby becomes a group, the nth homotopy group. One new feature is that, for all $n > 1$, $\pi_n(\mathcal{M})$ is always abelian.

Note that since the entire boundary of I_n is mapped to a single point, it is possible to collapse it, and talk instead about maps from the n-sphere S^n to $\mathcal{M}$, taking one designated point to m_0. The fact that $\pi_n(\mathcal{M})$ is nontrivial indicates the existence in $\mathcal{M}$ of closed n-surfaces that cannot be smoothly shrunk to a point. In particular, it is worth noting that, for any n, $\pi_n(S^n) = \mathbb{Z}$, the additive group of integers, while $\pi_m(S^n) = 0$ for all $m < n$.

A special case is $n = 0$. Here, S^0 comprises two points only, and since one of them is always mapped to m_0, we really have to consider maps from a single point to $\mathcal{M}$, that is, points in $\mathcal{M}$. Two points are homotopic if they can be joined by a path in $\mathcal{M}$. Thus, $\pi_0(\mathcal{M})$ may be identified with the set of path-connected components of $\mathcal{M}$. Note, however, that in general no product can be defined on $\pi_0(\mathcal{M})$, so $\pi_0(\mathcal{M})$ should be called the zeroth homotopy set (not group). There is an important exception, however: if G is a Lie group, and G_0 its connected subgroup (the subset of elements joined by paths to the identity e), then $\pi_0(\mathcal{M})$ may be identified with the quotient group G/G_0. Note, however, that this group $\pi_0(\mathcal{M}) = G/G_0$ is not necessarily abelian.

Classification of Defects

We now turn to the classification of defects by means of homotopy groups. It will be useful to start with simple specific examples in three-dimensional space, $\mathbb{R}^3$.

First, suppose again that $\boldsymbol{\phi}$ belongs to the vector representation of $G = \mathrm{SO}(3)$. Then $\mathcal{M} = \mathrm{SO}(3)/\mathrm{SO}(2) = S^2$ may be identified with the sphere $\mathcal{M} = \{\boldsymbol{\phi}: \boldsymbol{\phi}^2 = v^2\}$ in ϕ space. Consider a closed surface $\mathcal{S}$, an embedding of a 2-sphere S^2 in $\mathbb{R}^3$. Assume that everywhere on $\mathcal{S}$ the field $\boldsymbol{\phi}(\boldsymbol{r})$ has one of the ground-state values. In other words, we have a map $\boldsymbol{\phi}: \mathcal{S} \to \mathcal{M}$, from one 2-sphere to another. The map $\boldsymbol{\phi}$ can be extended to a map from the interior of $\mathcal{S}$ to $\mathcal{M}$ only if it belongs to the trivial homotopy class $[\boldsymbol{\phi}_0] \in \pi_2(\mathcal{M})$, where $\phi_0: I^2 \to \mathcal{M}: (t_1, t_2) \mapsto m_0 = eH$. In all other cases, there must be at least one point where $\boldsymbol{\phi}(\boldsymbol{r}) = 0$; this is a point defect. The second homotopy group in this case is $\pi_2(S^2) = \mathbb{Z}$, so the possible point defects, or monopoles, are labeled by an integer $n \in \mathbb{Z}$, the winding number. (An example of a map with winding number n is (in spherical polars) $(r, \theta, \varphi) \mapsto (v, \theta, n\varphi)$.)

More generally, point defects in $\mathbb{R}^d$ are classified by $\pi_{d-1}(\mathcal{M})$. A map ϕ from a closed $(d-1)$-dimensional surface $\mathcal{S} \subset \mathbb{R}^d$ to $\mathcal{M}$ can be extended to the interior of $\mathcal{S}$ if and only if it belongs to the trivial homotopy class $[\phi_0] \in \pi_{d-1}(\mathcal{M})$. If this is not the case, there must be at least one point around which $\phi(\boldsymbol{r})$ leaves the surface $\mathcal{M}$, although in general it is not required to vanish anywhere.

Second, take the case where ϕ is a single complex field, and G is the phase symmetry group U(1). In this case, H is the subgroup $\mathbf{1} = \{1\} \subset G$. Thus, $\mathcal{M} = \mathrm{U}(1)/\mathbf{1} = S^1$; this manifold may be identified with the circle $\{\phi: |\phi| = v\}$ in the order-parameter space. Now consider a closed loop $\mathcal{C}$ in space, an embedding of S^1 in $\mathbb{R}^3$ (see **Figure 2**). Suppose that on $\mathcal{C}$, $\phi(\boldsymbol{r})$ takes one of the ground-state values, say $\phi(\boldsymbol{r}) = v \exp[i\alpha(\boldsymbol{r})]$. If $\mathcal{S}$ is some surface with boundary $\mathcal{C}$, then the map $\phi: \mathcal{C} \to \mathcal{M}$ can be extended to a map $\phi: \mathcal{S} \to \mathcal{M}$ if and only if it belongs to the trivial homotopy class $[\phi_0] \in \pi_1(\mathcal{M})$. If it does not, then there must be at least one point on $\mathcal{S}$ within $\mathcal{C}$ where $\phi = 0$. Moreover, this

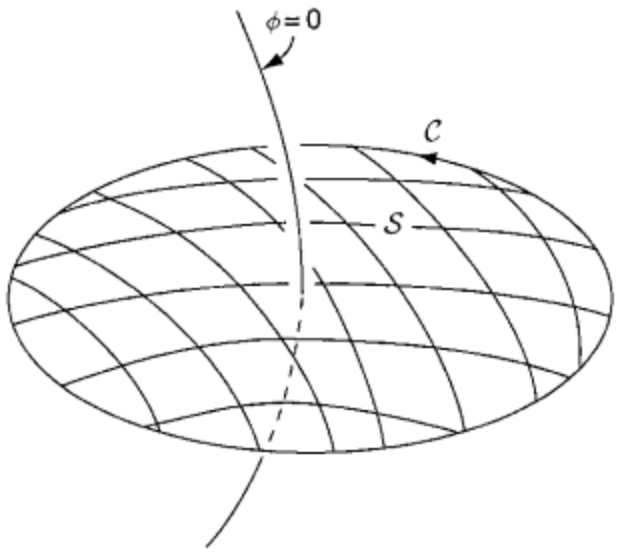


Figure 2 A linear defect.

must be true of every surface S spanning $\mathcal{C}$, so there must be a curve passing through $\mathcal{C}$ along which $\phi = 0$. This is a linear defect, a string or vortex line. In this case, the first homotopy group is $\pi_1(S^1) = \mathbb{Z}$, so we see that the possible linear defects are classified by an integer, the winding number n. An example of a map with winding number n is $\varphi \mapsto ve^{ni\varphi}$.

Again, this result can easily be generalized. Linear defects in $\mathbb{R}^d$ are classified by $\pi_{d-2}(\mathcal{M})$. If, on a $(d-2)$-dimensional surface $\mathcal{C}$, $\phi(r)$ takes values in $\mathcal{M}$, and if it does not belong to the trivial homotopy class, there must be a linear defect threading through $\mathcal{C}$, around which ϕ leaves the surface $\mathcal{M}$ – although again it need not necessarily vanish.

More generally yet, in the d-dimensional space $\mathbb{R}^d$, defects of dimension p are classified by the homotopy group $\pi_{d-p-1}(\mathcal{M})$. For example, in three dimensions, planar defects – domain walls – are classified by $\pi_0(\mathcal{M})$.

The Exact Sequence

There are mathematical theorems that greatly facilitate the computation of the homotopy group of homogeneous spaces, of the form $\mathcal{M} = G/H$.

We begin with the maps relating these spaces to each other. There is a canonical injective homomorphism $i: H \to G: h \mapsto h$, and a canonical projection associating each element of G with its coset: $p: G \to \mathcal{M}: g \mapsto gH$. Moreover, it is clear that the image of i, namely the subgroup H, is also the kernel of p, the inverse image $p^{-1}m_0$ of the distinguished element $m_0 = eH$ of $\mathcal{M}$. These statements can be summarized by saying that

$$1 \to H \xrightarrow{i} G \xrightarrow{p} \mathcal{M} \to 1$$

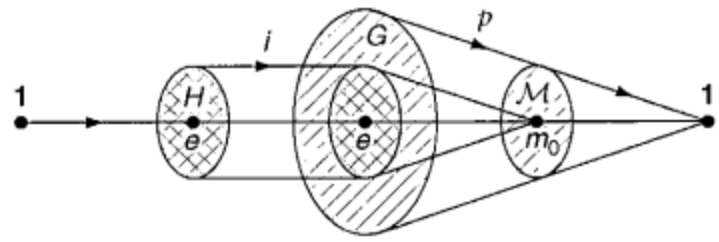


Figure 3 An exact sequence.

is an exact sequence: the image of each map is the kernel of the following one (see **Figure 3**).

Next, we note that since any closed loops (or n-surfaces) in H belonging to the same homotopy class are also homotopic as loops (or n-surfaces) in G, there is an induced homomorphism $i_*: \pi_n(H) \to \pi_n(G)$. Similarly, homotopic loops or n-surfaces in G project to homotopic loops or n-surfaces in $\mathcal{M}$, so there is an induced homomorphism $p_*: \pi_n(G) \to \pi_n(\mathcal{M})$. Moreover, it is easy to see that although i_* is not necessarily injective and p_* not necessarily projective, it is true that the image of i_* is the kernel of p_*. For example, any loop in G will be mapped to a homotopically trivial loop in $\mathcal{M}$ if and only if it is homotopic to the image of a loop in H.

In addition, there is a boundary map that relates homotopy groups of different dimension: $\partial: \pi_{n+1}(\mathcal{M}) \to \pi_n(H)$. To see this, it is useful to think of G as a fiber bundle with base space $\mathcal{M}$ and fiber H. Now consider a map $\phi: (I^{n+1}, \partial I^{n+1}) \to (\mathcal{M}, m_0)$. Since p is a projection, ϕ can always be lifted to a map $\hat{\phi}: (I^{n+1}, \partial I^{n+1}) \to (G, H)$, that is, we can find a (nonunique) map $\hat{\phi}$ such that $\phi = p \circ \hat{\phi}$ (see **Figure 4**). However, $\hat{\phi}$ does not necessarily map the boundary to a single point; what is true is that $\hat{\phi}$ must map the boundary to a subset of H, and since topologically $\partial I^{n+1} \simeq S^n$, this defines a map $\tilde{\phi}: S^n \to H$. If we allow ϕ to vary over some homotopy class of maps, and $\hat{\phi}$ to vary continuously, then $\tilde{\phi}$ will

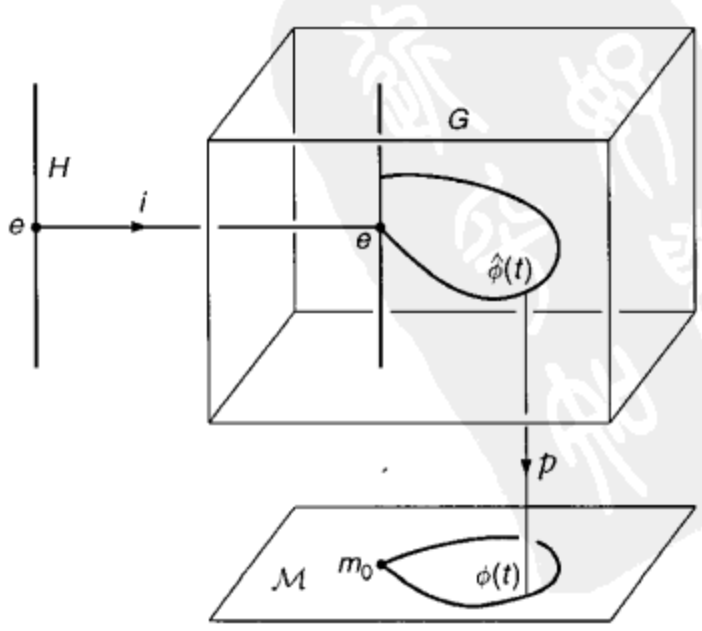


Figure 4 Lift of a loop.

also remain in one homotopy class. Thus, we have defined a map $\partial: \pi_{n+1}(\mathcal{M}) \rightarrow \pi_n(H): [\phi] \mapsto [\tilde{\phi}]$.

It is also easy to see that the image of $\partial: \pi_{n+1}(\mathcal{M}) \rightarrow \pi_n(H)$ is the kernel of $i_*: \pi_n(H) \rightarrow \pi_n(G)$, because the n-surface in H defined by $\tilde{\phi}$ is necessarily homotopically trivial in G. Similarly, one can see that the image of $p_*: \pi_{n+1}(G) \rightarrow \pi_{n+1}(\mathcal{M})$ is the kernel of $\partial: \pi_{n+1}(\mathcal{M}) \rightarrow \pi_n(H)$.

Putting all these results together, we see that there is a (semi-infinite) exact sequence connecting all the homotopy groups:

$$\begin{aligned} \cdots \xrightarrow{p_*} \pi_{n+1}(\mathcal{M}) \xrightarrow{\partial} \pi_n(H) \xrightarrow{i_*} \pi_n(G) \xrightarrow{p_*} \pi_n(\mathcal{M}) \\ \xrightarrow{\partial} \pi_{n-1}(H) \xrightarrow{i_*} \cdots \pi_1(G) \xrightarrow{p_*} \pi_1(\mathcal{M}) \xrightarrow{\partial} \pi_0(H) \\ \xrightarrow{i_*} \pi_0(G) \xrightarrow{p_*} \pi_0(\mathcal{M}) \end{aligned} \quad [9]$$

This sequence makes it easy to compute most of the low-dimensional homotopy groups of $\mathcal{M}$. Let us begin with $\pi_0(\mathcal{M})$, which merely labels its disconnected components. As noted earlier, for the Lie group G, $\pi_0(G)$ is the quotient group $\pi_0(G) = G/G_0$, where G_0 is the connected subgroup of G. Now the image of $\pi_0(H)$ under i_* is clearly the set of connected components of G that contain elements of H, so if G has m connected components, and n of them contain elements of H, then $\pi_0(\mathcal{M})$ has m/n elements (see **Figure 5**).

Next, we note that, for all the higher homotopy groups, disconnected pieces are irrelevant. Since a loop, for example, starting at m_0 must remain within its connected component $\mathcal{M}_0 \subset \mathcal{M}$, it follows that $\pi_1(\mathcal{M}) = \pi_1(\mathcal{M}_0)$, and similarly $\pi_n(\mathcal{M}) = \pi_n(\mathcal{M}_0)$ for all $n > 1$. So one can ignore any disconnected parts of the symmetry group G, and assume from now on that $\pi_0(G) = 0$. Moreover, it is always possible to replace G by its simply connected covering group, replacing SO(3), for example, by SU(2). Thus, we may also assume that $\pi_1(G) = 0$. Then the section of the exact sequence in the second line of [9] becomes

$$0 \xrightarrow{p_*} \pi_1(\mathcal{M}) \xrightarrow{\partial} \pi_0(H) \xrightarrow{i_*} 0$$

which implies that the two groups in the center are isomorphic:

$$\pi_1(\mathcal{M}) = \pi_0(H) \quad [10]$$

For example, if the symmetry group $G = \mathrm{SO}(3)$ is completely broken, so that $H = 1$, then replacing G by $\tilde{G} = \mathrm{SU}(2)$ requires replacing H by $\tilde{H} = \{+1, -1\} \simeq \mathbb{Z}_2$, hence also $\pi_1(\mathcal{M}) = \pi_0(\tilde{H}) = \mathbb{Z}_2$; there is only one nontrivial class of linear defects in this model.

To find $\pi_2(\mathcal{M})$, we need a standard theorem about Lie groups, namely that the second homotopy group of any Lie group is trivial: for any G, $\pi_2(G) = 0$. (No details of the proof are given here. It derives from the fact that a generic element $g \in G$ belongs to a unique one-parameter subgroup $\{\exp(tX), t \in \mathbb{R}\} \subset G$, where X is an element of the Lie algebra of G. Thus, all the points on a surface in G may be joined by these paths to the identity, and the surface may then be shrunk along the resulting cone. There are exceptional elements for which this is not true, but it can be shown that in a d-dimensional group they lie on $(d-3)$-dimensional surfaces, so any 2-surface can be smoothly deformed to avoid them.)

It follows from this theorem that another section of the exact sequence is

$$0 \xrightarrow{p_*} \pi_2(\mathcal{M}) \xrightarrow{\partial} \pi_1(H) \xrightarrow{i_*} 0$$

which again implies an isomorphism:

$$\pi_2(\mathcal{M}) = \pi_1(H) \quad [11]$$

For example, if $G = \mathrm{SO}(3)$ and $H = \mathrm{SO}(2)$, or equivalently $\tilde{G} = \mathrm{SU}(2)$ and $\tilde{H} = \mathrm{U}(1)$ (a double cover of the SO(2)), then $\pi_2(\mathcal{M}) = \pi_1(\tilde{H}) = \mathbb{Z}$, so point defects in this theory are labeled by an integer winding number.

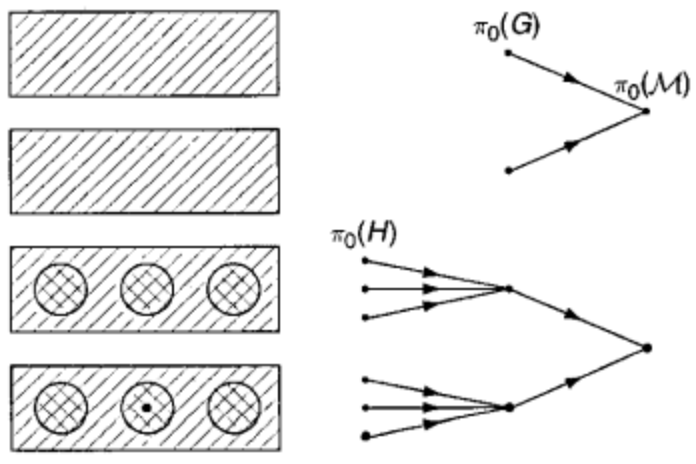


Figure 5 The disconnected components of G are shaded, those of H are cross-hatched. Here $\pi_0(\mathcal{M})$ has two elements.

Examples

The simplest continuous symmetry is the U(1) phase symmetry $\hat{\phi} \mapsto \hat{\phi} e^{i\alpha}$ of a complex field. In a weakly interacting Bose gas, below the Bose–Einstein condensation temperature, or in superfluid helium-4, a macroscopic fraction of the atoms occupies a single quantum state, and $\hat{\phi}$ acquires a nonzero expectation value, $\langle \hat{\phi} \rangle = \phi$, whose phase is arbitrary, so the symmetry is completely broken to $H = 1$. Thus, $\mathcal{M} = S^1$; we have a circle of equivalent degenerate ground states. (This corresponds to

spontaneous breaking of the particle-number symmetry. It is possible to describe the system in a U(1)-invariant way, by projecting out a state of definite particle number, a uniform superposition of all the states in $\mathcal{M}$, but it is generally less convenient to do so.) In this case, the only nontrivial homotopy group is $\pi_1(\mathcal{M}) = \mathbb{Z}$, so the only defects are linear defects classified by a winding number $n \in \mathbb{Z}$. The defects with $n = \pm 1$ are stable vortices. Those with $|n| > 1$ are in general unstable and tend to break up into $|n|$ single-quantum vortices.

Low-temperature superconductors also have a U(1) symmetry, although there are important differences. This is not a global symmetry but a local, gauge symmetry, with coupling to the electromagnetic field. Moreover, it is not single atoms that condense but Cooper pairs, pairs of electrons of equal and opposite momentum and spin. These systems too exhibit linear defects, magnetic flux tubes carrying a magnetic flux $4\pi n\hbar/e$.

A less trivial example is a nematic liquid crystal. These materials are composed of rod-shaped molecules that tend, at low temperatures, to line up parallel to one another. The nematic state is characterized by a preferred orientation, described by a unit vector $\boldsymbol{n}$, the director. (Note that $\boldsymbol{n}$ and $-\boldsymbol{n}$ are physically equivalent.) There is long-range orientational order, with molecules preferentially lining up parallel to $\boldsymbol{n}$, but unlike a solid crystal there is no long-range translational order – the molecules move freely past each other as in a normal liquid.

A convenient order parameter here is the mean mass quadrupole tensor Φ of a molecule. In the nematic state, Φ is proportional to $(3\boldsymbol{nn} - \mathbf{1})$; for example, if $\boldsymbol{n} = (0, 0, 1)$, then Φ is diagonal with diagonal elements proportional to $(-1, -1, 2)$. In this case, the symmetry group is SO(3) (or, more precisely, O(3); but the inversion symmetry is not broken, so we can restrict our attention to the connected part of the group). The subgroup H that leaves this Φ invariant is a semidirect product, $H = SO(2) \ltimes \mathbb{Z}_2$ (isomorphic to O(2)), composed of rotations about the z-axis and rotations through π about axes in the x–y plane. (If we enlarge G to its simply connected covering group $\tilde{G} = SU(2)$, then H becomes $\tilde{H} = [U(1) \ltimes \mathbb{Z}_4]/\mathbb{Z}_2$, where U(1) is generated as before by J_z. The essential difference is that the square of any of the elements in the disconnected piece of $\tilde{H}$ is not now the identity but the element $e^{2\pi i J_z} = -1 \in U(1)$.) The manifold $\mathcal{M}$ of degenerate ground states in this case is the projective space $\mathbb{R}P^2$ (obtained by identifying opposite points of S^2).

Since $\tilde{H}$ has disconnected pieces, we have $\pi_1(\mathcal{M}) = \pi_0(\tilde{H}) = \mathbb{Z}_2$. Thus, there can be topologically

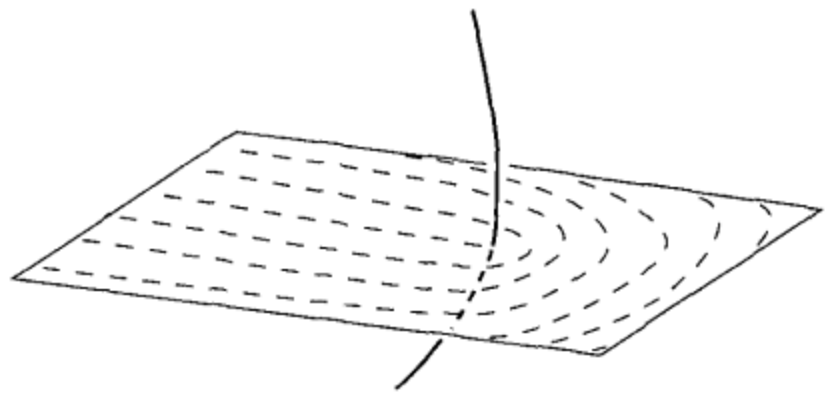


Figure 6 Orientation of molecules around a disclination line.

stable linear defects, here called disclination lines, around which the director $\boldsymbol{n}$ rotates by π (see **Figure 6**). The fact that these defects are classified by $\mathbb{Z}_2$ rather than $\mathbb{Z}$ means that a line around which $\boldsymbol{n}$ rotates by 2π is topologically trivial; indeed, $\boldsymbol{n}$ can be smoothly rotated near the line to run parallel to it, leaving a configuration with no defect.

There are also point defects; since $\pi_2(\mathcal{M}) = \pi_1(\tilde{H}) = \mathbb{Z}$, they are labeled by an integer winding number n. In a defect with $n = 1$, the vector $\boldsymbol{n}$ points radially outwards all round the defect position.

Helium-3

Finally, let us turn to helium-3, one of the most fascinating and complex examples of spontaneous symmetry breaking, which becomes a superfluid at a temperature of a few millikelvin. Unlike helium-4, this is, of course, a Fermi liquid, so it is not the atoms that condense, but bound pairs of atoms, analogous to Cooper pairs. In this case, however, the most attractive channel is not the ^{1}S, but the ^{3}P, so the pairs have both orbital and spin angular momentum, $L = S = 1$. Therefore, the order parameter is not a single complex scalar field but a 3×3 complex matrix Φ_{jk}, where the two indices label the orbital and spin angular momentum states.

To a good approximation, the system is invariant under separate rotations of $\boldsymbol{L}$ and $\boldsymbol{S}$ (the effects of the small spin–orbit coupling will be discussed later), so the symmetry group is

$$G = U(1)_Y \times SO(3)_L \times SO(3)_S \qquad [12]$$

where the subscripts denote the generators and $U(1)_Y$ represents multiplication by an overall phase factor, $e^{i\alpha Y} : \Phi_{jk} \mapsto \Phi_{jk} e^{i\alpha}$. This complicated symmetry allows much scope for a large variety of defects. There are, in fact, two distinct superfluid phases, A and B, with different symmetries (and indeed in the presence of a magnetic field there is a third, $A1$).

In the ^{3}He-A phase, the order parameter has the form $\Phi_{jk} \propto (m_j + in_j)d_k$, where $\boldsymbol{m}$, $\boldsymbol{n}$, $\boldsymbol{d}$ are unit vectors, with $\boldsymbol{m} \perp \boldsymbol{n}$; if we set $\boldsymbol{l} = \boldsymbol{m} \wedge \boldsymbol{n}$, then

$\boldsymbol{l}$ defines the orbital angular momentum state by $\boldsymbol{l}\cdot\boldsymbol{L}=1$, while $\boldsymbol{d}$ defines the spin quantization axis, such that $\boldsymbol{d}\cdot\boldsymbol{S}=0$. The manifold $\mathcal{M}_A$ for this phase is

$$\mathcal{M}_A = [\mathrm{SO}(3)\times S^2]/\mathbb{Z}_2 \qquad [13]$$

where the $\mathbb{Z}_2$ is present because $(\boldsymbol{m}, \boldsymbol{n}, \boldsymbol{d})$ and $(-\boldsymbol{m}, -\boldsymbol{n}, -\boldsymbol{d})$ represent the same state. If, for example, we take $\boldsymbol{l}$ and $\boldsymbol{d}$ in the z-direction, the unbroken symmetry subgroup is

$$H_A = \mathrm{SO}(2)_{L_z+Y}\times[\mathrm{SO}(2)_{S_z}\ltimes\mathbb{Z}_2] \qquad [14]$$

where the nontrivial element of $\mathbb{Z}_2$ may be taken to be $e^{i\pi(S_x+L_z)}$. The covering group of G is, of course,

$$\tilde{G} = \mathbb{R}_Y \times \mathrm{SU}(2)_L \times \mathrm{SU}(2)_S \qquad [15]$$

Correspondingly,

$$\tilde{H}_A = \mathbb{R}_{L_z+Y}\times[\mathrm{U}(1)_{S_z}\ltimes\mathbb{Z}_4] \qquad [16]$$

It follows that the homotopy groups are

$$\pi_0(\mathcal{M}_A)=0,\quad \pi_1(\mathcal{M}_A)=\mathbb{Z}_4,\quad \pi_2(\mathcal{M}_A)=\mathbb{Z} \qquad [17]$$

There are linear defects labeled by a mod-4 quantum number and point defects labeled by an integer.

For the ^{3}He-B phase, by contrast, the order parameter is of the form

$$\Phi_{jk}\propto R_{jk}e^{i\theta} \qquad [18]$$

where R is a rotation matrix, $R\in\mathrm{SO}(3)$. Here then,

$$\mathcal{M}_B = \mathrm{SO}(3)\times S^1 \qquad [19]$$

with homotopy groups

$$\pi_0(\mathcal{M}_B)=0,\quad \pi_1(\mathcal{M}_B)=\mathbb{Z}_2\times\mathbb{Z},\quad \pi_2(\mathcal{M}_B)=0 \qquad [20]$$

In this phase, there are two distinct types of linear defect, the mass vortices with an integer label, and the spin vortices with a mod-2 label. (One can also have a "spin–mass vortex" carrying both quantum numbers.)

Composite Defects

There are several cases, including in particular helium-3, that exhibit symmetry breaking with multiple length or energy scales. For example, there may be two order parameters, say ϕ, ψ, with $|\phi|\gg|\psi|$. If $|\psi|$ is negligible, the symmetry G is broken by ϕ to H, and the manifold of degenerate ground states is $\mathcal{M}=G/H$. However, these states are not all exactly degenerate: ψ breaks the symmetry further to $K\subset H$, so the precisely degenerate ground states form a submanifold $\mathcal{M}'=G/K$.

The case of helium-3 is slightly different. Here it is the small spin–orbit coupling, arising from long-range dipole–dipole interactions, that introduces the second scale. Its effect is only significant over large distances.

In the ^{3}He-A phase, at short range the $\boldsymbol{l}$ and $\boldsymbol{d}$ vectors are uncorrelated but, over large distances, they tend to be aligned parallel or antiparallel. We can use the $\mathbb{Z}_2$ symmetry mentioned earlier to choose $\boldsymbol{l}=\boldsymbol{d}$. Hence, the manifold $\mathcal{M}'_A$ of true ground states is only a submanifold of $\mathcal{M}_A$, namely $\mathcal{M}'_A=\mathrm{SO}(3)$, whose homotopy groups are

$$\pi_0(\mathcal{M}'_A)=0,\quad \pi_1(\mathcal{M}'_A)=\mathbb{Z}_2,\quad \pi_2(\mathcal{M}'_A)=0 \qquad [21]$$

Because of different behavior on different scales, "composite" defects can arise. For example, because $\pi_2(\mathcal{M}_A)=\mathbb{Z}$, there are short-range monopole configurations. For the $n=1$ monopole, we have a configuration with uniform $\boldsymbol{l}$, and with $\boldsymbol{d}$ pointing outwards from the center. But, eventually the misalignment of $\boldsymbol{d}$ with $\boldsymbol{l}$ is energetically disfavored, and at large distances $\boldsymbol{d}$ tends to rotate to align with $\boldsymbol{l}$ except around one particular direction where it is oppositely aligned (see **Figure 7**). We have a composite defect: a small monopole coupled to a relatively fat string.

To see how the small- and large-scale structures fit together, one has to look also at the relative homotopy groups $\pi_n(\mathcal{M},\mathcal{M}')$, whose elements are homotopy classes of maps from I^n to $\mathcal{M}$ such that one face of the boundary is mapped into $\mathcal{M}'$, and the remainder to the chosen base point m_0. For example, $\pi_1(\mathcal{M},\mathcal{M}')$ classifies paths that terminate at m_0 while beginning at any point of $\mathcal{M}'$. There is, in fact, a long exact sequence, similar to [9], relating these homotopy groups, of which a typical segment is

$$\cdots\xrightarrow{\partial}\pi_n(\mathcal{M}')\xrightarrow{i_*}\pi_n(\mathcal{M})\xrightarrow{p_*}\pi_n(\mathcal{M},\mathcal{M}')\xrightarrow{\partial}\pi_{n-1}(\mathcal{M}')\xrightarrow{i_*}\cdots \qquad [22]$$

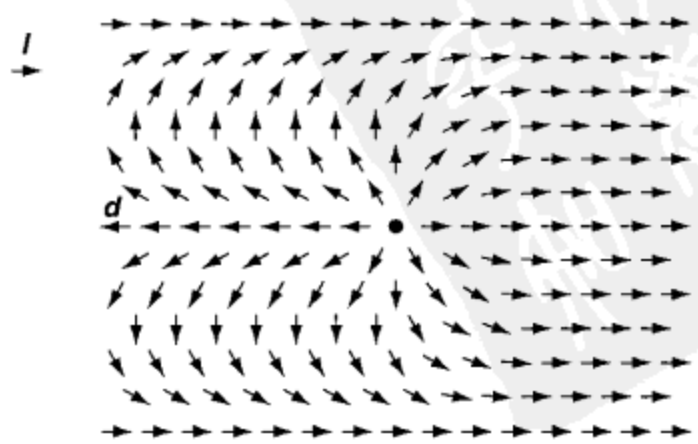


Figure 7 Cross-section of a short-range monopole attached to a fat string.

The relevant groups in the present case are

$$\pi_1(\mathcal{M}_A, \mathcal{M}'_A) = \mathbb{Z}_2, \quad \pi_2(\mathcal{M}_A, \mathcal{M}'_A) = \mathbb{Z} \qquad [23]$$

Because $\pi_1(\mathcal{M}_A) = \mathbb{Z}_4$, there are three distinct classes of linear defects at small scales, but only those with quantum number $n = 2 \,(\mathrm{mod}\, 4)$ survive unchanged to large scales; they correspond to the nontrivial element of $\pi_1(\mathcal{M}'_A) = \mathbb{Z}_2$. On the other hand, the homotopy classes $n = \pm 1 \,(\mathrm{mod}\, 4)$ are mapped to nontrivial elements of $\pi_1(\mathcal{M}_A, \mathcal{M}'_A) = \mathbb{Z}_2$, which indicates that the corresponding linear defects are coupled at long range to fat domain walls, across which $\boldsymbol{d}$ rotates through π with a compensating rotation through π about $\boldsymbol{l}$. Similarly, the nontrivial elements of $\pi_2(\mathcal{M}_A) = \mathbb{Z}$ are mapped to nontrivial elements of $\pi_2(\mathcal{M}_A, \mathcal{M}'_A)$, confirming that these short-range monopoles are coupled to fat strings, as in **Figure 7**.

For ^{3}He-B, the effect of the spin–orbit coupling is to make the most energetically favorable configurations those in which the rotation matrix R in [18] represents a rotation about an arbitrary axis $\boldsymbol{n}$ through the Leggett angle $\theta_L = \arccos(-1/4) = 104°$: $R = \exp(-i\theta_L \boldsymbol{n} \cdot \boldsymbol{J})$.

Consequently,

$$\mathcal{M}'_B = S^2 \times S^1 \qquad [24]$$

and so

$$\pi_0(\mathcal{M}'_B) = 0, \quad \pi_1(\mathcal{M}'_B) = \mathbb{Z}, \quad \pi_2(\mathcal{M}'_B) = \mathbb{Z} \qquad [25]$$

The relative homotopy groups are

$$\pi_1(\mathcal{M}_B, \mathcal{M}'_B) = \mathbb{Z}_2, \quad \pi_2(\mathcal{M}_B, \mathcal{M}'_B) = 0 \qquad [26]$$

Here the mass vortex persists at long range, but the configuration around the spin vortex deforms so that they become attached to fat domain walls. The "monopole" configurations corresponding to nontrivial elements of $\pi_2(\mathcal{M}'_B)$ have no short-range singularity at all.

See also: Abelian Higgs Vortices; Leray–Schauder Theory and Mapping Degree; Liquid Crystals; Phase Transition Dynamics; Quantum Field Theory: A Brief Introduction; Quantum Fields with Topological Defects; Solitons and Other Extended Field Configurations; String Topology: Homotopy and Geometric Perspectives; Symmetries and Conservation Laws; Symmetry Breaking in Field Theory; Variational Techniques for Ginzburg–Landau Energies.

Further Reading

Helgason S (2001) *Differential Geometry, Lie Groups and Symmetric Spaces*. Providence, RI: American Mathematical Society.

Hu S-T (1959) *Homotopy Theory*. New York: Academic Press.

Kibble TWB (1976) Topology of cosmic domains and strings. *Journal of Physics A: Mathematical and General* 9: 1387–1398.

Kibble TWB (2000) Classification of topological defects and their relevance to cosmology and elsewhere. In: Bunkov YM and Godfrin H (eds.) *Topological Defects and the Non-Equilibrium-Dynamics of Symmetry Breaking Phase Transitions*, NATO Science Series C: Mathematical and Physical Sciences. vol. 249, pp. 7–31. Dordrecht: Kluwer Academic Publishers.

Shellard EPS and Vilenkin A (1994) *Cosmic Strings and Other Topological Defects*. Cambridge: Cambridge University Press.

Tilley DR and Tilley J (1990) *Superfluidity and Superconductivity*, 3rd edn. Bristol: IoP Publishing. .

Toulouse G and Kléman M (1976) Principles of a classification of defects in ordered media. *Journal de Physique Lettres* 37: 149.

Vollhardt D and Wölfle P (1990) *The Superfluid Phases of Helium 3*. London: Taylor and Francis.

Volovik GE (1992) *Exotic Properties of Superfluid ^{3}He*. Singapore: World Scientific.

Volovik GE and Mineev VP (1977) Investigation of singularities in superfluid ^{3}He and in liquid crystals by the homotopic topology methods. *Zhurnal Eksperimentalnoi i Teoreticheskoi Fiziki* 72: 2256–2274 *(Soviet Physics–JETP 24:* 1186–1196).

PDG

数学物理学百科全书总目录

卷 1　数学物理学导言

经典力学
微分几何
电磁学
平衡态统计力学
泛函分析
闵可夫斯基时空与狭义相对论
量子力学
拓扑学

物 理 专 题

卷 2　经典力学；流体动力学

经典力学

边界控制方法与波传播的逆问题　3
约束系统　9
余切丛约化　15
引力 N 体问题(经典)　25
哈密顿流体动力学　33
哈密顿系统：可积性困难　41
无穷维哈密顿系统　48
经典力学中的逆问题　56
KAM 理论与天体力学　61
峰孤子　72
泊松约化　81
天体力学中的稳定性问题　87
对称与辛约化　93

流体动力学

流体动力学中的分岔　105
碎裂水波　110
毛细面　114
Burgers 型方程的柯西问题　129

可压缩流：数学理论 138
流体力学：数值方法 147
地球物理动力学 157
不可压缩欧拉方程：数学理论 163
界面与多组分流体 171
湍流间歇性 180
无粘性流 188
KdV 方程与其他调制方程 195
拉格朗日弥散（被动标量） 203
磁流体动力学 209
牛顿流体与热工水力学 218
非牛顿流体 229
偏微分方程：实例 237
流的稳定性 239
超流体 246
湍流理论 253
湍流中的变分法 262
无粘性不可压缩流体：数学理论 267
涡旋动力学 278
小波：在湍流中的应用 288

卷 3 可积系统；经典，共形与拓扑场论

可积系统

可积系统：综述 3
阿贝尔希格斯涡旋 21
仿射量子群 31
Bäcklund 变换 39
Bethe 拟设 46
孤子理论的双哈密顿量方法 51
可积方程的边值问题 58
非相对论和相对论型的 Calogero-Moser-Sutherland 系统 64
可积系统的 $\bar{\partial}$ 方法 73
量子完全可积系统的本征函数 81
泛函方程与可积系统 89
完整量子场 98
瞬子：拓扑理论 103
可积性与量子场论 109
可积的离散系统 118
可积系统与代数几何 124

可积系统与离散几何 137
可积系统与作用在辛和雅可比流形上的递归算符 146
可积系统与逆散射法 152
随机矩阵理论中的可积系统 161
同步系统 166
非线性薛定谔方程 174
Painlevé 方程 183
峰孤子 188
量子 Calogero-Moser 系统 197
可积系统中的黎曼-希尔伯特方法 204
正弦戈登方程 211
孤子与 Kac-Moody 李代数 219
Toda 格子 228
扭量理论:某些应用 238
Yang-Baxter 方程 247

经典,共形与拓扑场论

拓扑量子场论：综述 259
AdS/CFT 对应 272
拓扑量子场论的公理化方法 282
BF 理论 285
边界共形场论 293
Chern-Simons 模型：严格结果 301
Donaldson-Witten 理论 306
拓扑量子场论中的对偶 314
有限型不变量 322
四维流形不变量与物理学 331
规范理论的四维流形不变量 340
h-臆微分算符及其应用 347
Jones 多项式 359
纽结理论与物理学 368
Kontsevich 积分 379
大 N 与拓扑弦 388
Mathai-Quillen 形式体系 394
数学纽结理论 403
量子场论中的算符乘积展开 411
Schwarz 型拓扑量子场论 416
孤子与其他扩展场组态 426
拓扑缺陷及其同伦分类 433

拓扑引力，二维 440
拓扑纽结理论与宏观物理学 447
拓扑 σ 模型 455
二维共形场论与顶点算子代数 461
WDVV 方程与 Frobenius 流形 467

卷 4 规范理论；量子场论

规范理论

微分形式的阿贝尔和非阿贝尔规范理论 3
阿贝尔希格斯涡旋 13
AdS/CFT 对应 23
Aharonov-Bohm 效应 33
反常 41
BRST 量子化 49
Chern-Simons 模型：严格结果 56
引力和非阿贝尔规范理论中的狄拉克场 61
Donaldson-Witten 理论 69
有效场论 77
电磁对偶 86
电弱理论 94
精确重正化群 102
弦的规范理论 112
规范理论：数学应用 118
瞬子：拓扑理论 132
大 N 与拓扑弦 139
格点规范理论 145
圈空间测度 152
非对易几何与标准模型 157
规范理论的非微扰和拓扑理论 164
微扰重正化理论与 BRST 175
量子色动力学 182
量子电动力学及其精密验证 192
重正化：通论 201
Seiberg-Witten 理论 210
粒子物理的标准模型 220
超引力 227
超对称粒子模型 234
场论中的对称性破缺 240
扭量理论：某些应用 247

二维模型 256

量子场论

量子场论：导论 273
量子场论的代数方法 278
公理化量子场论 285
Batalin-Vilkōvisky 量子化 292
外场中的玻色子与费米子 299
约束系统 308
结构量子场论 314
流代数 329
狄拉克算符与狄拉克场 335
色散关系 348
欧几里德场论 363
量子场论中的 Gerbes 373
完整量子场 381
重正化量子场论的 Hopf 代数结构 386
不定度规 395
可积性与量子场论 402
量子场论中的算符乘积展开 412
不定度规量子场：非平庸模型 417
微扰理论及其技巧 423
有拓扑缺陷的量子场 437
弯曲时空中的量子场论 446
量子相变 457
重正化：统计力学与凝聚态物质 464
散射，渐进完备性与束缚态 472
相对论量子场论中的散射：基本概念和工具 485
相对论量子场论中的散射：分析程序 495
对称与守恒律 506
量子场论的对称性：代数理论 513
低时空维度量子场论的对称性 518
热量子场论 526
Tomita-Takesaki 模块理论 535
拓扑缺陷及其同伦分类 542

卷 5　广义相对论；量子引力；弦论与 M-理论广义相对论

广义相对论：综述 3
渐进结构与共形无穷 11

黑洞力学 17
时空边界 23
膜世界 31
正则广义相对论 37
引力坍缩中的临界现象 43
广义相对论的计算方法：理论 49
宇宙学：数学理论 57
引力狄拉克场与非阿贝尔规范场 63
Einstein-Cartan 理论 71
有物质的爱因斯坦方程 78
爱因斯坦方程：精确解 84
爱因斯坦方程：初值公式 93
广义相对论：实验验证 103
几何分析与广义相对论 110
几何流与 Penrose 不等式 118
引力透镜 129
引力波 138
爱因斯坦方程的哈密顿约化 143
最小子流形 160
广义相对论的牛顿极限 173
弯曲时空中的量子场论 180
包含高自旋场的相对论波动方程 191
Friedman-Robertson-Walker 度规的冲击波修正 199
时空拓扑，因果关系和奇异性 211
旋量与自旋系数 218
闵可夫斯基空间的稳定性 225
稳态黑洞 231
扭量 238

量子引力

纽结不变量与量子引力 247
纽结理论与物理学 253
圈量子引力 265
量子宇宙学 270
圈量子引力中的量子动力学 276
量子几何学及其应用 280
自旋泡沫 287
Wheeler-De Witt 理论 298

弦理论与 M-理论

AdS/CFT 对应 311
规范理论的膜结构 321
膜与黑洞统计力学 329
几何校正与特殊拉格朗(Lagrange)日子流形 339
超弦理论的紧致化 344
导出范畴 354
弦论中的 Fourier-Mukai 变换 360
弦的规范理论 368
大 N 与拓扑弦 374
大 N 对偶 380
镜像对称性:几何概论 386
弦的非对易几何学 396
随机代数几何，吸引子与通量真空 406
黎曼完整群与例外完整群 413
弦场论 419
弦理论：现象学 429
弦拓扑学：同伦与几何透视 437
超弦理论 441
扭量理论:某些应用 449
二维共形场论与顶点算子代数 458

卷 6 凝聚态物质与光学;量子信息与量子计算;量子力学

凝聚态物质与光学

玻色爱因斯坦凝聚 3
Falicov-Kimball 模型 10
分数量子霍尔效应 19
高 T_c 超导理论 29
Hubbard 模型 37
液晶 42
负折射与超衍射成像 51
核磁共振 60
光学刻蚀 69
量子相变 77
准周期系统 84
重正化：统计力学与凝聚态物质 92
短程自旋玻璃：亚稳态方法 100
拓扑缺陷及其同伦分类 107

量子信息与量子计算
纠缠引起的容量增加 117
量子信息容量 123
量子信息理论中的信道 131
纠缠 137
纠缠测量 143
有限 Weyl 系统 147
量子态的最优化克隆 154
量子信道：经典容量 159
量子熵 162
量子纠错与容错 169
量子信息理论中的源代码 175

量子力学
A-B 效应 187
算术量子混沌 195
相干态 205
几何相因子 214
h-膨微分算符及其应用 221
N 粒子量子散射 233
正规化形式与半经典近似 241
量子各态历经与本征函数混合 249
量子力学的散射理论 263
量子力学：基础 272
量子力学：综述 278
量子力学：弱测量 290
量子 N 体问题 298
量子自旋系统 305
薛定谔算符 312
物质稳定性 320
稳定位相近似 327
超对称量子力学 333

卷 7　无序系统;动力系统

无序系统
元胞自动机 3
拉格朗日弥散（被动标量） 16
平均场自旋玻璃与神经网络 22
逾渗理论 28

物理学中的随机矩阵理论 36
随机环境中的随机行走 46
短程自旋玻璃：亚稳态方法 65
自旋玻璃 72
随机 Loewner 演化 84
二维伊辛模型 91
Wulff 滴 98

动力系统

平均化方法 103
周期轨道的分岔 109
有界凸域 Billiards 115
中心流形，正规化形式 120
混沌与吸引子 126
余切丛约化 138
微扰理论的图形化方法 148
无限维耗散动力系统 155
动力系统与热力学 165
数学物理中的动力系统：水波实例 174
熵与定量横截性 180
各态历经理论 186
动力学中的分形维数 193
动力系统的普遍性质 202
引力 N 体问题(经典) 211
哈密顿流体动力学 219
哈密顿系统：稳定性与非稳定性理论 227
全纯动力学 233
圆的同胚与微分同胚 241
同宿现象 249
h-赝微分算符及其应用 256
双曲 Billiards 268
双曲动力系统 274
同单值变形 283
KAM 理论与天体力学 288
Lyapunov 指数与奇异吸引子 299
多标度方法 307
正规化形式与半经典近似 325
点涡动力学 333
泊松约化 347

多边 Billiards 353
准周期系统 357
随机动力系统 365
动力学 ζ 函数的正规化 374
共振 380
Riemann-Hilbert 问题 385
半经典谱与闭轨道 391
分离线分裂 398
天体力学中的稳定性问题 404
稳定理论与 KAM 410
动力学系统中的对称性与对称破缺 417
对称与辛约化 424
混沌同步 433
普适性与重正化 447
弱耦合振子 454

卷 8 平衡态统计力学;非平衡态统计力学

平衡态统计力学

Bethe 拟设 3
集团展开 8
二聚体问题 15
八顶点与硬六边形模型 22
Falicov-Kimball 模型 32
费米子系统 41
有限关联态 49
完整量子场 64
Hubbard 模型 69
平衡态统计力学中的大偏差 74
亚稳态 77
连续系统相变 81
Pirogov-Sinai 理论 88
量子中心极限定理 94
量子相变 107
量子自旋系统 114
量子统计力学：综述 121
反射正性与相变 128
短程自旋玻璃：亚稳态方法 139
统计力学与组合问题 146
界面统计力学 152

超流体 161
Toeplitz 行列式与统计力学 168
二维伊辛模型 176
Wulff 滴 183

非平衡态统计力学

非平衡态统计力学（稳定的）：综述 189
绝热活塞 200
玻耳兹曼方程(经典的和量子的) 215
玻璃态无序系统：动力学演化 222
傅里叶定律 230
相互作用粒子系统与流体动力学方程 236
相互作用随机粒子系统 243
运动学方程 249
宏观涨落与热力学泛函 258
非平衡态统计力学：动力学系统方法 268
非平衡态统计力学：理论与数值模拟 273
相变动力学 281
随机共振 287

数学专题

卷 9 代数技巧;李群和李代数；离散数学；量子群；随机方法

代数技巧

仿射量子群 3
辫子张量范畴 11
Clifford 代数及其表象 20
导出范畴 33
有限维代数与箭图 39
有限群对称破缺 49
Hopf 代数与 Q 形变量子群 55
Operads 69

李群和李代数

经典群与齐性空间 79
紧致群及其表象 91
李群：通论 102
李超代数及其表象 121
李，辛，泊松广群及其李代数体 128

伪黎曼幂零李代数 137
Riemann-Hilbert 问题 147
孤子与 Kac-Moody 李代数 153

离散数学

算术量子混沌 165
组合学：综述 175
物理中的数论 199
准周期系统 207

量子群

双交叉积 Hopf 代数与非对易时空 217
经典 r-矩阵，李双代数，及泊松李群 227
可重正化量子场论的 Hopf 代数结构 235
q-特殊函数 244
量子群微分、丛和规范理论 257
Yang-Baxter 方程 266

随机方法

行列式随机场 277
自由概率论 285
随机矩阵理论中的生长过程 293
随机矩阵理论中的可积系统 299
拉格朗日弥散（被动标量） 304
Malliavin 算法 310
圈空间测度 317
物理学中的随机矩阵理论 322
随机分配 332
随机环境中的随机行走 339
随机微分方程 358
随机流体力学 367
随机 Loewner 演化 376
随机矩阵理论中的超对称方法 383
随机矩阵理论中的对称类型 393

卷 10 复几何；微分几何；低维几何；非交换几何

复几何

导出范畴 3
规范理论：数学应用 9

弦论中的 Fourier-Mukai 变换 23
纽结同调 31
镜像对称性:几何概论 39
模空间:引论 49
Quillen 行列式 59
黎曼面 67
Riemann-Hilbert 问题 78
多复变量:基本几何理论 84
多复变量:紧流形 95
扭量理论:某些应用 104

微分几何

几何校正与特殊拉格朗(Lagrange)日子流形 115
毛细面 120
示性类 135
Einstein 流形 144
指标定理 152
相交理论 162
K-理论 168
Lorentz 几何 177
Mathai-Quillen 形式体 184
伪黎曼幂零李代数 194
Riemann-Hilbert 问题 204
黎曼和乐群与例外和乐群 209
奇性与分歧理论 215
超流形 222
扭量理论:某些应用 228

低维几何

三维流形的有限型不变量 239
Floer 同调 247
四维流形不变量与物理学 257
四维流形的规范理论不变量 266
Jones 多项式 273
纽结不变量与量子重力 282
大 N 与拓扑弦 288
量子三维流形不变量 294
Ricci 流的奇性 300

非交换几何

重正化量子场论的 Hopf 代数结构 307
非交换几何与标准模型 316
来自弦的非交换几何 322
非交换环面，Yang-Mills，弦论 331
非交换几何中的路径积分 338
量子群微分，丛与规范理论 343
量子 Hall 效应 351
Riemann-Hilbert 问题 359

卷 11 代数拓扑；辛几何与拓扑；常微分与偏微分方程

代数拓扑

特征曲线类型 3
上同调论 12
导出范畴 21
等价的上同调与 Cartan 模型 27
弦论中的 Fourier-Mukai 变换 36
指数定理 44
相交理论 54
K-理论 60
Mathai-Quillen 形式体系 69
Operads 79
谱序列 87
弦拓扑学：同伦与几何透视 98

辛几何与拓扑

经典 r-矩阵，李双代数，及泊松李群 105
切触流形 113
辛流形上泊松括号的形变 119
Fedosov 量子化 130
Floer 同调 140
分次泊松代数 150
哈密顿群作用量 158
镜像对称性：几何概论 166
多哈密顿系统 176
经典力学中的递归算符 183
奇性与分歧理论 189

常微分与偏微分方程

分歧理论 199
玻耳兹曼方程(经典的和量子的) 206
边界控制方法与波传播的逆问题 213
毛细面 219
Burgers 型方程的柯西问题 234
椭圆微分方程：线性理论 243
演化方程：线性与非线性 256
流体力学：数值方法 264
金兹堡-朗道方程 274
图像处理：数学 280
Sobolev 空间中的不等式 290
同单值变形 295
运动学方程 300
准周期势的定域化 309
磁共振成像 315
最小子流形 324
Painlevé 方程 337
偏微分方程：实例 342
包含高自旋场的相对论波动方程 344
Riemann-Hilbert 问题 352
半线性波动方程 358
微分方程的分离变量 366
对称双曲系统与冲击波 376
波动方程与衍射 383

卷 12 泛函分析和算子代数；量子化方法和路径积分；变分技术

泛函分析和算子代数

Bäcklund 变换 3
C^* 代数及其分类 10
相干态 16
自由概率论 25
量子物理中的泛函积分 33
规范理论：数学应用 47
h-赝微分算符及其应用 61
Jones 多项式 73
K-理论 82
Leray-Schauder 理论与映射度 91
Ljusternik-Schnirelman 理论 96
常见特殊函数 102

C^* 代数正映射 111
量子动力学半群 118
鞍点问题 125
线性算子谱理论 135
Tomita-Takesaki 模理论 148
冯诺依曼代数：导论，模理论，分类理论 155
冯诺依曼代数：子因子理论 161
小波：应用 167
小波：数学理论 173

量子化方法和路径积分

形变量子化 189
形变量子化与表象理论 197
形变理论 204
辛流形上泊松括号的形变 212
Fedosov 量子化 223
费曼路径积分 233
量子物理中的泛函积分 240
非对易几何中的路径积分 254
动力学 ζ 函数的正规化 259

变分技术

毛细面 267
数学物理中的控制问题 282
凸分析与对偶方法 288
自由界面与自由间断：变分问题 299
Γ-收敛与均一化 306
几何测度理论 315
哈密度雅可比方程与动力学系统：变分观点 324
变分计算中的极小极大原理 333
优化输运 341
变分法求金兹堡-朗道能量 346
微结构变分技术 354

（穆良柱 译 陈熙谋 校）

[General Information]
书名=ENCYCLOPEDIA OF MATHEMATICAL PHYSICS 4 GAUGE THEORY:QUANTUM FIELD THEORY
作者=Jean-PierreFran等著
页数=566
出版社=北京市：科学出版社
出版日期=2008
SS号=40213689
DX号=000006542262
URL=http://book.szdnet.org.cn/bookDetail.jsp?dxNumber=000006542262&d=3A0072D8F6028909A86D46585510ED946